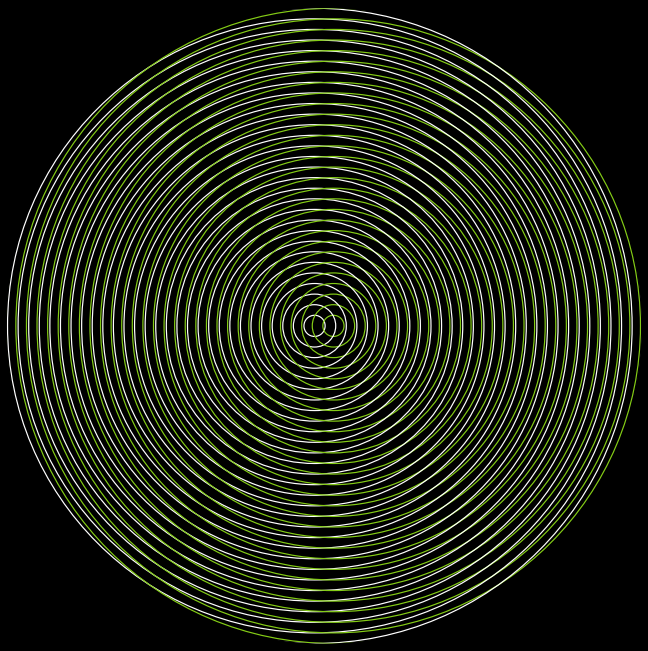
Piatra . Institute

Introduction to Metaxics

Relations, Readers, and the Origin of Perspective



2026

Piatra . Institute

Introduction to Metaxics

Relations, Readers, and the Origin of Perspective

> Two prisoners whose cells adjoin communicate with each other by knocking on the wall. The wall is the thing which separates them but it is also their means of communication. [. . .] Every separation is a link.
>
> Simone Weil, *Gravity and Grace*, "Metaxu"

2026

Contents

Preface **xix**

Notation **xxv**

I The Physics of Relations **1**

1 Objects, Relations, and Relational Observables **3**
1.1 Absolute and Relational Descriptions 3
1.2 Relational Coordinates 5
1.3 Invariance 6
1.4 Pairwise and Higher-Order Relations 7
1.5 When Is a Relational Variable Useful? 7
1.6 Same Parts, Different Relations 8
1.7 From Relation to Reader 9

2 Moiré: The Prototype of Relational Structure **13**
2.1 Two Gratings 13
2.2 Translation Moiré 14
2.3 Rotational Moiré 15
2.4 Amplification 15
2.5 What Belongs to the Layers? 16
2.6 What Belongs to the Configuration? 17
2.7 Multiplicative Versus Additive Superposition 17
2.8 Moiré as a Relational Measuring Device 19
2.9 Limits of the Analogy 20

3 Beats: Temporal Relations **23**
3.1 Two Nearby Frequencies 23
3.2 The Envelope Identity 24
3.3 Fourier Reality Versus Envelope Appearance 25
3.4 Difference and Sum Frequencies 26
3.5 Intermodulation 26
3.6 Phase Relations 27
3.7 Binaural Beats 27
3.8 The First Appearance of the Reader Problem 28

4 How New Frequencies Become Physical **31**
4.1 Linear Systems . . . 31
4.2 Quadratic Nonlinearities . . . 32
4.3 Cubic Nonlinearities . . . 33
4.4 Rectification . . . 34
4.5 Thresholding . . . 35
4.6 Multiplication . . . 35
4.7 Heterodyning . . . 36
4.8 Filtering . . . 37
4.9 Lock-in Detection . . . 37
4.10 Volterra Series . . . 38
4.11 Effective Observables . . . 39

II Special Relations Are Not One Thing **45**

5 Magnitude Balance and Crossover **47**
5.1 Dimensionless Numbers . . . 47
5.2 Order-One Transitions . . . 48
5.3 The Reynolds Number as Paradigm . . . 49
5.4 Competing Energy Scales . . . 50
5.5 Coupling Versus Heterogeneity . . . 50
5.6 Why “the Optimum Is in the Middle” Is Usually Too Weak 52

6 Resonance **55**
6.1 The Driven Oscillator . . . 55
6.2 Natural Frequency . . . 56
6.3 The Resonance Curve . . . 56
6.4 Damping . . . 56
6.5 The Quality Factor . . . 57
6.6 Nonlinear Resonance . . . 57
6.7 Parametric Resonance . . . 58
6.8 Neural Resonance . . . 59
6.9 The Resonator as a Reader . . . 60
6.10 Why Resonance Is Not Commensurability . . . 60

7 Commensurability **65**
7.1 Rational Ratios and the Torus . . . 65
7.2 Locking: The Adler Equation . . . 66
7.3 The Circle Map . . . 66
7.4 Arnold Tongues . . . 68
7.5 The Devil’s Staircase . . . 68
7.6 Locking in Living Tissue . . . 70
7.7 Three Special Relations . . . 70
7.8 Irrational Ratios and the Golden Mean . . . 71
7.9 The Golden-Ratio Hypothesis for Brain Rhythms . . . 72

7.10 A Relation the Fish Avoids 72

8 Magic Angles Without Magic **77**
8.1 Two Sheets and a Twist 77
8.2 The Moiré Brillouin Zone 78
8.3 The Bistritzer–MacDonald Model 78
8.4 What Happens on a Flat Band 81
8.5 Turning the Other Knob 81
8.6 The Chiral Limit and the Pattern of Magic Angles . . . 82
8.7 Not the Perfect Amount of Disorder 82
8.8 A Moiré Where Commensurability Does Matter 83
8.9 Why Graphene Is Not a Theory of Consciousness 83

III Collective Variables **89**

9 Synchronization **91**
9.1 From Clocks to Populations 91
9.2 Phase Oscillators 92
9.3 The Order Parameter 92
9.4 The Self-Consistency Argument 93
9.5 The Threshold as a Balance 94
9.6 Partial Synchrony and Clusters 95
9.7 Measuring Synchrony in Data 95
9.8 What Reads the Order Parameter? 96

10 Coordination Dynamics **101**
10.1 The Bimanual Experiment 101
10.2 Relative Phase as Order Parameter 102
10.3 The HKB Model 102
10.4 From the Components Up 103
10.5 Fluctuations and Critical Slowing Down 104
10.6 Breaking the Symmetry 105
10.7 Coordination Between People 105
10.8 What Reads the Relative Phase? 106

11 Metastability **109**
11.1 Integration Without Permanent Synchrony 109
11.2 The Ghost of a Fixed Point 110
11.3 Coordination Without Stable Patterns 111
11.4 Two Ways to Dwell 111
11.5 Heteroclinic Channels 112
11.6 Metastability in Brain Models 113
11.7 Two Senses of the Word 113
11.8 What Reads a Metastable Relation? 114

12 Chimera States and Frustration **117**
12.1 Identical Parts, Split Behavior . . . 117
12.2 The Minimal Chimera . . . 119
12.3 Frustration . . . 120
12.4 Chimeras in the Brain? . . . 122
12.5 What the Coupling Reads . . . 123

13 Synergetics **127**
13.1 Instability and the Separation of Timescales . . . 127
13.2 Adiabatic Elimination: A Worked Example . . . 128
13.3 The Slaving Principle . . . 129
13.4 The Laser . . . 130
13.5 Order Parameters of Behavior . . . 131
13.6 What the Vocabulary Adds . . . 132

14 Causal Emergence and Coarse-Graining **135**
14.1 Micro, Macro, and Coarse-Graining Maps . . . 135
14.2 Effective Information . . . 136
14.3 A Worked Example . . . 136
14.4 Causal Emergence and Its Critics . . . 138
14.5 Emergence as Synergy . . . 139
14.6 Causal States . . . 139
14.7 Whose Coarse-Graining? . . . 140

IV The Reader Problem **147**

15 What Is a Reader? **149**
15.1 Five Senses of "Reader" . . . 149
15.2 From Interaction to Interpretation . . . 150
15.3 The Reading Loop . . . 152
15.4 Four Loops . . . 154
15.5 Reading Is Implemented . . . 155
15.6 What the Loop Leaves Open . . . 156

16 Reader-Relative Variables **161**
16.1 One Mixture, Five Readers . . . 161
16.2 Explicit Variables . . . 163
16.3 What a Reader Cannot Make Explicit . . . 164
16.4 No Metaphysics Required . . . 164
16.5 Description and Dynamics . . . 165
16.6 Counterfactual Response . . . 165
16.7 Channels and Bottlenecks . . . 166
16.8 When the Reader Is a Relatum . . . 167

17 Does Everything Compute Everything? 171
17.1 Physical Computation . . . 172
17.2 The Triviality Problem . . . 173
17.3 Putnam . . . 173
17.4 Searle . . . 174
17.5 Counterfactual Causal Structure . . . 174
17.6 Mechanistic Accounts . . . 176
17.7 Computation Versus Description . . . 176
17.8 What an Endogenous Reader Adds . . . 178

18 Polycomputation 185
18.1 One Substrate, Multiple Computations . . . 185
18.2 Observer-Dependent Decompositions . . . 186
18.3 Granular Polycomputation . . . 187
18.4 Latent Computations . . . 187
18.5 Exploited Versus Merely Describable Mappings . . . 188
18.6 Internal Readers . . . 189
18.7 Can One Subsystem Read Another? . . . 190
18.8 Can a Substrate Read Itself? . . . 190

V The Electric Fish: A Complete Reader 197

19 Active Electrosensation 199
19.1 The Electric Organ Discharge . . . 199
19.2 Electroreceptors . . . 200
19.3 The Self-Generated Carrier . . . 200
19.4 The Reading Loop of Electrolocation . . . 202
19.5 Active and Passive Sensing . . . 203
19.6 The Vulnerability of a Self-Generated Carrier . . . 204

20 The Jamming Problem 207
20.1 Two Nearby Discharge Frequencies . . . 207
20.2 Why a Slow Beat Is a Problem . . . 208
20.3 The Jamming Avoidance Response . . . 209
20.4 Changing a Relatum to Change a Relation . . . 210
20.5 The Loop Closed and Opened in the Laboratory . . . 211
20.6 Participation Without a Participant . . . 211

21 Reading the Sign 215
21.1 Why the Magnitude Is Not Enough . . . 215
21.2 Amplitude and Differential Phase . . . 216
21.3 The Sign Is in the Joint Behavior . . . 217
21.4 The Neural Sign Computation . . . 218
21.5 A Neuronal Democracy . . . 221
21.6 No Internal Reference Is Needed . . . 221

21.7 The Counterfactual Test 222

22 Reafference **225**
22.1 The Reafference Principle 225
22.2 Reafference Across the Animal Kingdom 226
22.3 Negative Images in Pulse-Type Fish 227
22.4 Reafference as Carrier 227
22.5 When No Efference Copy Is Required 228
22.6 Measurement Indexed to One's Own Action 229
22.7 The Minimal Agent-Centered Frame 229

23 Phantom Jamming and Error **233**
23.1 Mimicking Relational Cues 233
23.2 Behavioral Misclassification 234
23.3 When Is an Error Really an Error? 234
23.4 False Positives and False Negatives 235
23.5 Teleology Versus Mere Mismatch 236
23.6 The Beginning of Semantic Questions 236

24 Convergent Evolution of Reading **241**
24.1 Two Independent Origins 241
24.2 The Same Problem 242
24.3 The Same Computational Rules 242
24.4 Different Implementations 243
24.5 Marr's Levels, Made Empirical 244
24.6 Why the Algorithm Converged 244
24.7 Multiple Realizability, Carefully 245

VI From Reading to Meaning **251**

25 The Epistemic Cut **253**
25.1 Pattee 253
25.2 Dynamics Versus Symbols 254
25.3 Measurement 255
25.4 Constraints 256
25.5 Control 257
25.6 Physical Embodiment of Measurement 258

26 Semantic Closure **263**
26.1 Closure 263
26.2 Self-Construction 264
26.3 Codes 265
26.4 Interpretation 265
26.5 The Cell as Minimal Observer? 267
26.6 Where the Metaphor Becomes Dangerous 267

27 Autopoiesis **271**
27.1 Organizational Closure . . . 271
27.2 Self-Production . . . 272
27.3 Boundary Creation . . . 273
27.4 Operational Autonomy . . . 274
27.5 Environment Versus Umwelt . . . 275
27.6 Criticisms . . . 275

28 Sense-Making **279**
28.1 Viability . . . 279
28.2 Normativity . . . 280
28.3 Adaptive Regulation . . . 281
28.4 The Minimal Case: Bacterial Chemotaxis . . . 282
28.5 Sense-Making and Enactivism . . . 283
28.6 Does Normativity Require Consciousness? . . . 284

29 Naturalizing Aboutness **289**
29.1 Intentionality . . . 289
29.2 Causal Theories and the Disjunction Problem . . . 290
29.3 Information . . . 291
29.4 Teleosemantics . . . 292
29.5 Structural Representation . . . 293
29.6 Predictive Representation . . . 294
29.7 Three Indeterminacies . . . 294
29.8 Representation Without a Homunculus . . . 294
29.9 What Naturalized Content Does Not Deliver . . . 295

30 Misrepresentation **299**
30.1 Deviation and Error . . . 299
30.2 Kinds of Error . . . 300
30.3 Signal Detection and Optimal Error . . . 301
30.4 Phantom Jamming . . . 302
30.5 What Separates Misrepresentation from Deviation . . . 303
30.6 Error Without Experience . . . 304

VII Detachment and World Models **309**

31 Representation Without the Present Stimulus **311**
31.1 Cued and Detached Representations . . . 311
31.2 Memory: The Simplest Detachment . . . 312
31.3 Prediction and Forward Models . . . 313
31.4 Emulation . . . 314
31.5 Cognitive Maps and Internally Generated Sequences . . . 315
31.6 A Scale of Detachment . . . 315
31.7 What Detachment Adds to the Reader Problem . . . 316

32 Counterfactuals **319**
32.1 What Might Be . . . 319
32.2 Prospective Control . . . 321
32.3 Planning in a Model . . . 321
32.4 Deliberation in the Hippocampus . . . 322
32.5 Mental Time Travel . . . 322
32.6 Why Controllers Need Models . . . 323
32.7 Counterfactual Capacity Without Experience . . . 324

33 Generative Models **327**
33.1 Perception as Inference . . . 327
33.2 Predictive Coding . . . 328
33.3 Free Energy and Active Inference . . . 329
33.4 Hidden States and Precision . . . 330
33.5 Action as Inference . . . 330
33.6 Allostasis and Interoceptive Inference . . . 331
33.7 What Generative Models Explain, and What They Do Not 331

VIII From World-Centered to Agent-Centered Variables 337

34 Here **339**
34.1 Coordinates with an Origin . . . 339
34.2 Chains of Frames . . . 340
34.3 Gain Fields . . . 341
34.4 Allocentric Maps and a Located Self . . . 343
34.5 Active Sensing Builds the Frame . . . 343
34.6 Why Location Is Not Perspective . . . 344

35 Now **349**
35.1 Every Reader Has a Window . . . 349
35.2 The Specious Present . . . 350
35.3 Postdiction and Time Labels . . . 351
35.4 Temporally Centered Variables . . . 352
35.5 The Fish's Now . . . 353
35.6 Does a System Require a Modeled Now? . . . 354

36 Me **357**
36.1 Self and Non-Self as Source Attribution . . . 357
36.2 Interoception and the Material Me . . . 359
36.3 Proprioception and the Body Model . . . 359
36.4 Agency . . . 359
36.5 Ownership . . . 360
36.6 Minimal and Narrative Self . . . 361
36.7 The Fish's Me . . . 362
36.8 What "Me" Does Not Yet Include . . . 362

37 The Essential Indexical **367**
37.1 The Messy Shopper . . . 367
37.2 Quasi-Indicators . . . 368
37.3 Character and Content . . . 368
37.4 Centered Worlds . . . 369
37.5 Why the Center Matters for Action . . . 370
37.6 Indexicals as Computational Variables . . . 371
37.7 What the Indexical Does Not Settle . . . 372

38 Projective Perspective **375**
38.1 The Central Projection . . . 375
38.2 What Projection Keeps . . . 376
38.3 The Projective Consciousness Model . . . 377
38.4 Moving the Center . . . 378
38.5 A Centered Arena for Action . . . 379
38.6 A Point of View Without Projection . . . 379
38.7 Is Geometry Enough? . . . 380

IX Self-Reading Systems **385**

39 The Reader Reads the Reader **387**
39.1 From Reading the World to Reading a Reading . . . 387
39.2 A Signal-Detection Model of Confidence . . . 388
39.3 Confidence in Animals . . . 389
39.4 The Value of Reading One's Own Reading . . . 390
39.5 Recursion Is Cheap . . . 391
39.6 Does the Fish Read Its Own Reading? . . . 392
39.7 What the Second Reader Adds . . . 392

40 Higher-Order Representation **395**
40.1 The Transitivity Intuition . . . 395
40.2 Varieties of Higher-Order Theory . . . 396
40.3 Unconscious First-Order States . . . 397
40.4 Evidence and Clarification . . . 397
40.5 Objections . . . 398
40.6 The Higher-Order State Is Indexical . . . 399
40.7 The Fish and the Higher-Order Question . . . 400

41 Self-Models **403**
41.1 A Reader Whose Object Includes Itself . . . 403
41.2 Metzinger's Self-Model Theory . . . 405
41.3 Building and Breaking the Model . . . 406
41.4 Model and Model-Owner . . . 407
41.5 Why Recursion Is Not Enough . . . 408

42 Attention Models **413**
42.1 Attention as Selection 413
42.2 Controlling a Process Requires a Model of It 414
42.3 Awareness as the Content of the Schema 416
42.4 A Reader of Readers, and What It Reads 417
42.5 Limits 417

X Consciousness **423**

43 Functional Perspective Versus Phenomenal Perspective **425**
43.1 Two Perspectives 425
43.2 Access and Phenomenal Consciousness 426
43.3 The Hard Problem and the Explanatory Gap 427
43.4 Deflating the Gap 428
43.5 What Would Count as Closing the Gap? 429
43.6 The Word "For" 430

44 Major Theories of Consciousness **435**
44.1 Five Questions for Every Theory 435
44.2 Global Neuronal Workspace 436
44.3 Integrated Information Theory 437
44.4 Recurrent Processing 438
44.5 Higher-Order Theories 438
44.6 Predictive Processing Accounts 439
44.7 Attention Schema, Self-Model, and Projective Theories . 439
44.8 Merker's Egocentric Integration 440
44.9 Illusionism 440
44.10 The Profiles Side by Side 441
44.11 Adversarial Testing 442
44.12 What Each Theory Says About the Fish 443

45 Neural Integration and Differentiation **447**
45.1 Segregation and Integration 447
45.2 Criticality 448
45.3 Perturbational Complexity 449
45.4 Anesthesia, Sleep, Coma, and Seizures 451
45.5 Why None of This Establishes Phenomenality 452

46 Neural Frequency Relations **457**
46.1 Oscillations and Their Background 457
46.2 Cross-Frequency Coupling 458
46.3 Frequency Ratios 459
46.4 Entrainment and Arnold Tongues 460
46.5 Intermodulation 461
46.6 Bispectrum, Bicoherence, and the Bispectral Index . . . 462

46.7 Signatures of Propofol Unconsciousness 463
46.8 The Rule 464
46.9 What Relational Measures Have Not Yet Shown 465

47 Neural Moiré Theories 469
47.1 Grid Cells 469
47.2 Oscillatory Interference 470
47.3 Moiré Grids 471
47.4 Tests and Failures 472
47.5 Retinal Moiré and Orientation Maps 474
47.6 Lessons for Theory Building 475

XI The Comparative Ladder 481

48 The Radio 483
48.1 Carrier and Modulation 483
48.2 Interference 484
48.3 Demodulation and the Superheterodyne 484
48.4 Reading the Sign of a Frequency Offset 485
48.5 Channel Selection, Feedback, and Adaptive Frequency Hopping 487
48.6 Cognitive Radio 487
48.7 What the Radio Lacks 488
48.8 The First Rung 489

49 The Cell 493
49.1 Self-Production and the Boundary 493
49.2 Chemotaxis: Reading the Present Against the Past . . . 494
49.3 The Reader Is Built by the System It Serves 496
49.4 Norms Without Minds 496
49.5 Has Meaning Appeared? 497
49.6 The Second Rung 498

50 The Electric Fish 501
50.1 What the Fish Adds 501
50.2 The Loop, Written Out 502
50.3 Perspective in the Minimal Sense 504
50.4 Is There Anything It Is Like to Be the Fish? 505
50.5 The Third Rung 506

51 The Mammalian Brain 511
51.1 Integrating Several Senses in One Model 511
51.2 Hierarchical Models of the World 512
51.3 Simulating What Is Absent 512
51.4 Models of the Body 513

51.5 Monitoring One's Own Readings . . . 514
51.6 Which Additions Matter? . . . 514
51.7 The Fourth Rung . . . 516

52 Humans 521
52.1 Language and Shared Intentionality . . . 521
52.2 "I": Explicit Self-Reference . . . 522
52.3 The Narrative Self . . . 522
52.4 Recursive Metacognition . . . 523
52.5 Reports and Their Limits . . . 523
52.6 Reveal or Constitute? . . . 524
52.7 The Complete Ladder . . . 525

XII Is There a Metaxic Principle? 531

53 Candidate Unifications 533
53.1 Seven Ingredients . . . 533
53.2 Unifications That Already Exist . . . 534
53.3 Ladders That Have Already Been Drawn . . . 536
53.4 Testing a Conjunction . . . 537
53.5 Recursion and the Self . . . 539
53.6 The Verdict . . . 540

54 Failed Grand Unifications 545
54.1 General System Theory . . . 545
54.2 Catastrophe Theory . . . 546
54.3 The Edge of Chaos . . . 547
54.4 Cybernetics: A Partial Success . . . 548
54.5 Information Metaphysics . . . 548
54.6 Premature Theories of Consciousness . . . 549
54.7 What the Failures Teach . . . 550

55 What Existing Fields Already Explain 555
55.1 Disciplinary Amnesia . . . 555
55.2 The Map . . . 556
55.3 Physics and Dynamics . . . 558
55.4 Computation, Sensing, and Control . . . 559
55.5 Biology, Representation, and Perspective . . . 559
55.6 Mind and Experience . . . 560
55.7 Using the Map . . . 560
55.8 The Joints . . . 561

56 What Would Metaxics Have to Add? 565
56.1 What Makes a Field . . . 565
56.2 A New Formalism . . . 566

56.3 A Transferable Theorem 567
56.4 A New Experimental Operation 567
56.5 A New Quantitative Prediction 568
56.6 A New Empirical Regularity 569
56.7 What a Synthetic Perspective Is For 570
56.8 The Verdict 570

XIII Research Programme **577**

57 Relational Intervention **579**
57.1 The Logic of Intervening on a Relation 579
57.2 A Catalogue 581
57.3 The Fish as the Model Case 582
57.4 Relational Interventions in Brains 584
57.5 What a Relational Intervention Shows 584

58 Reader Intervention **589**
58.1 The Operation 589
58.2 Replacement: Swapping the Reader 590
58.3 Lesion and Stimulation in the Fish 591
58.4 Modulation: Adaptation Changes the Reader 592
58.5 Learning: Rewriting the Reader 592
58.6 Reader Intervention in a Closed Loop 593
58.7 Why Reader Intervention Is the Distinctive Operation . 594

59 Closed-Loop Experiments **599**
59.1 Open and Closed Loops 599
59.2 A Toy Model of the Jamming Loop 600
59.3 Held and Hein's Carousel 602
59.4 Active Sensing Movements in the Fish 603
59.5 What the Loop Adds 604

60 Consciousness Experiments **609**
60.1 The Rule of the Null Model 609
60.2 Proposal 1: Entrainment Tongues Across Sedation . . . 611
60.3 Proposal 2: Intermodulation with Configural Controls . 612
60.4 Proposal 3: States that Separate Responsiveness from Experience 613
60.5 Proposal 4: Active Versus Passive Contingencies 615
60.6 Proposal 5: Relational Against Univariate, Everywhere . 615
60.7 What a Positive Result Would Show 616

61 Adversarial Testing **621**
61.1 Covariation Does Not Discriminate 621
61.2 Preregistration 622

61.3 Adversarial Collaboration . . . 623
61.4 The Global Workspace and Integrated Information Contest 623
61.5 Four Hypotheses and What Separates Them . . . 624
61.6 Risky Commitments and Noncommittal Rivals . . . 625
61.7 Rules for a Fair Contest . . . 627

62 What Would Falsify the Programme? 631
62.1 Falsification and Research Programmes . . . 631
62.2 Six Claims and How Each Could Fail . . . 632
62.3 How the Programme Could Cheat . . . 635
62.4 How the Central Idea Could Die . . . 636

XIV The Remaining Question 643

63 From Reading to Seeing 645
63.1 The Progression . . . 645
63.2 Grading the Arrows . . . 645
63.3 Where the Fish Stands . . . 648
63.4 What Would Count as Progress . . . 649
63.5 The Moiré, Revisited . . . 649

64 The First View from Somewhere 653
64.1 A Detector Is Not Yet a Reader . . . 653
64.2 A Reader Is Not Necessarily an Agent . . . 654
64.3 An Agent Is Not Necessarily a Representer . . . 654
64.4 A Representer Is Not Necessarily Self-Aware . . . 654
64.5 A Self-Model Is Not Obviously a Subject . . . 655
64.6 A Functional Perspective Is Not Obviously a Phenomenal One . . . 655
64.7 The Question . . . 655
64.8 What Metaxics Is, at the End . . . 656

Appendices 659

Appendices 661

A Mathematical Prerequisites 661
A.1 Linear Algebra . . . 661
A.2 Complex Numbers and Phasors . . . 662
A.3 Trigonometric Identities Used in the Book . . . 663
A.4 Calculus of Oscillations . . . 663
A.5 Probability . . . 664
A.6 Entropy and Mutual Information . . . 664
A.7 Notes and References . . . 665
A.8 Exercises . . . 665

B Fourier and Spectral Methods **667**
B.1 Conventions . . . 667
B.2 Rules . . . 667
B.3 Sums and Products of Sinusoids . . . 668
B.4 Sampling and Aliasing . . . 669
B.5 Finite Windows, Leakage, and Resolution . . . 669
B.6 Power Spectra . . . 670
B.7 The Analytic Signal . . . 670
B.8 Cross-Spectra and Coherence . . . 671
B.9 Higher-Order Spectra . . . 672
B.10 Notes and References . . . 672

C Nonlinear Systems **675**
C.1 Fixed Points and Their Stability . . . 675
C.2 Limit Cycles . . . 675
C.3 Elementary Bifurcations . . . 676
C.4 Center Manifolds and Slow Manifolds . . . 677
C.5 Nonlinear Input–Output Systems . . . 678
C.6 Notes and References . . . 679
C.7 Exercises . . . 679

D Phase Reduction **681**
D.1 Asymptotic Phase and Isochrons . . . 681
D.2 The Phase Response Curve . . . 681
D.3 A Worked Example . . . 682
D.4 Weak Coupling and Averaging . . . 683
D.5 Periodic Forcing . . . 684
D.6 When the Reduction Fails . . . 684
D.7 Notes and References . . . 685
D.8 Exercises . . . 685

E The Kuramoto Model **687**
E.1 The Model and Its Mean Field . . . 687
E.2 The Continuum Limit . . . 687
E.3 Stationary States . . . 688
E.4 The Self-Consistency Condition and the Threshold . . . 688
E.5 The Ott–Antonsen Reduction . . . 690
E.6 Finite Populations . . . 690
E.7 A Numerical Recipe . . . 691
E.8 Fitting the Model to Data . . . 691
E.9 Beyond the Basic Model . . . 692
E.10 What the Model Contains . . . 692
Notes and References . . . 693

F Arnold Tongues **695**

F.1 Circle Maps and Their Lifts . . . 695
F.2 The Rotation Number . . . 696
F.3 Tongue Boundaries . . . 697
F.4 Widths for Weak Coupling . . . 697
F.5 Farey Ordering . . . 698
F.6 The Critical Line and the Complete Staircase . . . 698
F.7 Beyond the Critical Line . . . 698
F.8 Measuring Tongues in Experiments . . . 699
F.9 A Numerical Recipe . . . 699
Notes and References . . . 700

G Hypergraphs and Higher-Order Interactions 701
G.1 Graphs and Their Limits . . . 701
G.2 Hypergraphs . . . 701
G.3 Simplicial Complexes . . . 702
G.4 Higher-Order Kuramoto Dynamics . . . 703
G.5 Where Higher-Order Interactions Come From . . . 704
G.6 When Pairwise Models Fail . . . 705
G.7 A Higher-Order Relation in the Fish . . . 705
G.8 Notes and References . . . 706
G.9 Exercises . . . 706

H Information Decomposition 709
H.1 Entropy and Mutual Information . . . 709
H.2 Interaction Information . . . 710
H.3 A Worked Example: Exclusive Or . . . 710
H.4 Partial Information Decomposition . . . 711
H.5 Synergy and Relations . . . 712
H.6 Transfer Entropy and Granger Causality . . . 712
H.7 Information in Time: The Anatomy of a Bit . . . 713
H.8 Integrated Information Decomposition . . . 713
H.9 Estimation . . . 714
H.10 Notes and References . . . 714

I Causal Inference and Interventions 715
I.1 Structural Causal Models . . . 715
I.2 Interventions on Relations . . . 716
I.3 The Interventionist Account . . . 717
I.4 Potential Outcomes . . . 718
I.5 Granger Causality and Transfer Entropy . . . 718
I.6 Causal Discovery and Feedback . . . 719
I.7 Notes and References . . . 720

J Experimental Surrogate Methods 721
J.1 The Logic of a Surrogate Test . . . 721

J.2 Shuffle Surrogates . . . 722
J.3 Phase-Randomized Surrogates . . . 722
J.4 Amplitude-Adjusted Surrogates . . . 723
J.5 Multivariate Surrogates . . . 723
J.6 Time-Shift and Trial-Shuffle Surrogates . . . 724
J.7 Relational Measures and Their Nulls . . . 724
J.8 Multiple Comparisons and Nonstationarity . . . 725
J.9 Notes and References . . . 725

K EEG and MEG Spectral Analysis 727
K.1 What the Sensors Record . . . 727
K.2 Estimating a Spectrum . . . 728
K.3 Periodic and Aperiodic Components . . . 728
K.4 Phase and the Analytic Signal . . . 729
K.5 Coupling Between Signals . . . 729
K.6 Cross-Frequency Coupling . . . 730
K.7 Steady-State Responses and Intermodulation . . . 730
K.8 Higher-Order Spectra . . . 731
K.9 Statistics Across Sensors, Frequencies, and Time . . . 731
K.10 Notes and References . . . 732

L Comparative Glossary 733
L.1 Terms . . . 733
L.2 Terms Declined . . . 738
L.3 Notes and References . . . 739

References 741

Index 777

Preface

Lay one fine grating over another and turn it by a degree. Broad dark bands appear, many times wider than the lines that make them, and they move a long way when the top grating moves a little. The bands are in neither grating. They belong to the configuration: to the small difference in angle, and to nothing else.

This book began with a question provoked by that experience: is consciousness a moiré effect? The question is better than it sounds. Experience is unified while the brain that supports it is distributed; experience has a slow "now" while neural activity is fast; perception is often settled by competition among representations rather than by any single sensory stream. A structure that exists only in the relation among many partially independent processes, slower and larger than any of them, has some of the right shape. It also has a fatal weakness. A moiré pattern explains how a new observable structure can arise from interacting structures. It does not explain why that structure should be experienced, and radio receivers are full of interference patterns that nobody supposes to be experienced.

The book that follows is what remained after that weakness was taken seriously.

Why "Metaxics"?

The Greek word μεταξύ (*metaxu*, often written *metaxy*) means "between." In Plato's *Symposium*, Diotima calls right opinion something between wisdom and ignorance, and calls Eros a spirit that is between the mortal and the immortal, carrying messages from each to the other (Plato 1989, 202a–203a). In Aristotle's *De Anima* the same word names the medium of perception: a colored object laid directly on the eye is not seen, because vision requires something between the object and the organ, and the transparent medium is what the color sets in motion (Aristotle 1984, II.7, 419a). Simone Weil used *metaxu* for the intermediaries through which a separated being is connected to what it is separated from, and gave the idea its sharpest form in the image of two prisoners who talk by knocking on the wall between their cells (Weil 2002, p. 145). Eric Voegelin made *metaxy* a name for the structure of human existence itself, lived in the tension between poles it never occupies (Voegelin 1974).

We borrow the word for a narrower purpose. *Metaxics*, in this book, names a research perspective on relational observables and on the physical systems that read them. The word marks a question. Whether it deserves to name a discipline is the subject of Part XII, and the answer given there is that it does not yet.

From Relations to Readers

An earlier plan for this book ran to forty chapters, from moiré patterns to consciousness, and treated "relations matter" as the thesis. That plan failed on inspection, and the way it failed shaped everything here.

Relationality is nearly universal. Isomers share their atoms and differ in arrangement; sentences share their words and differ in order; proteins share a sequence and differ in fold. Every science that deals with structure already deals with relations, and most of them have better tools for doing so than a new name could supply. Synchronization, coordination dynamics, synergetics, network science, information theory, and theoretical biology each own a large region of the territory the first plan claimed. A book that rebrands their results adds nothing and loses the precision those fields have earned.

The same inspection turned up a question that the established fields leave in an interesting state. Take the sum of two tones, $\cos \omega_1 t + \cos \omega_2 t$. It can be rewritten as a product, an envelope at half the difference frequency times a carrier at the mean frequency, and the envelope is plain to anyone who plots it. Yet the Fourier spectrum of the sum contains only ω_1 and ω_2. There is no component at the difference frequency until something nonlinear receives the signal: a square-law detector, a rectifying synapse, a cochlea. Square the sum and terms at $\omega_1 - \omega_2$, $\omega_1 + \omega_2$, $2\omega_1$, and $2\omega_2$ appear. A different receiver makes different variables explicit.

We state the result carefully, because it is easy to overstate:

> *Which relational variables become dynamically explicit depends jointly on the physical mixture and on the transduction performed by the receiving system.*

Nothing in that sentence says that relations exist only when observed. It says that a relation implicit in a mixture becomes a variable that can drive further dynamics only through a physical operation, and that different operations make different relations available. That operation is what this book calls a *reader*.

The question that organizes the book is therefore:

> *How does a relation become a variable for a system that participates in that relation?*

The weakly electric fish answers it more completely than any other system we know. *Eigenmannia*, a South American knifefish, emits a quasi-sinusoidal electric field hundreds of times a second and senses the field's distortions through electroreceptors in its skin. When a neighbor discharges at a nearby frequency, the two fields beat, and the fish shifts its own frequency away from its neighbor's: up if the neighbor is lower, down if the neighbor is higher (Watanabe and Takeda 1963; Bullock, Hamstra, et al. 1972a; Heiligenberg 1991a). To do this it must read the sign of the frequency difference. The amplitude of the mixture carries no trace of that sign, and its phase carries it only through small distortions; the sign is carried exactly by the joint behavior of the two, and the fish reads it from that relation without consulting a copy of its own motor command (Heiligenberg 1991a; Kawasaki 1993). The African *Gymnarchus*, whose electric sense evolved independently, uses the same computational rules with different brain regions (Kawasaki 1993; Kawasaki 1996; Kawasaki and Y. Guo 1996). The circuits have been mapped from receptor to pacemaker. The fish changes one of the relata in order to change the relation it is sensing, and in that loop a physical relation becomes a variable for the system that is one of its terms. The fish is this book's hydrogen atom: the simplest system rich enough to show the whole structure, returned to in every part.

The Method

One rule governs the book:

Subtract the established explanations before naming the residue.

A corollary follows for vocabulary. A new concept earns its name only when we can state a distinction, a prediction, or an operation that the existing vocabulary cannot express. The unification of several concepts under one feeling of connectedness does not earn a name. Where existing terms fit (reafference, active sensing, closed-loop control, semantic closure, implementation, intentionality) we use them.

The rule is enforced by the structure of every chapter. Each chapter closes with an *audit*, seven questions asked in the same order throughout:

1. *Canonical phenomenon*: the concrete system the chapter is about.
2. *Minimal model*: the smallest mathematics that captures it.
3. *What is relational here?*: which variable belongs to a configuration and to no single component.
4. *What physically reads it?*: which operation, if any, makes that variable dynamically explicit.
5. *What existing field explains it?*: the prior art.
6. *What remains unexplained?*: the residue.

7. *A discriminating experiment*: a test whose outcome would differ depending on the answer.

The last three questions carry the method: prior art, residue, test. Many audits end with a small residue, or with none. That is the expected outcome, and a chapter whose audit finds nothing new has done its job.

The formal environments follow one convention. *Theorems* follow exactly from their stated assumptions. *Propositions* follow from a model whose fit to the world is an open question. *Models* are explicit simplifications, useful for structure and prediction and not to be confused with descriptions of reality. *Conjectures* are claims we believe but have not proved or tested. *Hypotheses* are empirical claims stated so that an experiment could fail them. *Definitions* fix terminology and are stipulative. Numerical values are sourced, or else labeled as illustrative.

Two Ladders

The word "relation" has at least four meanings in what follows, and they are easy to confuse: a *geometric* relation, such as a relative angle; a *dynamical coupling*, such as the interaction term between two oscillators; an *informational dependence*, such as the mutual information between two signals; and a relation *for an agent*, a variable that some system reads and uses.

The word "reader" has at least four meanings too: a *detector*, whose output changes with some property of its input; a *decoder*, which maps an input to a variable that other processes can use; an *endogenous controller*, a reader built and maintained by the system whose behavior it regulates; and a *phenomenal subject*, for whom there is something it is like to read.

The book climbs both ladders at once. It is cumulative: each part adds one capability to the systems under study, from physical mixing through readout, implemented computation, active sensing, closed-loop control, biological closure, misrepresentation, world modeling, and self-modeling. At the top of both ladders stands the question the book began with, now in a form precise enough to be answered or to be shown unanswered:

Why is an endogenous reader ever accompanied by a point of view?

We do not answer it. The contribution of the book is to separate the steps that lead up to it, to show which of them existing science already explains, and to say exactly which step remains.

The Structure of This Book

Part I (*The Physics of Relations*) separates what belongs to components from what belongs to their configuration, using moiré patterns, beats, and the nonlinear operations that turn implicit relations into physical frequencies.

Part II (*Special Relations Are Not One Thing*) separates magnitude balance, resonance, and commensurability, three phenomena often run together as "the right relation," and uses magic-angle graphene to show the discipline that relational explanation requires. **Part III** (*Collective Variables*) covers synchronization, coordination dynamics, metastability, chimera states, synergetics, and causal emergence: the settings in which a relation becomes a variable with its own dynamics.

Part IV (*The Reader Problem*) introduces the reader, shows how different readers make different variables explicit, and confronts the claim that every physical system computes every computation. **Part V** (*The Electric Fish*) follows the jamming avoidance response from waveform to behavior, through the sign computation, reafference, the fish's errors, and the independent evolution of the same algorithm in two lineages.

Part VI (*From Reading to Meaning*) asks what makes a reading matter to the system that performs it, through Pattee's epistemic cut and semantic closure, autopoiesis, sense-making, theories of content, and misrepresentation. **Part VII** (*Detachment and World Models*) follows readings that concern what is absent: memory, prediction, counterfactuals, and generative models. **Part VIII** (*From World-Centered to Agent-Centered Variables*) studies here, now, and I as coordinates with an origin, from active sensing to the essential indexical. **Part IX** (*Self-Reading Systems*) covers metacognition, higher-order representation, self-models, and attention schemas.

Part X (*Consciousness*) states the difference between a functional and a phenomenal perspective, compares the major theories by asking each what extra property it claims, and reviews the neural evidence, including relational measures and the history of neural moiré theories. **Part XI** (*The Comparative Ladder*) sets radio, cell, fish, mammal, and human side by side and asks what changes at each step. **Part XII** (*Is There a Metaxic Principle?*) turns the method on the book's own name. **Part XIII** (*Research Programme*) proposes experiments that intervene on relations, on readers, and on loops, and states what results would falsify the programme. **Part XIV** (*The Remaining Question*) returns to the full progression and asks which arrows have been explained.

Twelve appendices supply the mathematics (Fourier analysis, nonlinear systems, phase reduction, the Kuramoto model, Arnold tongues, higher-order interactions, information decomposition, causal inference, surrogate data, and electrophysiological spectral analysis) and a comparative glossary.

Readers can take several paths. A short course on the book's argument reads Parts I, IV, V, VI, X, and XIV. A physics course reads Parts I to III with Appendices B to F. A course in philosophy of mind reads Parts IV and VI to XII, returning to Part V whenever the fish is invoked. The chapters assume calculus, linear algebra, and elementary probability; the appendices supply the rest.

This book is a starting point, written at the Piatra Institute and open

to revision as the questions it isolates are taken up by others. The reader is invited to test it, and to break it where it deserves breaking.

Piatra Institute, 2026

Notation

The following symbols and conventions are used throughout the text. Symbols local to a single chapter are defined where they appear. Where a standard notation from a neighboring field collides with ours (the Kuramoto order parameter r, the twist angle θ), we keep the standard notation and say so on first use.

The Reading Loop

Symbol	Name	Description
x_i	Relatum	A component whose own properties are given independently of the others
$r_{ij} = \mathcal{R}(x_i, x_j)$	Relation	A variable defined on two or more relata and on none alone
s_{A}, s_{E}	Agent and environment states	The states of a system and of what it is coupled to
$y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})$	Mixing	The physical superposition or interaction that delivers agent and environment together to a receiving surface
$z = \mathsf{D}(y)$	Reading	The physically implemented transduction, detection, or decoding that makes a variable of y dynamically explicit
$a = \pi(z)$	Policy	The map from the read variable to action
$\dot{s}_{\mathrm{A}} = \mathsf{F}(s_{\mathrm{A}}, a)$	Intervention	The dynamics by which action changes the agent (and, through it, the next mixture)
$\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$	Reading loop	The closed loop in which a reader participates in producing the relation it reads

Symbol	Name	Description
R, $\hat{R}$	World variable, estimate	A variable of the world and the system's internal estimate of it; misrepresentation is $\hat{R} \neq R$ with functional consequences

Signals, Spectra, and Oscillators

Symbol	Name	Description
ω, f	Angular and ordinary frequency	$\omega = 2\pi f$
$\Delta\omega = \omega_1 - \omega_2$	Detuning	Frequency difference between two oscillations
$\bar{\omega}$	Mean frequency	$(\omega_1 + \omega_2)/2$
$\hat{x}(\omega)$	Fourier transform	Spectrum of a signal $x(t)$; conventions in Appendix B
θ_i	Phase	Phase of oscillator i (Part III); twist angle between lattices in Chapters 2 and 8
$\phi = \theta_1 - \theta_2$	Relative phase	The collective variable of coordination dynamics
$r\,\mathrm{e}^{\mathrm{i}\psi}$	Kuramoto order parameter	$N^{-1}\sum_j \mathrm{e}^{\mathrm{i}\theta_j}$; $r \in [0,1]$ measures phase coherence
K	Coupling strength	Strength of mutual coupling among oscillators
ρ	Rotation number	Mean phase advance per cycle of a circle map
Q	Quality factor	Sharpness of a resonance, ω_0/γ
Π	Dimensionless group	A ratio of magnitudes invariant under change of units

The Electric Fish

Symbol	Name	Description
f_s, ω_s	Self frequency	Frequency of the fish's own electric organ discharge (EOD)
f_e, ω_e	External frequency	Frequency of the neighbor's EOD

Symbol	Name	Description
$\Delta f = f_{\mathrm{e}} - f_{\mathrm{s}}$	Frequency difference	Positive when the neighbor's frequency is higher; the jamming avoidance response moves f_{s} so as to increase $\lvert\Delta f\rvert$
A_{s}, A_{e}	Amplitudes	Local amplitudes of the self and external fields at a patch of skin
$m = A_{\mathrm{e}}/A_{\mathrm{s}}$	Contamination ratio	Relative strength of the external field at a patch of skin; it differs from patch to patch
$A(t)$, $\varphi(t)$	Amplitude and phase modulation	Envelope and phase of the local mixture, both modulated at $\lvert\Delta f\rvert$

Information, Inference, and Hypotheses

Symbol	Name	Description
$H(X)$	Entropy	Shannon entropy of a random variable
$I(X;Y)$	Mutual information	Information shared by X and Y
$\mathbb{E}[\cdot]$, $\mathbb{P}(\cdot)$	Expectation, probability	
$\langle\cdot\rangle$	Average	Time or ensemble average, as stated
$\mathrm{do}(X{=}x)$	Intervention	Pearl's operator for setting a variable from outside its usual causes
H_{X}	Hypothesis	A named hypothesis in the research programme, e.g. H_{GNW}, H_{IIT}

Conventions

- Sans-serif capitals (M, D, F) denote the operators of the reading loop.
- Calligraphic letters ($\mathcal{R}$, $\mathcal{L}$) denote relations, loops, and structured sets.
- Bold lowercase ($\boldsymbol{x}$) denotes vectors; bold uppercase ($\boldsymbol{A}$) denotes matrices.
- i is the imaginary unit and e is Euler's number; d is a differential.
- Time dependence is written $x(t)$; discrete time steps are written x_n.
- Physical units follow the SI and are typeset with their values, as in $400\,\mathrm{ns}$.

Part I

The Physics of Relations

What physical operation exposes the pattern?

1

Objects, Relations, and Relational Observables

> As for my own opinion, I have said more than once, that I hold space to be something merely relative, as time is; that I hold it to be an order of coexistences, as time is an order of successions.
>
> Gottfried Wilhelm Leibniz, third paper to Samuel Clarke (1716)

A physical description usually begins with a list of things and their properties: this particle is here, that oscillator has this phase, this grating has that period. Many of the variables that do the explanatory work belong to none of the things on the list. The distance between two particles, the phase difference between two oscillators, and the angle between two gratings are defined on pairs, and they change when either member of the pair changes. We call such variables *relational observables*.

The thesis of this chapter has three parts. A relational observable is a variable of a configuration that is a function of no single component. Relational observables are forced on us whenever a system has no privileged reference frame, and they earn their keep when they compress, predict, and respond to intervention. And relationality by itself is far too common to found a discipline. The question the rest of the book pursues is narrower: which relational variables become available to which physical processes, and what happens when the process that reads a relation is itself one of its terms.

1.1 Absolute and Relational Descriptions

Five pairs of variables fix the distinction.

- *Position* and *displacement*. The position of a particle is its coordinate in some frame; the displacement between two particles is the difference of

their coordinates, and it is the same in every frame related to the first by a translation.

- *Phase* and *phase difference.* The phase of an oscillator is its position on its cycle measured from an arbitrary origin; the phase difference between two oscillators does not depend on where that origin is placed.
- *Orientation* and *relative orientation.* The angle of a grating is measured from an axis someone chose; the angle between two gratings is measured from neither.
- *Frequency* and *detuning.* Each tone has a frequency; the detuning $\Delta\omega = \omega_1 - \omega_2$ belongs to the pair.
- *Scale* and *scale ratio.* A length is a number of units; a ratio of two lengths is a pure number, the same in every system of units.

The first member of each pair is "absolute" only in a qualified sense. It is relational too, but one of its relata is a reference (an origin, an axis, a unit) that lies outside the system under study and is fixed by convention. The second member of each pair is relational among the system's own components. The distinction that matters is therefore between relations to an external reference and relations internal to the system, and the difference becomes consequential when the system itself has no access to the external reference.

The idea is old. Leibniz held space and time to be orders of coexistence and succession rather than containers (Alexander 1956, third paper, § 4); Mach's criticism of absolute space, and its later formalization, showed that dynamics can be written entirely in terms of relative configurations (Barbour and Bertotti 1982); Cassirer argued that the history of physics is a history of substance concepts giving way to relational concepts (Cassirer 1923); and structural realism holds that what our best theories describe is structure rather than the intrinsic nature of relata (Ladyman and Ross 2007). None of that philosophy is needed here, and we take no side in it. The claim of this chapter is operational: for many systems the variables that enter the laws, and the variables available to parts of the system, are relational ones.

Definition 1.1 (Relational observable)**.** Let a system consist of components with states $x_1, \dots, x_n$ taking values in spaces $X_1, \dots, X_n$. A *relational observable* is a function

$$r = \mathcal{R}(x_{i_1}, \dots, x_{i_k}), \qquad k \geq 2,$$

that cannot be written as a function of the state of any single component. We call the components $x_{i_1}, \dots, x_{i_k}$ its *relata* and the joint state $(x_1, \dots, x_n)$ the *configuration.*

The definition is deliberately thin. It does not say that relational observables are special, deep, or emergent; it says only that they are defined on more than one component and are not secretly properties of one of them. Displacement, phase difference, relative orientation, detuning, and scale ratio all qualify. So does the sum $x_1 + x_2$ of two positions, which is relational in this sense and

physically uninteresting in most settings. The criteria that separate useful relational observables from arbitrary ones come in Section 1.5.

Example 1.2 (Huygens's clocks). In 1665 Christiaan Huygens noticed that two pendulum clocks hung from a common beam fell into step, swinging in opposition, and returned to that state when disturbed (Pikovsky et al. 2001, ch. 1). The variable that settled was neither clock's phase. Each phase kept advancing; what came to rest was their difference. The beam coupled the clocks, and the coupled dynamics had an attractor in the relative phase and none in the absolute phases. Part III builds the theory of such variables, from the Kuramoto model (Chapter 9) to the relative-phase dynamics of coordination (Chapter 10).

1.2 Relational Coordinates

A configuration of n point particles in $\mathbb{R}^d$ is specified by nd numbers. Many problems are simpler in coordinates built from relations among the particles.

Definition 1.3 (Relational coordinates)**.** Given component states $x_1, \dots, x_n$, a set of *relational coordinates* is a collection of relational observables $r_{ij} = \mathcal{R}(x_i, x_j)$ (and, where needed, observables on larger sets of components) from which the configuration can be reconstructed up to the action of a stated group of transformations.

The standard example is the reduction of the two-body problem. Two particles with masses m_1, m_2 interacting through a potential that depends only on their separation obey

$$m_1 \ddot{\boldsymbol{x}}_1 = -\nabla V(\boldsymbol{x}_1 - \boldsymbol{x}_2), \qquad m_2 \ddot{\boldsymbol{x}}_2 = +\nabla V(\boldsymbol{x}_1 - \boldsymbol{x}_2). \tag{1.1}$$

In the coordinates $\boldsymbol{X} = (m_1\boldsymbol{x}_1 + m_2\boldsymbol{x}_2)/(m_1 + m_2)$ and $\boldsymbol{r} = \boldsymbol{x}_1 - \boldsymbol{x}_2$ the system separates:

$$\ddot{\boldsymbol{X}} = 0, \qquad \mu\, \ddot{\boldsymbol{r}} = -\nabla V(\boldsymbol{r}), \qquad \mu = \frac{m_1 m_2}{m_1 + m_2}. \tag{1.2}$$

The center of mass drifts uniformly and carries no information about the interaction. Everything the interaction does is written in the relational coordinate $\boldsymbol{r}$. The relational variable is the one whose dynamics *closes*: its future depends only on its own present value and velocity.

Counting relational coordinates shows how much of a configuration is fixed by convention. For $n \geq d$ points in general position in $\mathbb{R}^d$, the rigid motions of space form a group of dimension $d + d(d-1)/2$, and the configuration modulo rigid motions has

$$nd - d - \tfrac{1}{2}d(d-1) \tag{1.3}$$

degrees of freedom: $2n - 3$ in the plane and $3n - 6$ in space, the familiar count of internal coordinates of a nonlinear molecule. The subtracted degrees of freedom are exactly the relations to an external frame.

1.3 Invariance

When the laws of a system do not change under some transformation of all its components at once, the transformation's parameters cannot be read off from anything inside the system. Translations, rotations, and shifts of a common phase origin are the standard cases.

Definition 1.4 (Frame-independent observable)**.** Let a group G act on each component space X_i, and on configurations componentwise: $g \cdot (x_1, \ldots, x_n) = (gx_1, \ldots, gx_n)$. An observable $\mathcal{O}$ is *frame-independent* (with respect to G) if $\mathcal{O}(g \cdot x) = \mathcal{O}(x)$ for every $g \in G$ and every configuration x.

The following result explains why frame-independent quantities are relational.

Theorem 1.5 (Invariance forces relationality)**.** *Suppose G acts transitively on each component space X_i. Then every non-constant frame-independent observable is a relational observable.*

Proof. Suppose $\mathcal{O}$ is frame-independent and can be written as $\mathcal{O}(x) = g_i(x_i)$ for some single component i and some function g_i. For any two states $x_i, x_i' \in X_i$, transitivity gives some $h \in G$ with $x_i' = hx_i$, and frame independence gives $g_i(x_i') = g_i(hx_i) = g_i(x_i)$. So g_i is constant, and so is $\mathcal{O}$. A non-constant frame-independent observable therefore cannot be a function of any single component. □

The theorem is elementary and the hypothesis matters. Translations act transitively on $\mathbb{R}^d$, rotations of a common phase origin act transitively on the circle, and rescalings act transitively on the positive reals, so in each of these settings the only non-trivial quantities that do not depend on a convention are relations among the components. The converse fails: the sum $x_1 + x_2$ is relational and changes under translation.

Example 1.6 (Only phase differences are observable). The Kuramoto model of N coupled phase oscillators,

$$\dot{\theta}_i = \omega_i + \frac{K}{N} \sum_{j=1}^{N} \sin(\theta_j - \theta_i), \tag{1.4}$$

is unchanged when every phase is shifted by the same constant, because the coupling depends only on differences $\theta_j - \theta_i$ (Kuramoto 1984; Strogatz 2000). By Theorem 1.5, the only non-trivial frame-independent observables are functions of phase differences. The model has N phases and $N-1$ independent phase differences; the missing degree of freedom is the choice of origin on the circle, and no experiment inside the system can recover it.

The same fact has a consequence for any physical system that must *use* a relation rather than merely exhibit one. A reader with no access to an external reference can respond only to frame-independent variables, and by the theorem those are relational. Chapter 21 shows an animal in exactly this position. The weakly electric fish has no internal clock against which to measure the absolute phase of the field on its skin, and it determines the sign of the frequency difference between itself and a neighbor by comparing phase *between* regions of its own body (Heiligenberg 1991a; C. Carr et al. 1986).

1.4 Pairwise and Higher-Order Relations

Relations need not be pairwise. A relation $\mathcal{R}(x_i, x_j, x_k)$ on three components can be irreducible, in the sense that no combination of the pairwise relations determines it. Network science represents pairwise relations as edges; higher-order relations need hyperedges (sets of any size) or simplicial complexes (hyperedges closed under taking subsets), and dynamics on such structures can differ qualitatively from dynamics on graphs (Battiston et al. 2020; Bick et al. 2023). Appendix G collects the definitions.

The cleanest example is informational.

Example 1.7 (An irreducibly three-way relation). Let X and Y be independent fair bits and let $Z = X \oplus Y$ (exclusive or). Every pair is independent: $I(X;Y) = I(X;Z) = I(Y;Z) = 0$. Yet the three variables are jointly constrained, since any two determine the third: $I(X,Y;Z) = 1$ bit. Nothing in the pairwise statistics reveals the relation, and a reader that monitors pairs of variables cannot detect it. The quantity McGill called interaction information registers exactly this kind of dependence (McGill 1954), and partial information decomposition names it *synergy* (Williams and Beer 2010). Appendix H develops the measures and their disputes.

The example carries a lesson for the rest of the book. Whether a relation is detectable depends on what the detector can combine. A reader that sees only pairs is blind to the exclusive or; a reader that sees all three variables at once is not. The same pattern returns in Chapter 3, where a linear filter cannot find the difference frequency in a sum of two tones and a square-law detector can, and in Chapter 21, where the sign of a frequency difference is carried only by the joint behavior of two modulations.

1.5 When Is a Relational Variable Useful?

Relational observables are cheap: any function of two components is one. Five criteria separate the ones that belong in a scientific description.

Principle 1.8 (Criteria for a useful relational variable)**.** A relational variable $r = \mathcal{R}(x_{i_1}, \ldots, x_{i_k})$ earns a place in a description of a system to the extent that it satisfies the following.

1. *Compression*: the system's behavior can be described with fewer variables when r is among them.
2. *Prediction*: r predicts future states at least as well as the relata do, so that, for the purposes at hand, knowing r screens off the finer detail.
3. *Causal relevance*: interventions that change r change outcomes of interest (Woodward 2003).
4. *Intervention sensitivity*: r can be manipulated while the relata's own properties are held as fixed as physically possible, so that its causal role can be tested separately from theirs.
5. *Robustness*: r keeps its role under perturbation, coarse-graining, and moderate changes in the underlying components.

The relative coordinate of the two-body problem satisfies all five, which is why it appears in every mechanics textbook. Relative phase in bimanual coordination satisfies them as well: it compresses the joint motion of two limbs to one variable, it has its own dynamics, perturbing it produces predictable relaxation, and its transition from anti-phase to in-phase coordination survives changes of limb, speed, and subject (Chapter 10). The average distance between randomly chosen pairs of atoms in a table satisfies none of them.

The fourth criterion deserves emphasis because it is the one most often skipped. To show that a relation, and no property of its relata, does causal work, one must change the relation while holding the relata fixed. Sometimes physics makes this easy: the twist angle between two graphene sheets can be set without changing either sheet (Chapter 8). Sometimes it is hard: one cannot change the phase difference between two neural populations without touching at least one of them. Part XIII is organized around this criterion (Chapters 57 and 58).

1.6 Same Parts, Different Relations

Ethanol and dimethyl ether are built from the same atoms, two carbons, six hydrogens, and one oxygen. They differ in how the atoms are bonded, and the difference is enough to make one a liquid and the other a gas at room temperature. "Dog bites man" and "man bites dog" share their words and differ in order. A protein's properties depend on how its chain folds, and a network's on which of its nodes are joined. Examples of the pattern "same components, different relations, different whole" can be multiplied without limit, because the pattern is close to universal.

This is the first caution of the book, and it shapes everything that follows. If relationality were enough to define a field, chemistry, linguistics, structural biology, and network science would all be branches of it, and a new name

would add nothing to any of them. Every science that studies structure already studies relations, usually with better tools than a general theory of relations could supply. The Preface records how an earlier plan for this book foundered on exactly this point.

Something narrower survives the caution. In each example above, the relation is a fact about the configuration, available to an outside observer who can see the whole. The question that no structural science takes as its subject is how a relation becomes available to a physical process *inside* the configuration, a process that responds to the relation and not merely to its relata. The two-body relative coordinate is available to the equations of motion; is it available to either body? The phase difference between Huygens's clocks is available to the beam that couples them; is it available to anything that could use it? These questions have physical answers, and they differ from case to case.

1.7 From Relation to Reader

The questions of the previous section organize the book. We state them informally here; Chapter 15 makes them precise.

A system and its environment, with states s_{A} and s_{E}, are delivered together to some receiving surface by a physical mixing process M:

$$y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}}).$$

The mixture y implicitly contains relations between agent and environment. A physical transduction D makes some of them explicit, as variables that can drive further dynamics:

$$z = \mathsf{D}(y).$$

If the system acts on z, through a policy $a = \pi(z)$, and its action changes its own state, $\dot{s}_{\mathrm{A}} = \mathsf{F}(s_{\mathrm{A}}, a)$, then the next mixture depends on what was read. We call the closed structure (Figure 1.1) a *reading loop*, $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$.

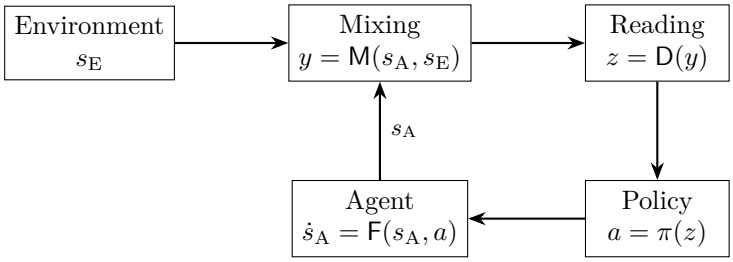


Figure 1.1: The reading loop. A reader inside the relation it reads: what the agent does changes the next mixture. Formal treatment in Chapter 15.

A loop of this shape is a reader inside the relation that the reader subsequently modifies. Most of its parts have established names: sensing and

transduction, reafference, active sensing, closed-loop control, and, when the loop maintains the system that contains it, semantic closure. The book uses those names. The weakly electric fish of Part V instantiates every part of the loop and has had each part mapped: its own discharge is one relatum, a neighbor's is the other, the beat is the mixture, midbrain neurons read the sign of the frequency difference, and the fish changes its own frequency to change the relation it senses (Heiligenberg 1991a).

1.8 Audit

Canonical phenomenon

Two pendulum clocks on a common beam settle into anti-phase: each phase keeps advancing while their difference comes to rest.

Minimal model

Relational observables $r = \mathcal{R}(x_{i_1}, \ldots, x_{i_k})$ (Definition 1.1); separation of the two-body problem into center-of-mass and relative coordinates (1.2); the invariance theorem (Theorem 1.5).

What is relational here?

Displacement, phase difference, relative orientation, detuning, and scale ratio; in the clock example, the relative phase $\theta_1 - \theta_2$.

What physically reads it?

In the clock example, the beam: it couples the clocks through a force that depends on their relative motion, so the relative phase enters the dynamics. Nothing reads it in any richer sense. The chapter's examples are relations available to equations of motion or to an outside observer, which is exactly why they are the starting point and not the subject.

What existing field explains it?

Classical mechanics (relative coordinates and reduction), the relational tradition in the philosophy of space and time (Alexander 1956; Barbour and Bertotti 1982), symmetry and invariance, synchronization theory (Pikovsky et al. 2001), network science and higher-order interactions (Battiston et al. 2020), and information theory (McGill 1954; Williams and Beer 2010). Together they explain everything in this chapter.

What remains unexplained?

Nothing in the physics. The residue is a question: how a relation implicit in a configuration becomes available to a process that is one of the configuration's terms. The chapter states it; the rest of the book pursues it.

A discriminating experiment

For a proposed relational variable in any system, perform the manipulation of Principle 1.8(4): change the relation while holding the relata as fixed as

possible (for coupled oscillators, drive the phase difference through a delay while clamping each oscillator's frequency). If outcomes follow the relation, the variable is causally relevant; if they follow the relata, the relational description is a convenient redescription and nothing more.

1.9 Notes and References

The relational view of space and time runs from the Leibniz–Clarke correspondence (Alexander 1956) through Mach to the modern formulation of relational dynamics by Barbour and Bertotti (1982). Cassirer (1923) gives the classic philosophical argument that science moves from substance concepts to function and relation concepts; Ladyman and Ross (2007) develop structural realism in its contemporary form. In quantum theory, Rovelli (1996) proposes that the states of physical systems are relative to other systems; that proposal concerns the interpretation of quantum mechanics and makes claims far stronger than anything in this book, and we mention it only so that the two are not confused.

Relative coordinates and the reduction of symmetric systems are standard in classical mechanics; Landau and Lifshitz (1976) is a compact source. Anderson (1972) is the classic statement that new levels of description bring new variables without any failure of reduction. On synchronization and Huygens's clocks, see Pikovsky et al. (2001); on the Kuramoto model, Kuramoto (1984) and Strogatz (2000). Higher-order interactions are surveyed by Battiston et al. (2020) and Bick et al. (2023); the information-theoretic treatment of irreducible dependence begins with McGill (1954) and continues in the partial information decomposition of Williams and Beer (2010). The interventionist criterion of causal relevance is developed by Woodward (2003).

1.10 Exercises

Exercise 1.1. Derive (1.2) from (1.1). Show that the kinetic energy splits as $\tfrac{1}{2}(m_1+m_2)|\dot{\boldsymbol{X}}|^2 + \tfrac{1}{2}\mu|\dot{\boldsymbol{r}}|^2$. Which of the five criteria of Principle 1.8 does $\boldsymbol{r}$ satisfy, and why?

Exercise 1.2. Verify the count (1.3) for $n = 3$ points in the plane by exhibiting three relational coordinates that determine a triangle up to rigid motion. Which additional relational coordinate is needed if reflections are not allowed, and what kind of relation (pairwise or higher-order) is it?

Exercise 1.3. Show that (1.4) is invariant under $\theta_i \mapsto \theta_i + c$ for all i. Rewrite the model for $N = 2$ as a single equation for $\phi = \theta_1 - \theta_2$, and find the condition on K and $\omega_1 - \omega_2$ under which ϕ has a fixed point.

Exercise 1.4. For $Z = X \oplus Y$ with X, Y independent fair bits, compute $I(X;Y)$, $I(X;Z)$, $I(Y;Z)$, $I(X,Y;Z)$, and $I(X;Y \mid Z)$. Then let X and Y be biased,

$\mathbb{P}(X{=}1) = \mathbb{P}(Y{=}1) = p$, and plot $I(X;Z)$ and $I(X,Y;Z)$ against p. At which values of p is the relation purely three-way?

Exercise 1.5. A patch of skin receives a signal of the form $\cos\omega_{\mathrm{s}}t + m\cos(\omega_{\mathrm{e}}t + \psi_0)$ with $0 < m < 1$. Show that its envelope is $A(t) = \sqrt{1 + m^2 + 2m\cos(\Delta\omega\, t + \psi_0)}$, with $\Delta\omega = \omega_{\mathrm{e}} - \omega_{\mathrm{s}}$. Show that the envelope is unchanged under $(\Delta\omega, \psi_0) \mapsto (-\Delta\omega, -\psi_0)$, and conclude that no reader of the envelope alone can tell whether the second frequency is above or below the first. (Chapter 21 shows how the electric fish solves this problem.)

Exercise 1.6. Construct, for a system of your choice, one relational variable that satisfies all five criteria of Principle 1.8 and one that is well defined but fails the third. For each, describe the intervention that would demonstrate its status.

Exercise 1.7. Take any relation from Section 1.6 and ask whether any process inside the configuration responds to it. If one does, describe it in the notation of Figure 1.1: what is M, what is D, and does anything close the loop?

2

Moiré: The Prototype of Relational Structure

Lay a fine grating over a copy of itself, turn the top one by a few degrees, and broad bands appear that are many times wider than the lines that make them. Slide the top grating by a fraction of a line spacing and the bands sweep across the whole field. The bands are the moiré pattern, and they are the cleanest visible case of a structure that belongs to a configuration and to neither of its parts.

This chapter makes that statement exact and then limits it. A moiré fringe is a low-frequency term created when two fine periodic structures are *multiplied*; its period and orientation are functions of the relative period and relative angle of the layers and of nothing else; and it becomes visible because the observer, or the camera, discards everything fine. Each of these three facts matters for the rest of the book. The first says that the physical operation combining the layers decides whether the relation is present as a spatial frequency at all. The second says what belongs to the configuration. The third says that even a relation physically present in a mixture needs a reader with the right properties to become an observable. The chapter ends by closing the analogy between moiré and mind before the analogy can do damage.

2.1 Two Gratings

Model a transparency ruled with parallel lines as a transmittance that varies sinusoidally across the lines:

$$t_i(x) = \tfrac{1}{2}\left[1 + \cos(k_i x)\right], \qquad k_i = \frac{2\pi}{p_i}, \tag{2.1}$$

where p_i is the period of grating i and k_i its spatial frequency. Real rulings are closer to square waves than to sinusoids; their extra harmonics add higher-order fringes without changing anything in the argument below (Amidror 2009).

Light passing through both transparencies is attenuated by each in turn, so the transmitted intensity is proportional to the *product* of the transmittances:

$$\begin{aligned} t_1(x)\,t_2(x) = \tfrac{1}{4}\Big[1 + \cos(k_1 x) + \cos(k_2 x) \\ + \tfrac{1}{2}\cos\big((k_1 - k_2)x\big) + \tfrac{1}{2}\cos\big((k_1 + k_2)x\big)\Big]. \end{aligned} \tag{2.2}$$

The product contains a term that neither factor contains, at the difference frequency $k_1 - k_2$. When the periods are close, that frequency is small, and its period is long.

Definition 2.1 (Moiré period)**.** For two parallel gratings with periods $p_1 \neq p_2$, the *moiré period* is the period of the difference-frequency term of their product,

$$p_M = \frac{2\pi}{|k_1 - k_2|} = \frac{p_1\,p_2}{|p_1 - p_2|}. \tag{2.3}$$

Two gratings with periods 1.00 mm and 1.01 mm, to take illustrative values, produce fringes 101 mm apart. A mismatch of one percent becomes a pattern a hundred times coarser than either grating.

Whether an observer sees the fringes depends on what the observer's optics do with (2.2). An eye or a camera at a distance averages the transmitted light over a patch much larger than p_1 and p_2 and much smaller than p_M. That averaging is a low-pass filter. It removes the terms at k_1, k_2, and $k_1 + k_2$ and keeps the constant and the difference term:

$$\big\langle t_1 t_2 \big\rangle_{\text{low-pass}}(x) = \tfrac{1}{4} + \tfrac{1}{8}\cos\big((k_1 - k_2)x\big). \tag{2.4}$$

The fringes are what survives the reader's blurring.

The same result has a geometric form due to Oster, Wasserman, et al. (1964). Index the lines of the first grating by integers j and those of the second by integers l, so that the lines are the level curves $\Phi_1(x, y) = j$ and $\Phi_2(x, y) = l$. The moiré fringes are the level curves of the index difference, $\Phi_1 - \Phi_2 = \text{const}$: the loci where the two families are in step. The indicial description works for curved families as well as straight ones and is the easiest way to predict the shape of fringes from arbitrary line patterns; the spectral description (2.2) is the easiest way to predict their contrast and to see why they appear at all (Amidror 2009).

2.2 Translation Moiré

Keep the two gratings parallel and slide the second by a distance δ across its lines. Its transmittance becomes $t_2(x - \delta)$, and the difference term of (2.2) becomes

$$\tfrac{1}{8}\cos\big((k_1 - k_2)x + k_2\delta\big). \tag{2.5}$$

The phase of the fringe pattern advances by $k_2\delta$, so each fringe moves by

$$\Delta x_M = \frac{k_2\,\delta}{|k_1 - k_2|} = \frac{p_1}{|p_1 - p_2|}\,\delta. \tag{2.6}$$

The ratio $M = p_1/|p_1 - p_2|$ is a magnification. With the illustrative periods above, $M = 100$: a shift of the grating by 10 µm moves the fringes by a millimeter. The magnification is again a function of the configuration alone. Neither grating has it.

2.3 Rotational Moiré

The more familiar case is two identical gratings of period a, one rotated by a small angle θ relative to the other. Represent each grating by its wavevector $\boldsymbol{k}_i$, of length $2\pi/a$ and perpendicular to its lines. The product of the transmittances contains a term $\cos\big((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{x}\big)$, and since $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$ have equal length and differ by the angle θ,

$$|\boldsymbol{k}_1 - \boldsymbol{k}_2| = \frac{2\pi}{a}\,2\sin\frac{\theta}{2}.$$

The fringe spacing is therefore

$$L_M = \frac{a}{2\sin(\theta/2)} \approx \frac{a}{\theta} \quad (\theta \ll 1). \tag{2.7}$$

The fringes are perpendicular to $\boldsymbol{k}_1 - \boldsymbol{k}_2$, which for equal wavevectors is perpendicular to their bisector; at small angles the fringes therefore run nearly at right angles to the lines that produce them, which is the first thing one notices on looking through two layers of fine mesh.

The general case, gratings of different periods p_1, p_2 at relative angle θ, follows from the same vector difference:

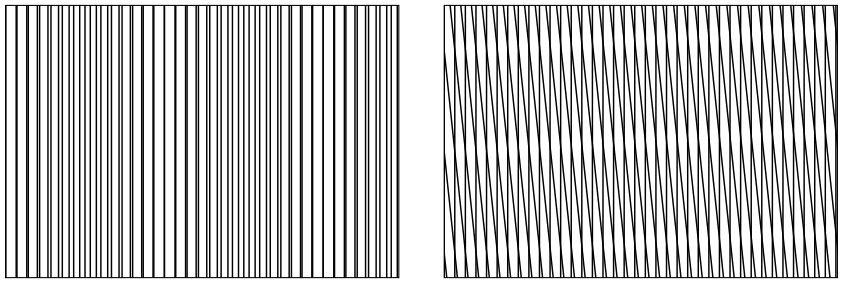
$$L_M = \frac{p_1\,p_2}{\sqrt{p_1^2 + p_2^2 - 2p_1p_2\cos\theta}}, \tag{2.8}$$

which reduces to (2.3) at $\theta = 0$ and to (2.7) at $p_1 = p_2$. Figure 2.1 shows both limits.

2.4 Amplification

Equation (2.7) converts a small angle into a large length: the fringe spacing is about $1/\theta$ times the period. The displacement magnification of Section 2.2 has a rotational counterpart. Translating one of two identical gratings by δ across its lines shifts the phase of the difference term by $(2\pi/a)\,\delta$, and the fringes move by

$$\Delta x_M = \frac{\delta}{2\sin(\theta/2)} \approx \frac{\delta}{\theta}. \tag{2.9}$$



(a) periods 0.14 and 0.15 cm (b) equal periods, rotated by 6°

Figure 2.1: Moiré fringes from two line gratings. (a) Translation moiré: fringes spaced $p_1 p_2/|p_1 - p_2| = 2.1\,\text{cm}$, parallel to the lines. (b) Rotational moiré: fringes spaced $a/(2\sin 3°) \approx 1.3\,\text{cm}$, nearly perpendicular to the lines. In both panels the fringes are properties of the pair.

At $\theta = 0.01$ radians, a displacement of one micrometer moves the fringes by a tenth of a millimeter.

Both amplifications have the same structure. A small mismatch between two nearly identical layers defines a new length scale, long compared with either layer's own scale, and a small relative motion becomes a large motion of that new structure. The size of the amplification is set by how *nearly* the layers match: the closer the match, the larger the moiré scale and the larger the magnification.

The same arithmetic operates far below the scale of transparencies.

Example 2.2 (Twisted bilayer graphene). Graphene is a single sheet of carbon atoms in a honeycomb lattice with a lattice constant of about 0.246 nm, $\sqrt{3}$ times the carbon–carbon distance of about 0.142 nm (Castro Neto et al. 2009). Two sheets stacked with a twist of 1.1° produce, by (2.7), a moiré superlattice with a period of about 12.8 nm, some fifty times the atomic spacing. In such superlattices the moiré is read by the electrons themselves, which move in the long-period potential it creates (Bistritzer and MacDonald 2011); in graphene on hexagonal boron nitride the same kind of superlattice produces the fractal electronic spectrum known as Hofstadter's butterfly (Dean et al. 2013).

Chapter 8 follows what happens when that relationally generated length scale reaches a size at which the electrons' kinetic energy becomes small; for now the point is only that the moiré length is set by the twist, a relation, and by the lattice constant, a property of each layer, and by nothing else.

2.5 What Belongs to the Layers?

Each layer has properties that it carries whether or not the other layer is present: its period, the orientation of its lines relative to whatever frame one uses to describe it, the profile of each line (sinusoidal, square, blurred), its

contrast, and its defects. These properties determine the layer's own spectrum, the set of spatial frequencies in (2.1) and its harmonics.

Two consequences follow. The line profile of each layer determines which higher-order moirés can appear, since a square-wave grating contributes harmonics at $2k_i, 3k_i, \ldots$, and the product contains differences such as $2k_1 - k_2$ as well as $k_1 - k_2$. The contrast of each layer multiplies the contrast of the fringes: in (2.2) the difference term is the product of the two layers' modulation amplitudes. A layer with no modulation produces no moiré with anything. So the fringe *contrast* and *profile* depend on the layers.

2.6 What Belongs to the Configuration?

The fringe *geometry* depends on the layers only through their relation. The moiré wavevector is $\boldsymbol{k}_M = \boldsymbol{k}_1 - \boldsymbol{k}_2$, and three facts about it can be checked directly from that formula.

Proposition 2.3 (The moiré geometry is relational)**.** *Let two gratings have wavevectors $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$.*

1. *Rotating both gratings by the same rotation R rotates $\boldsymbol{k}_M$ by R and leaves the fringe spacing $2\pi/|\boldsymbol{k}_M|$ unchanged.*
2. *Translating both gratings by the same vector $\boldsymbol{d}$ translates the fringes by $\boldsymbol{d}$.*
3. *Translating one grating by $\boldsymbol{d}$ relative to the other shifts the fringe phase by $\boldsymbol{k}_2 \cdot \boldsymbol{d}$, which moves the fringes by a distance amplified by the ratio of the grating frequency to the moiré frequency.*

Proof. (1) $\mathsf{R}\boldsymbol{k}_1 - \mathsf{R}\boldsymbol{k}_2 = \mathsf{R}(\boldsymbol{k}_1 - \boldsymbol{k}_2)$ and rotations preserve length. (2) Translating both layers by $\boldsymbol{d}$ replaces $\boldsymbol{x}$ by $\boldsymbol{x} - \boldsymbol{d}$ in every term of the product, including the difference term. (3) Replacing $t_2(\boldsymbol{x})$ by $t_2(\boldsymbol{x} - \boldsymbol{d})$ adds the phase $\boldsymbol{k}_2 \cdot \boldsymbol{d}$ to the difference term; dividing by $|\boldsymbol{k}_M|$ gives the displacement, as in (2.6) and (2.9). □

Parts (1) and (2) say that the fringe pattern moves rigidly with the pair when the pair moves rigidly: its spacing is a frame-independent quantity in the sense of Chapter 1. Part (3) says that relative motion of the layers is what moves the fringes relative to the layers. The spacing, orientation, and phase of the moiré are therefore determined by the relative period, relative angle, and relative displacement of the layers. These are relational observables in the strict sense of Chapter 1, and the moiré is a device that makes them large and visible.

2.7 Multiplicative Versus Additive Superposition

Everything so far depended on the layers being *multiplied*. Light passing through two transparencies is attenuated twice, and attenuation multiplies.

Other ways of combining two periodic patterns add them instead. Two slide projectors throwing gratings onto the same screen with incoherent light produce an intensity that is the *sum* of the two patterns; two sound sources produce a pressure that is the sum of the two pressures. The difference matters, and the Fourier transform shows why.

Multiplication in space corresponds to convolution of spectra (Appendix B). The spectrum of each sinusoidal grating consists of an impulse at zero and impulses at $\pm\boldsymbol{k}_i$; convolving two such spectra places impulses at every sum and difference, $\pm\boldsymbol{k}_1 \pm \boldsymbol{k}_2$, and the impulses at $\pm(\boldsymbol{k}_1 - \boldsymbol{k}_2)$ lie close to the origin when the gratings nearly match. That low-frequency pair is the moiré (Amidror 2009). Addition in space corresponds to addition of spectra. The sum of two gratings has impulses at 0, $\pm\boldsymbol{k}_1$, and $\pm\boldsymbol{k}_2$, and nowhere else. There is no component at the difference frequency.

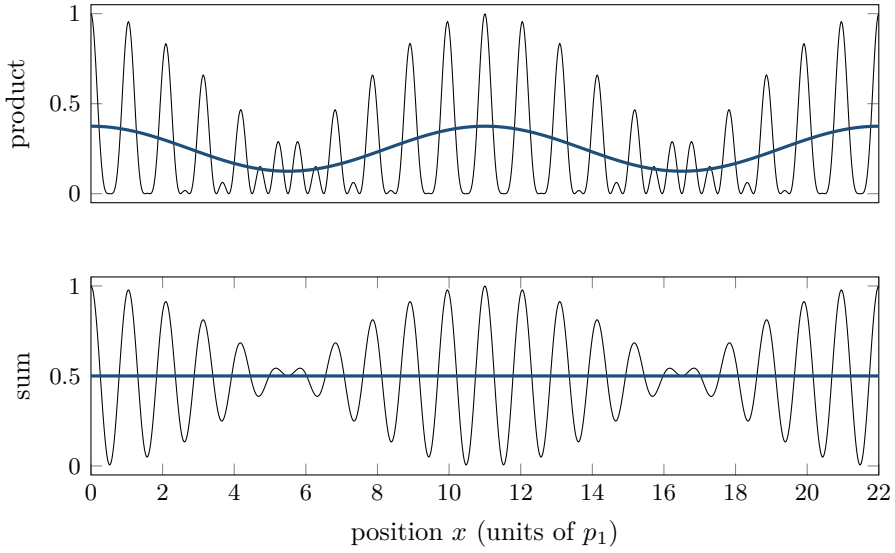


Figure 2.2: Two gratings with periods $p_1 = 1$ and $p_2 = 1.1$, combined two ways (thin lines) and then low-pass filtered (thick lines). Top: the product of the transmittances contains a component at the difference frequency, and the low-pass reader recovers fringes with period $p_M = 11$. Bottom: the normalized sum has an envelope at the same period, but no component at the difference frequency, and the same low-pass reader returns a constant.

Figure 2.2 shows the consequence for a low-pass reader. The sum of two gratings still has an *envelope*: there are places where the two patterns are in step and their peaks add, and places where they are out of step and the modulation cancels. Anyone plotting the sum sees the envelope. But a reader that averages over distances longer than the grating periods returns a uniform gray, because every term of the sum oscillates on the fine scale. To turn the envelope of an additive superposition into visible fringes, something nonlinear must intervene: a threshold, a saturation, a squaring detector. The relation

is present in both cases, as a fact about where the patterns coincide. In the multiplicative case it is present as a spatial frequency that a linear reader can pass; in the additive case it is present only implicitly, as a pattern of coincidences that a linear reader destroys.

This distinction is the first precise version of the book's central claim. Whether a relation between two structures is available as an observable depends on two things jointly: the physical operation that combined them and the operation performed by whatever receives the combination. Chapter 3 finds the same distinction in time, and Chapter 4 turns it into a general account of how difference frequencies are manufactured.

Sampling is a third way of combining two periodic structures, and it is multiplicative. A digital camera records the image only at the points of its pixel grid, which amounts to multiplying the image by a periodic array of impulses. When the image contains a pattern finer than the grid can represent, the product contains difference frequencies that fall inside the recorded band, and a photograph of a finely woven fabric or of a computer screen shows spurious fringes. Signal processing calls this aliasing (Oppenheim and Schafer 2010); it is moiré between the scene and the sensor. Printers meet the same problem when overlaying the halftone screens of different inks, and choose the relative angles of the screens so as to avoid the most visible moirés (Amidror 2009).

2.8 Moiré as a Relational Measuring Device

Because a moiré magnifies small relative differences, it has long served as an instrument. Rayleigh (1874) used the bands formed by superposing two gratings to test the regularity of their rulings: a perfectly matched pair gives uniform fringes, and any error in either ruling distorts them. The vernier scale, which reads a fraction of a division from the place where two slightly different scales coincide, works on the same principle in one dimension.

Moiré methods in experimental mechanics make the principle quantitative. A grating is bonded to the surface of a specimen and deforms with it; a reference grating stays undeformed; the fringes of the pair are contours of the specimen's in-plane displacement, one fringe for each period of relative displacement. Moiré interferometry, which uses very fine gratings and optical interference in place of shadows, measures such displacements with sub-micrometer sensitivity across a whole field (Post et al. 1994).

The instrument reads a relation. The fringes report the displacement of the specimen grating *relative to* the reference, and the reference is supplied by the instrument. This structure recurs throughout the book: many readers bring one relatum with them and measure the other against it. A superheterodyne receiver brings a local oscillator (Chapter 4), a resonator brings its natural

frequency (Chapter 6), and an electric fish brings its own discharge, against which it measures the field of a neighbor (Chapters 19 and 20).

2.9 Limits of the Analogy

The moiré is the most persuasive image in this book, and for that reason it needs the firmest limits. The proposal that consciousness is a moiré pattern, a structure arising from the mutual misalignment of many internal processes and present in none of them, has four problems that the analysis above makes explicit.

First, a moiré requires specific ingredients: two layers with nearby periods or orientations, a multiplicative combination (or a nonlinearity), and a reader that discards fine structure. Saying that experience is a moiré without naming the layers, the combining operation, and the reader says only that experience is relational, and Chapter 1 showed that relationality is too common to explain anything by itself.

Second, the visibility of a moiré is reader-relative. The fringes of (2.4) appear for a low-pass reader and would not appear for a reader that resolved each line. An explanation of experience by moiré therefore presupposes a reader of the right kind, and the nature of that reader is the question at issue. The analogy moves the problem without solving it.

Third, interference patterns are everywhere. Radios, radar receivers, and audio equipment are full of beats, difference frequencies, and aliasing, and nobody supposes that they are experienced. Whatever distinguishes an experienced relation from an unexperienced one, it cannot be the mere presence of an interference structure.

Fourth, moiré mechanisms have been proposed in neuroscience, for grid cells and for the orientation maps of visual cortex, and tested. Some predictions failed and others survived in modified form, and Chapter 47 tells that history as a lesson in how a beautiful relational mechanism can be wrong about biology.

What survives the analysis is a question rather than an answer. For any proposed relational structure, we can ask what physical operation exposes it: what combines the layers, and what receives the combination. That question is precise, it has different answers in different systems, and it organizes Part I.

2.10 Audit

Canonical phenomenon

Two line gratings, overlaid with a small difference in period or angle, produce broad fringes that move far when either grating moves a little.

Minimal model

Sinusoidal transmittances (2.1), multiplied (2.2), then low-pass filtered (2.4); fringe spacings (2.3), (2.7), and (2.8); displacement magnification (2.6).

What is relational here?

The relative period, relative angle, and relative displacement of the layers, which fix the moiré wavevector $\boldsymbol{k}_1 - \boldsymbol{k}_2$ and its phase (Proposition 2.3).

What physically reads it?

Two operations in sequence: the multiplication performed by light passing through both layers, which creates a component at the difference frequency, and the spatial low-pass filtering of the eye or camera, which keeps that component and discards the rest. In twisted bilayer graphene the reader is the electron system, which responds to the long-period potential. Under additive superposition, a linear low-pass reader finds nothing, and a nonlinearity is required.

What existing field explains it?

Wave optics and Fourier analysis (Amidror 2009; Oster, Wasserman, et al. 1964); experimental mechanics (Post et al. 1994); sampling theory (Oppenheim and Schafer 2010); solid-state physics for moiré superlattices (Bistritzer and MacDonald 2011). The explanation is complete.

What remains unexplained?

None in the physics. The chapter leaves one sharpened question: in any system proposed to exploit a relational structure, which operation combines the relata and which operation reads the combination.

A discriminating experiment

Superpose the same two gratings additively (two incoherent projectors) and multiplicatively (two transparencies), photograph each with a lens defocused to blur the lines, and measure the fringe contrast. The analysis predicts fringes of contrast set by (2.4) in the multiplicative case and none in the additive case until a nonlinear step (for example, a hard threshold on the recorded image before blurring) is inserted.

2.11 Notes and References

Amidror (2009) is the standard modern treatment, built on the spectral approach used here, with the superposition of periodic layers analyzed through the convolution of their spectra; it also covers moiré in color printing. Oster and Nishijima (1963) is an accessible account by one of the pioneers, and Oster, Wasserman, et al. (1964) introduced the indicial description of fringes as level curves of index differences. Rayleigh (1874) is the early use of superposed gratings to test rulings. For moiré interferometry in experimental mechanics, see Post et al. (1994). Aliasing is treated in any signal-processing text; Oppenheim and Schafer (2010) is thorough. Twisted bilayer graphene enters in Chapter 8; Bistritzer and MacDonald (2011) is the model paper, and Dean et al. (2013)

shows electrons responding to a moiré superlattice in graphene on hexagonal boron nitride.

2.12 Exercises

Exercise 2.1. Derive (2.2) from (2.1). Compute the moiré period for gratings with periods 0.50 mm and 0.52 mm, and the distance the fringes move when the second grating is shifted by 5 µm.

Exercise 2.2. Derive (2.8) from $|\boldsymbol{k}_1 - \boldsymbol{k}_2|$. Find the angle between the fringes and the lines of the first grating as a function of p_1, p_2, and θ, and show that it approaches 90° as $\theta \to 0$ with $p_1 = p_2$.

Exercise 2.3. Using a lattice constant of 0.246 nm, compute the moiré period of twisted bilayer graphene for twist angles of 0.5°, 1.1°, 2°, and 5°, and plot L_M against θ on logarithmic axes. What power law does the plot show at small angles, and where does it fail?

Exercise 2.4. Simulate the two superpositions of Figure 2.2 on a grid of 2000 points. Apply a moving-average filter of width $5p_1$ to each and confirm the results in the figure. Then threshold the additive superposition at its mean value before filtering, and measure the period and contrast of the fringes that appear. Explain in terms of Chapter 4 which operation created them.

Exercise 2.5. Two families of concentric circles are centered at $(-c, 0)$ and $(c, 0)$, with radii jp and lp for integers j and l (the cover of this book). Using the indicial description, show that the fringes $j - l = \text{const}$ are hyperbolas with foci at the two centers, and that the fringes $j + l = \text{const}$ are ellipses with the same foci. Which family is coarse, and why?

Exercise 2.6. A sinusoidal grating with a period of 0.9 pixel is photographed by a sensor that samples once per pixel. Find the period of the recorded pattern. How must the grating period be changed to make the recorded fringes twice as wide?

Exercise 2.7. Describe a physical reader for which the moiré of two superposed transparencies is invisible even though the difference-frequency term is present in the transmitted light. State precisely what property of the reader removes it.

3

Beats: Temporal Relations

Strike two tuning forks that are almost, but not quite, in tune, and the combined sound swells and fades several times a second. Neither fork is doing that. Each produces a steady tone; the slow waxing and waning belongs to the pair, and its rate is the difference of their frequencies. The beat is the temporal counterpart of the moiré fringe of Chapter 2: a slow structure generated by two fast ones, with a scale set by their mismatch.

The beat has a second property that makes it the right place to state the book's central problem in its first precise form. Its envelope is plain in the waveform and absent from the waveform's spectrum. Plot the sum of two tones and the envelope is the most obvious thing in the picture; compute the Fourier transform and there is no component at the difference frequency at all. Both descriptions are exact. Which one a physical system responds to depends on the operation it performs on the waveform, and this chapter shows five different operations on the same input making five different variables explicit.

3.1 Two Nearby Frequencies

Consider the sum of two unit-amplitude tones,

$$x(t) = \cos \omega_1 t + \cos \omega_2 t, \tag{3.1}$$

with ω_1 and ω_2 close together. Over a short stretch of time the sum looks like a single tone at a frequency between the two. Over a longer stretch, its amplitude rises and falls periodically, and at moments the signal almost vanishes. Figure 3.1 shows the waveform for tones at 10 Hz and 11 Hz, chosen for visibility.

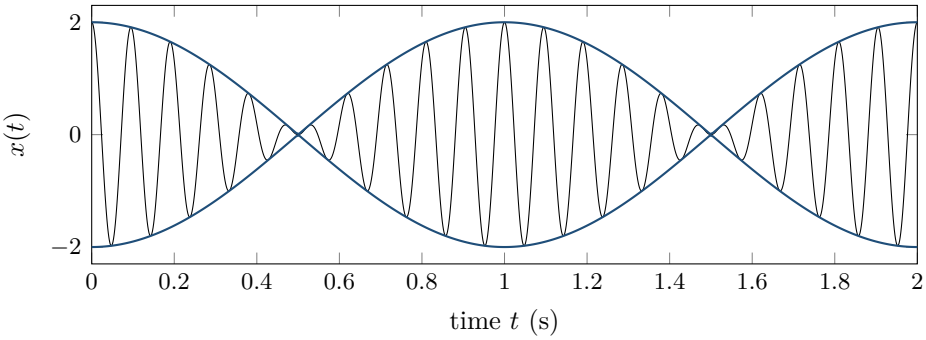


Figure 3.1: The sum of two tones at 10 Hz and 11 Hz (thin line) and its envelope $\pm 2|\cos(\pi t)|$ (thick lines). The envelope has one maximum per second, the difference of the two frequencies. The spectrum of the thin line contains only the two tones.

3.2 The Envelope Identity

The trigonometric identity for a sum of cosines rewrites (3.1) as a product:

$$\begin{aligned} \cos\omega_1 t + \cos\omega_2 t &= 2\cos\Big(\frac{\Delta\omega}{2}\,t\Big)\cos(\bar\omega\, t), \\ \Delta\omega = \omega_1 - \omega_2, &\qquad \bar\omega = \frac{\omega_1+\omega_2}{2}. \end{aligned} \tag{3.2}$$

The right-hand side is a fast carrier at the mean frequency $\bar\omega$, multiplied by a slow factor at half the difference frequency. The loudness maxima come twice per cycle of the slow factor, once when it is near $+2$ and once near -2, so the audible beat rate is the full difference, $|f_1 - f_2|$, and not half of it. The quantity that repeats at the beat rate is the magnitude $2|\cos(\Delta\omega\, t/2)|$.

Tones of unequal amplitude beat less deeply. Write the stronger tone as the reference, with amplitude a_1, and the weaker as $a_2 = m\,a_1$ with $0 < m < 1$:

$$\begin{aligned} x(t) &= a_1\cos\omega_1 t + a_2\cos(\omega_2 t + \psi_0) \\ &= a_1\ \mathrm{Re}\Big\{\mathrm{e}^{\mathrm{i}\omega_1 t}\big(1 + m\,\mathrm{e}^{\mathrm{i}\psi(t)}\big)\Big\}, \\ \psi(t) &= (\omega_2-\omega_1)t + \psi_0. \end{aligned} \tag{3.3}$$

The complex factor $1 + m\,\mathrm{e}^{\mathrm{i}\psi}$ traces a circle of radius m around the point 1 once per beat cycle. Its length is the envelope and its angle is a phase modulation of the carrier:

$$A(t) = a_1\sqrt{1 + m^2 + 2m\cos\psi(t)}, \qquad \varphi(t) = \arg\big(1 + m\,\mathrm{e}^{\mathrm{i}\psi(t)}\big). \tag{3.4}$$

The envelope swings between $a_1(1-m)$ and $a_1(1+m)$. The phase swings back and forth without drifting, because the circle never encloses the origin.

Definition 3.1 (Beat envelope and beat frequency)**.** For a signal composed of two tones with positive frequencies ω_1 and ω_2, the *beat envelope* is the

magnitude $A(t)$ of (3.4), and the *beat frequency* is $|f_1 - f_2|$, the rate at which $A(t)$ repeats.

Proposition 3.2 (Phase of a beating pair)**.** *For $0 < m < 1$ the phase modulation $\varphi(t)$ is periodic with the beat period, its largest excursion is* $\arcsin m$*, and the mean frequency of the combined signal, measured by the average rate of advance of its total phase $\omega_1 t + \varphi(t)$, equals ω_1, the frequency of the stronger tone.*

Proof. The point $1 + m\,\mathrm{e}^{\mathrm{i}\psi}$ moves on a circle of radius $m < 1$ centered at 1, so its angle is a single-valued periodic function of ψ and returns to its starting value after each beat cycle; the average of $\dot{\varphi}$ over a cycle is therefore zero. The angle is largest where the line from the origin is tangent to the circle, and there $\sin \varphi_{\max} = m$. □

The proposition has a consequence that returns in Chapter 20. A detector that counts zero crossings of a beating pair reports the frequency of the stronger tone, with a small periodic wobble at the beat rate; the weaker tone shows up as modulation of the stronger one's amplitude and phase. An electric fish receiving its own discharge together with a weaker discharge from a neighbor is in exactly this position, with its own signal as the stronger tone (Chapter 19).

3.3 Fourier Reality Versus Envelope Appearance

The Fourier transform of (3.1) consists of four impulses, at $\pm\omega_1$ and $\pm\omega_2$ (Appendix B). There is nothing at $\omega_1 - \omega_2$, and nothing at $\bar{\omega}$ either. The product form (3.2) is the same function written differently, and its factors are not spectral components: multiplying a cosine at $\Delta\omega/2$ by a cosine at $\bar{\omega}$ yields cosines at $\bar{\omega} \pm \Delta\omega/2$, which are ω_1 and ω_2 again.

The envelope is defined by an operation. The standard construction takes the *analytic signal* $z(t) = x(t) + \mathrm{i}\,\mathcal{H}[x](t)$, where $\mathcal{H}$ is the Hilbert transform, and defines the envelope as $|z(t)|$ and the instantaneous phase as $\arg z(t)$ (Appendix B). For the two-tone signal, $z(t) = a_1\mathrm{e}^{\mathrm{i}\omega_1 t} + a_2\mathrm{e}^{\mathrm{i}(\omega_2 t + \psi_0)}$, and $|z|$ is exactly the $A(t)$ of (3.4). The Hilbert transform is linear. The modulus is not. The envelope is a nonlinear functional of the waveform, and that is why a list of the waveform's frequency components does not contain it.

Both descriptions are complete, and they answer different questions. The spectrum says which sinusoids, held for all time, add up to the signal. The envelope says how the signal's amplitude varies over time. Which description a physical system tracks depends on how long it integrates. A system that averages over a window T short compared with $1/|f_1 - f_2|$ cannot separate the two frequencies; it sees a single tone whose loudness changes. A system that integrates over a window long compared with $1/|f_1 - f_2|$ can resolve two spectral lines, and for it the "beat" dissolves into two steady tones. The

trade-off between time resolution and frequency resolution is a theorem about windows (Appendix B); it holds for ears and instruments alike, and it means that "is there a beat?" has no answer until the observer's integration time is specified.

3.4 Difference and Sum Frequencies

A component at the difference frequency appears only when the sum passes through an operation that is not linear and time-invariant. Squaring is the simplest:

$$x^2(t) = 1 + \cos\big((\omega_1 - \omega_2)t\big) + \cos\big((\omega_1 + \omega_2)t\big) + \tfrac{1}{2}\cos 2\omega_1 t + \tfrac{1}{2}\cos 2\omega_2 t$$

for unit amplitudes, with a new term at the difference frequency, a new term at the sum frequency, harmonics, and a constant. Chapter 4 develops this systematically. Here we need only the qualitative point: the difference frequency is *manufactured* by the receiver, and a receiver without the right nonlinearity does not manufacture it.

Human hearing shows the three regimes that a single receiver passes through as the frequency difference grows. When two tones differ by a few hertz, listeners hear one tone whose loudness fluctuates at the beat rate. As the difference grows to tens of hertz, the fluctuation becomes too fast to follow as loudness and is heard as roughness. When the tones are far enough apart to excite separate regions of the cochlea, listeners hear two tones. Helmholtz traced musical dissonance to the roughness of beats between the partials of complex tones (Helmholtz 1885). Plomp and Levelt (1965) measured the dissonance of pairs of pure tones and found it greatest when the frequency difference is about a quarter of the critical bandwidth, the frequency range within which the ear's filters combine inputs. The beat is heard as a beat only while both tones fall within one auditory filter, whose output then carries the envelope; once the tones are resolved into separate filters, each filter sees one steady tone and the beat is gone.

3.5 Intermodulation

When a sum of tones passes through a nonlinear element, the output contains components at combinations $|k_1\omega_1 + k_2\omega_2|$ for integers k_1, k_2, the *intermodulation products*. Radio engineers treat them as a nuisance, since an amplifier driven by two strong stations produces spurious signals at frequencies neither station occupies (Horowitz and Hill 2015). The ear produces them too. Listeners presented with two tones hear additional tones that are not in the stimulus, of which the most prominent is the cubic difference tone at $2f_1 - f_2$ (Goldstein 1967), and these combination tones originate in the nonlinear mechanics of the cochlea itself (Robles and Ruggero 2001). The receiver adds relations to the

signal: it creates components whose frequencies are relational quantities of the input.

3.6 Phase Relations

The relative phase ψ_0 of the two tones does not change the beat rate. It changes *when* the envelope peaks. Delay the second tone by a time τ and the whole envelope pattern shifts by $\tau\,\omega_2/|\omega_2 - \omega_1|$, a large shift for a small delay when the tones are close, in exact analogy with the fringe magnification of Chapter 2.

Phase relations become informative when the same pair of tones is received at two places with different amplitude ratios. At a place where the weaker tone is relatively strong, both the envelope and the phase modulation are deep; where it is relatively weak, both are shallow. The two phase modulations have the same timing and different depths, so their difference is itself a signal modulated at the beat rate, and the relation between that difference and the envelope carries information that neither place carries alone. Chapter 21 shows that this is how an electric fish determines whether its neighbor's frequency is above or below its own, a question that the envelope at any single place cannot answer (Chapter 1, Exercise 3.6 below).

3.7 Binaural Beats

Everything so far concerned a single physical mixture: two tones added in the air and delivered to one ear. A different experiment separates the tones. Present a tone of frequency f_1 to the left ear and a tone of frequency f_2 to the right ear through headphones, so that no acoustic mixing occurs and each ear receives a pure, steady tone. Listeners nonetheless report a beat at $|f_1 - f_2|$ (Oster 1973). The effect is confined to relatively low carrier frequencies, and its limits were the subject of classic measurements by Licklider et al. (1950). Oster (1973) attributed binaural beats to the brainstem nuclei where signals from the two ears first converge.

Example 3.3 (Monaural and binaural beats). In a *monaural* beat the relation between the two tones is present in a single physical signal, the pressure at one eardrum, and the cochlea's filtering and nonlinear transduction make it explicit. In a *binaural* beat no physical signal contains the relation. The two tones never mix outside the head. Each ear transduces its own steady tone, and the relation between the two tones exists only as a relation between two neural signals. It becomes an explicit variable where neurons receive both signals and respond to their timing difference. In the monaural case the reader makes explicit a relation already present in a mixture; in the binaural case the reader constructs the mixture.

The distinction between these two cases will recur. The electric fish does both. Every patch of its skin receives a physical mixture of its own field and its neighbor's, as an ear receives two tones in the air; and its nervous system compares the timing of signals from different patches of skin, as the brainstem compares the two ears (Chapter 21).

3.8 The First Appearance of the Reader Problem

Take one input, the unequal two-tone signal (3.3), and pass it through five different physical operations D. Table 3.1 lists what each makes explicit.

Reader D	**Variable made explicit,** $z = \mathsf{D}[x]$
Narrow band-pass filter at ω_1	A steady tone at ω_1; the beat is invisible
Square-law detector, then low-pass filter	A sinusoid at $\lvert\omega_1 - \omega_2\rvert$: the beat as a signal
Magnitude of the analytic signal	The envelope $A(t)$ of (3.4)
Zero-crossing counter	The frequency ω_1 of the stronger tone, wobbling at the beat rate (Proposition 3.2)
Spectrum analyzer with window $T \gg 1/\lvert f_1 - f_2\rvert$	Two steady lines at ω_1 and ω_2; no beat

Table 3.1: Five readers of the same two-tone input and the variables they make explicit.

The table states the reader problem in the form it will keep for the rest of the book. The input contains the relation between the two tones implicitly, in the sense that each of the five outputs is some function of it. Each reader makes a different variable explicit, available to drive whatever the reader is connected to. A linear filter tuned to one tone discards the relation entirely. A square-law detector turns the relation into a frequency. None of this requires any claim about what the relation "really" is: every entry in the table is a statement about what a physical operation does to a waveform.

Two lessons carry forward. The first is that the envelope and the difference frequency are distinct variables: the envelope $A(t)$ of an unequal pair is a distorted periodic function at the beat rate, while the output of a square-law detector followed by a filter can be a pure sinusoid at the difference frequency. Different readers make different variables out of the same relation, and the variables differ in kind as well as in fidelity. The second is that no reader in the table can tell whether ω_2 lies above or below ω_1 from the envelope alone, because the envelope is unchanged when the sign of $\omega_2 - \omega_1$ and the sign of ψ_0 are reversed together. The sign is present in the input, in the direction in which the point $1 + m\,\mathrm{e}^{\mathrm{i}\psi}$ circles, but reading it requires combining amplitude with phase. Part V follows an animal that does exactly that.

3.9 Audit

Canonical phenomenon

Two tuning forks slightly out of tune produce a sound that swells and fades at the difference of their frequencies.

Minimal model

The product identity (3.2); for unequal amplitudes, amplitude and phase modulation (3.4); the envelope as the modulus of the analytic signal.

What is relational here?

The frequency difference $\omega_1 - \omega_2$, which sets the beat rate, and the relative phase ψ_0, which sets the timing of the envelope.

What physically reads it?

A receiver that integrates over less than a beat period and applies a nonlinearity: the cochlear filter containing both tones, followed by nonlinear transduction; for binaural beats, brainstem neurons that receive both ears. A linear receiver, or one that integrates long enough to resolve both lines, does not make the beat explicit.

What existing field explains it?

Acoustics and psychoacoustics (Helmholtz 1885; Plomp and Levelt 1965; Goldstein 1967); signal processing (analytic signals, time-frequency resolution) (Oppenheim and Schafer 2010); auditory neuroscience for binaural beats (Oster 1973; Licklider et al. 1950). All of the chapter's claims are standard.

What remains unexplained?

None. The chapter's contribution is the arrangement in Table 3.1: one input, several readers, several explicit variables, which is the reader problem stated for signals.

A discriminating experiment

Present listeners with the same pair of tones monaurally and binaurally while varying the carrier frequency and the frequency difference. The account predicts that the monaural beat survives wherever both tones share an auditory filter, while the binaural beat requires neural timing comparison and should fade at carrier frequencies where phase locking fails. A dissociation of the two limits confirms that the two beats are read by different operations.

3.10 Notes and References

Helmholtz (1885) remains a rewarding account of beats, combination tones, and roughness, and of the ear as an analyzer of periodic sound. Plomp and Levelt

(1965) is the classic measurement tying dissonance to the critical bandwidth. Combination tones and their origin in cochlear nonlinearity are covered by Goldstein (1967) and, for the mechanics, by Robles and Ruggero (2001). For binaural beats, see Oster (1973) and the measurements of Licklider et al. (1950). The analytic signal, the Hilbert transform, and the trade-off between time and frequency resolution are treated in Oppenheim and Schafer (2010) and Bracewell (2000); Appendix B collects what this book uses.

3.11 Exercises

Exercise 3.1. Prove (3.2). Show that $2|\cos(\Delta\omega\, t/2)|$ has period $2\pi/|\Delta\omega|$, and explain why the audible beat rate is $|f_1 - f_2|$ even though the slow factor in (3.2) oscillates at half that frequency.

Exercise 3.2. Derive (3.4) from (3.3). For $m = 0.3$, compute the ratio of the largest to the smallest envelope value and the largest phase excursion in degrees.

Exercise 3.3. Show that when $m > 1$ the mean frequency of (3.3) equals ω_2, and that at $m = 1$ the phase jumps by π at each zero of the envelope. Interpret the result in terms of the winding of $1 + m\,\mathrm{e}^{\mathrm{i}\psi}$ around the origin.

Exercise 3.4. Sample (3.1) with $f_1 = 100\,\mathrm{Hz}$ and $f_2 = 104\,\mathrm{Hz}$ at $2\,\mathrm{kHz}$. Compute magnitude spectra over rectangular windows of 0.1, 0.25, 0.5, and 2 s. For which windows are two peaks resolved, and how does the answer compare with $1/|f_1 - f_2|$?

Exercise 3.5. Pass the unequal pair (3.3) through a square-law detector. List every frequency in the output with its amplitude, and specify a low-pass cutoff that isolates the difference frequency when $\omega_1 = 2\pi \times 440\,\mathrm{Hz}$ and $\omega_2 = 2\pi \times 443\,\mathrm{Hz}$.

Exercise 3.6. Show that the envelope $A(t)$ in (3.4) is unchanged under the substitution $(\omega_2 - \omega_1, \psi_0) \mapsto (\omega_1 - \omega_2, -\psi_0)$, while the phase modulation $\varphi(t)$ changes sign. Conclude that the sign of the frequency difference cannot be read from the envelope alone, and describe what additional quantity would suffice.

Exercise 3.7. A model listener computes the envelope of the signal at each ear separately and then compares the two envelopes. Show that this listener cannot hear a binaural beat. Specify the minimal operation on the two ears' signals that yields a variable oscillating at $|f_1 - f_2|$, and state what property of the neural signals it requires.

4

How New Frequencies Become Physical

The moiré fringe of Chapter 2 and the beat of Chapter 3 are relations implicit in a mixture. In one case the mixture already contained a component at the difference frequency, because the layers were multiplied; in the other it did not, because the tones were added. This chapter asks what it takes, in general, to make such a relation explicit: to put energy at a frequency where the input had none, so that a filter can pass it and a downstream process can be driven by it.

The answer fits in two sentences. A linear time-invariant system can only reweight the frequencies already present in its input, so every new frequency, and with it every difference-frequency variable, is manufactured by a nonlinearity, by a multiplication, or by a time variation inside the receiver. Which of the manufactured frequencies a receiver makes available depends on its symmetry, on the filters around its nonlinearity, and on any reference it supplies itself. The chapter works through the standard devices (square-law and cubic nonlinearities, rectifiers, thresholds, mixers, heterodyne receivers, lock-in amplifiers) and ends with the Volterra series, which describes all of them at once and turns "which relations can this receiver read?" into a question about a measurable function.

4.1 Linear Systems

A system is *linear* if its response to a sum of inputs is the sum of its responses, and *time-invariant* if delaying the input delays the output and changes nothing else. Every linear time-invariant system acts by convolution with an impulse response h,

$$y(t) = \int h(\tau)\, x(t-\tau)\, \mathrm{d}\tau, \qquad \hat{y}(\omega) = H(\omega)\, \hat{x}(\omega), \tag{4.1}$$

where H is the transfer function, the Fourier transform of h (Appendix B).

Theorem 4.1 (Linear time-invariant systems create no frequencies)**.** *If $y = h * x$, then the spectrum of y vanishes wherever the spectrum of x vanishes. In particular, the response to $a_1 \cos \omega_1 t + a_2 \cos \omega_2 t$ contains components at ω_1 and ω_2 only.*

Proof. By (4.1), $\hat{y}(\omega) = H(\omega)\hat{x}(\omega) = 0$ wherever $\hat{x}(\omega) = 0$. □

A linear time-invariant reader can still change how visible a relation is. A filter that attenuates one of two beating tones reduces the depth of the envelope, and a filter that delays one tone relative to the other shifts the envelope in time. What it cannot do is convert the envelope into a component at the difference frequency. No arrangement of linear filters, however elaborate, will produce from two tones a signal that oscillates at their difference.

Both qualifiers in the theorem matter. A system can be linear in its input and still create frequencies if it varies in time: multiplying the input by a periodic gain $g(t) = \cos \omega_{\mathrm{LO}} t$ obeys superposition and moves every input component at ω to $\omega \pm \omega_{\mathrm{LO}}$. Such a system contains a clock, and the clock is a relatum supplied by the receiver. Mixers, heterodyne receivers, and lock-in amplifiers are of this kind (Sections 4.6, 4.7 and 4.9). A nonlinear system creates new frequencies with no clock at all, from the input alone.

4.2 Quadratic Nonlinearities

The simplest nonlinear reader squares its input. For a two-tone input $x = a_1 \cos \omega_1 t + a_2 \cos \omega_2 t$, expanding and applying the product formulas for cosines gives

$$\begin{aligned} x^2(t) = {} & \tfrac{1}{2}(a_1^2 + a_2^2) + \tfrac{1}{2} a_1^2 \cos 2\omega_1 t + \tfrac{1}{2} a_2^2 \cos 2\omega_2 t \\ & + a_1 a_2 \cos\big((\omega_1 - \omega_2)t\big) + a_1 a_2 \cos\big((\omega_1 + \omega_2)t\big). \end{aligned} \tag{4.2}$$

The output contains a constant, the second harmonic of each tone, and two new components at the difference and sum frequencies (Figure 4.1). The difference and sum terms have amplitude $a_1 a_2$: they exist only when both tones are present, and they grow with the product of the two amplitudes. Their frequencies are relational quantities of the input, and their amplitudes are joint quantities.

The same bookkeeping works for any polynomial. Write each cosine as a sum of exponentials, $\cos \theta = \tfrac{1}{2}(\mathrm{e}^{\mathrm{i}\theta} + \mathrm{e}^{-\mathrm{i}\theta})$, and expand.

Proposition 4.2 (Combination frequencies of a polynomial reader)**.** *Let $y = \sum_{n=0}^{N} c_n x^n$ and $x = a_1 \cos \omega_1 t + a_2 \cos \omega_2 t$. Then y contains only the frequencies $|k_1 \omega_1 + k_2 \omega_2|$ with integers k_1, k_2 satisfying $|k_1| + |k_2| \le N$, and the power x^n contributes only combinations with $|k_1| + |k_2| \le n$ and $|k_1| + |k_2|$ of the same parity as n.*

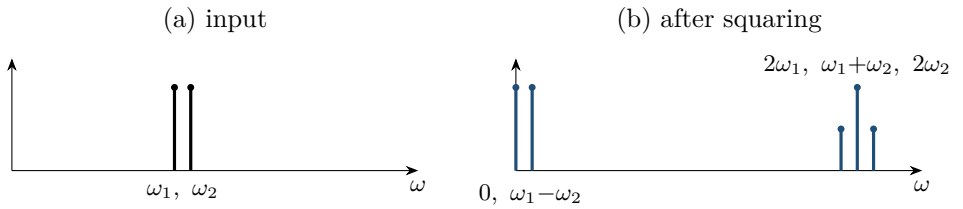


Figure 4.1: Spectra of a two-tone input with equal amplitudes (a) and of its square (b). The squarer creates a constant, a component at the difference frequency, one at the sum frequency, and second harmonics. A low-pass filter after the squarer keeps the constant and the difference term.

Proof. Each term in the expansion of x^n is a product of n exponentials, each of the form $\mathrm{e}^{\pm \mathrm{i}\omega_1 t}$ or $\mathrm{e}^{\pm \mathrm{i}\omega_2 t}$. If $n_1^{\pm}$ and $n_2^{\pm}$ count the factors of each kind, the product is $\mathrm{e}^{\mathrm{i}(k_1\omega_1 + k_2\omega_2)t}$ with $k_1 = n_1^+ - n_1^-$ and $k_2 = n_2^+ - n_2^-$. Then $|k_1| + |k_2| \le n_1^+ + n_1^- + n_2^+ + n_2^- = n$, and $k_1 + k_2$ differs from n by twice the number of negative factors, so it has the parity of n; $|k_1| + |k_2|$ has the parity of $k_1 + k_2$. □

The number $|k_1| + |k_2|$ is the *order* of the combination: the difference and sum frequencies are second-order products, $2\omega_1 - \omega_2$ is third-order, and so on. Engineers call the unwanted ones intermodulation distortion (Horowitz and Hill 2015). For our purposes they are the catalogue of relational variables that a polynomial receiver can make explicit.

4.3 Cubic Nonlinearities

Cubing the same two-tone input and collecting terms gives

$$\begin{aligned} x^3(t) = {} & \left(\tfrac{3}{4}a_1^3 + \tfrac{3}{2}a_1a_2^2\right)\cos\omega_1 t + \left(\tfrac{3}{4}a_2^3 + \tfrac{3}{2}a_1^2a_2\right)\cos\omega_2 t \\ & + \tfrac{1}{4}a_1^3\cos 3\omega_1 t + \tfrac{1}{4}a_2^3\cos 3\omega_2 t \\ & + \tfrac{3}{4}a_1^2a_2\left[\cos\big((2\omega_1 - \omega_2)t\big) + \cos\big((2\omega_1 + \omega_2)t\big)\right] \\ & + \tfrac{3}{4}a_1a_2^2\left[\cos\big((2\omega_2 - \omega_1)t\big) + \cos\big((2\omega_2 + \omega_1)t\big)\right]. \end{aligned} \tag{4.3}$$

Three features stand out. The cubic term changes the response at the input frequencies themselves, and the change at ω_1 depends on a_2: each tone modulates the gain seen by the other. It creates the third-order products $2\omega_1 \pm \omega_2$ and $2\omega_2 \pm \omega_1$; when the tones are close, $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$ lie just outside the pair, on either side, where a filter tuned near the primaries cannot exclude them. And it creates no constant and no component at $\omega_1 - \omega_2$.

The last feature is a consequence of symmetry, and it holds far beyond the cube.

Proposition 4.3 (Odd readers cannot make the plain difference frequency)**.** *Let f be a memoryless nonlinearity that is odd, $f(-x) = -f(x)$, and let*

$x = a_1 \cos\theta_1 + a_2 \cos\theta_2$ with $\theta_i = \omega_i t$. Then the output $f(x)$ contains only combinations $k_1\omega_1 + k_2\omega_2$ with $k_1 + k_2$ odd. In particular, when ω_1 and ω_2 are incommensurate, so that distinct combinations have distinct frequencies, it contains no constant term and no components at $\omega_1 \pm \omega_2$. If f is even, only combinations with $k_1 + k_2$ even appear.

Proof. As a function of the two phases, $F(\theta_1, \theta_2) = f(a_1 \cos\theta_1 + a_2 \cos\theta_2)$ is 2π-periodic in each and has a double Fourier series $\sum c_{k_1 k_2} \mathrm{e}^{\mathrm{i}(k_1\theta_1 + k_2\theta_2)}$. Shifting both phases by π reverses the sign of x, so for odd f, $F(\theta_1 + \pi, \theta_2 + \pi) = -F(\theta_1, \theta_2)$. The shift multiplies each Fourier term by $(-1)^{k_1+k_2}$, so every term with $k_1 + k_2$ even must vanish. The even case is the same argument with the sign reversed. □

The proposition says that which relations a receiver can make explicit depends on its symmetry. A symmetric saturating amplifier, whose response to a negative input is the mirror image of its response to a positive one, turns two tones into third-order products at $2\omega_1 - \omega_2$ but never into a component at $\omega_1 - \omega_2$. To obtain the plain difference frequency, the receiver needs an asymmetric nonlinearity with an even part: a square law, a rectifier, a threshold set away from zero.

Human hearing illustrates both halves. Listeners presented with two tones hear combination tones that are not in the stimulus, and the most prominent of them is the cubic difference tone at $2f_1 - f_2$, audible when the two primaries are close in frequency (Goldstein 1967). The nonlinearity that produces it lies in the mechanics of the cochlea (Robles and Ruggero 2001), and the distortion products it generates travel back out of the ear: Kemp (1978) showed that the ear emits sound in response to stimulation, which opened the way to recording cochlear distortion products with a microphone in the ear canal. The receiver adds relational components to its own input, and they can be measured outside it.

4.4 Rectification

A half-wave rectifier passes the positive part of its input, $r(x) = \max(x, 0)$; a full-wave rectifier returns the magnitude, $|x|$. Neither is odd, so both can produce the difference frequency. The full-wave rectifier is even, and by Proposition 4.3 it produces only even combinations: a constant, second harmonics, and the sum and difference frequencies, together with higher even orders.

A rectifier followed by a low-pass filter is an *envelope detector*, the circuit that demodulates amplitude-modulated radio. Suppose the input is a carrier whose amplitude varies slowly, $x(t) = A(t)\cos\omega_c t$, with $A(t) > 0$ changing little over a carrier cycle. The average of $\max(A\cos\phi, 0)$ over one cycle of ϕ is

A/π, so the filtered output of a half-wave rectifier is

$$\mathrm{LPF}\big[\max(x,0)\big](t) = \frac{A(t)}{\pi}, \tag{4.4}$$

and a full-wave rectifier gives $2A(t)/\pi$. For the equal-amplitude beating pair of Chapter 3, the slow factor $\cos(\Delta\omega\, t/2)$ changes sign, but the average of $\max(A\cos\phi, 0)$ is $|A|/\pi$ for either sign of A, so both rectifiers deliver a multiple of $|\cos(\Delta\omega\, t/2)|$. That function repeats at $|\Delta\omega|$: the envelope detector turns the implicit beat into an explicit slow signal.

Nerve cells rectify. A firing rate cannot be negative, and a neuron driven by an oscillating input fires on one half of the cycle and falls silent on the other. Sensory neurons whose firing rate follows the amplitude of a carrier act, to a first approximation, as rectifiers followed by smoothing. In the electric fish, one class of electroreceptor afferents, the P-units, changes its probability of firing with the amplitude of the fish's own electric field (Scheich et al. 1973), which makes it an amplitude reader of this general kind. Chapter 21 follows what the fish does with the result.

4.5 Thresholding

A threshold element produces an output only when its input exceeds a level ϑ. Applied to the beating pair, it responds on those carrier cycles whose peaks rise above ϑ and falls silent when the envelope drops below it. If it responds at most once per carrier cycle, at the moment the input crosses ϑ from below, its output is a train of pulses with two independent properties. Whether a pulse occurs on a given cycle depends on the envelope. When it occurs depends on the phase of the carrier, including the slow phase modulation $\varphi(t)$ of Chapter 3.

A single threshold therefore makes two relational variables available through two different codes: the probability of a pulse carries the amplitude modulation, and the timing of the pulse carries the phase modulation. Nervous systems exploit both codes, and the electric fish assigns them to different receptor classes. Its P-units change their probability of firing with the amplitude of the field, and its T-units fire one spike per cycle of the fish's discharge, locked in time to a fixed phase of the cycle, so that their spike times encode phase (Scheich et al. 1973). The two codes travel along separate pathways into the brain, and the sign of a frequency difference emerges only where they are combined (Chapter 21).

4.6 Multiplication

An ideal *mixer* multiplies two inputs:

$$\cos\omega_1 t\, \cos\omega_2 t = \tfrac{1}{2}\cos\big((\omega_1-\omega_2)t\big) + \tfrac{1}{2}\cos\big((\omega_1+\omega_2)t\big). \tag{4.5}$$

The product contains the difference and sum frequencies and nothing else: no constant and no harmonics. Electronic mixers approximate this with diodes or transistors (Horowitz and Hill 2015). A mixer has two input ports; the square-law device of Section 4.2 is a one-port mixer, since $(x_1 + x_2)^2 = x_1^2 + x_2^2 + 2x_1x_2$ contains the product together with the squares of each input. The two transparencies of Chapter 2 form a spatial mixer, which is why their superposition contains the moiré frequency.

From the point of view of either input alone, a mixer is linear and time-varying: the other input acts as a gain that changes in time. This is the loophole in Theorem 4.1 noted at the start of the chapter, and it is the basis of heterodyne reception.

4.7 Heterodyning

A heterodyne receiver supplies the second input itself. It generates a local oscillation at ω_{LO} and mixes it with the incoming signal at ω_s, producing a component at the *intermediate frequency* $\omega_{\mathrm{IF}} = |\omega_s - \omega_{\mathrm{LO}}|$. Armstrong's superheterodyne receiver used this to convert every incoming station to one fixed intermediate frequency, where amplification and selective filtering could be designed once and done well (E. Armstrong 1921; Horowitz and Hill 2015).

The heterodyne receiver is the first reader in this book that deliberately supplies one relatum and reads the relation between the incoming signal and it. The intermediate frequency is a detuning: a relation between the world's frequency and the receiver's own.

It also has a characteristic blind spot. A real mixer followed by a filter at ω_{IF} responds identically to signals at $\omega_{\mathrm{LO}} + \omega_{\mathrm{IF}}$ and at $\omega_{\mathrm{LO}} - \omega_{\mathrm{IF}}$, because the product formula (4.5) depends only on the magnitude of the difference. The unwanted one is called the *image*, and a simple receiver cannot tell on which side of its local oscillator a signal lies. Receivers remove the ambiguity in one of two ways. They filter out the image band before the mixer, or they mix in *quadrature*: multiply the input by $2\cos\omega_{\mathrm{LO}}t$ and by $-2\sin\omega_{\mathrm{LO}}t$, low-pass both products, and treat the two results as the real and imaginary parts of one complex signal. For an input $\cos\omega_s t$ this yields $\mathrm{e}^{\mathrm{i}(\omega_s-\omega_{\mathrm{LO}})t}$, whose frequency carries its sign: it rotates one way in the complex plane when the signal is above the local oscillator and the other way when it is below.

Example 4.4 (The image problem in a fish). A weakly electric fish is a heterodyne receiver whose local oscillator is its own electric discharge. A neighbor's discharge at a nearby frequency mixes with the fish's own field at the skin and produces a beat at $|\Delta f|$, the fish's intermediate frequency. The fish needs to know the sign of Δf, because it must shift its own frequency away from its neighbor's (Chapter 20). That is the image problem. The fish has no image filter, and it has no internal quadrature copy of its own discharge to mix against (Heiligenberg 1991a). It resolves the sign by combining amplitude

modulation with phase differences between regions of its own body, which together play the role of the two quadrature channels (Chapter 21).

4.8 Filtering

A nonlinearity creates many frequencies at once; a filter chooses which of them becomes the output. Practical readers are cascades of the form

$$z = L_2\big[\, N\big(L_1[y]\big)\big], \tag{4.6}$$

a linear filter L_1, a memoryless nonlinearity N, and a second linear filter L_2; the cascades with only one of the filters are known as Wiener and Hammerstein systems (Schetzen 1980). The same nonlinearity combined with different filters yields different readers. A squarer followed by a low-pass filter is a beat detector; a squarer followed by a band-pass filter at 2ω is a frequency doubler; a mixer followed by a filter at ω_{IF} is a frequency converter.

The filter before the nonlinearity matters as much as the one after it. If L_1 separates the two tones, sending each to a different branch before any nonlinearity acts, then no branch ever contains both tones, and no difference frequency is produced anywhere. This is the physical reason for the observation in Chapter 3 that widely separated tones, which excite different regions of the cochlea, do not beat. Section 4.10 gives the exact form of the statement.

4.9 Lock-in Detection

A lock-in amplifier measures a weak signal at a known frequency buried in noise. It multiplies the input $y(t) = A\cos(\omega_r t + \phi) + n(t)$ by a reference $\cos(\omega_r t + \phi_r)$ and by its quadrature $-\sin(\omega_r t + \phi_r)$, and low-pass filters both products with a long time constant T. The outputs are

$$X = \tfrac{1}{2} A\cos(\phi - \phi_r), \qquad Y = \tfrac{1}{2} A\sin(\phi - \phi_r), \tag{4.7}$$

from which the amplitude A and the phase difference $\phi - \phi_r$ follow. A noise component at $\omega_r + \delta$ appears after mixing at frequency δ and is removed by the filter when $\delta \gg 1/T$, so the instrument behaves as a band-pass filter centered on the reference with a width of order $1/T$, which can be made far narrower than any fixed filter (Scofield 1994). Dicke (1946) used the same principle in his microwave radiometer, switching the receiver between the antenna and a reference load at a known rate and detecting synchronously, which removed the effect of slow drifts in the receiver's gain.

The lock-in makes explicit a relation between its input and its own clock: the phase of the signal relative to the reference. It is a powerful reader precisely because it brings a relatum of its own, and it is useless without one. The contrast with the electric fish is instructive. The fish reads the sign of a

frequency difference without any internal reference to its own discharge; the relevant relation is between two parts of its input, the fields at two regions of its skin, and not between the input and an internal clock (Heiligenberg 1991a; Kawasaki 1993). Chapter 22 returns to what that absence means.

4.10 Volterra Series

The devices above are special cases of a general description of time-invariant nonlinear systems with fading memory, the Volterra series:

$$\begin{aligned} y(t) &= h_0 + \int h_1(\tau)\, x(t-\tau)\, \mathrm{d}\tau \\ &\quad + \iint h_2(\tau_1,\tau_2)\, x(t-\tau_1)\, x(t-\tau_2)\, \mathrm{d}\tau_1 \mathrm{d}\tau_2 + \cdots \end{aligned} \tag{4.8}$$

(Schetzen 1980). The first-order kernel h_1 is the impulse response of the system's linear part. The second-order kernel h_2, which can be taken symmetric in its arguments, describes how pairs of input values at different delays combine, and its Fourier transform

$$H_2(\omega_1,\omega_2) = \iint h_2(\tau_1,\tau_2)\, \mathrm{e}^{-\mathrm{i}(\omega_1\tau_1+\omega_2\tau_2)}\, \mathrm{d}\tau_1 \mathrm{d}\tau_2$$

describes how pairs of input frequencies combine. For a pure squarer, h_2 is a product of two delta functions and $H_2 \equiv 1$.

Proposition 4.5 (The difference-frequency term of a second-order reader)**.** *Let a system be described by* (4.8) *with symmetric h_2, and let $x = a_1 \cos\omega_1 t + a_2 \cos\omega_2 t$. The second-order part of the output contains, at the difference frequency, the component*

$$a_1 a_2\, |H_2(\omega_1,-\omega_2)|\, \cos\big((\omega_1-\omega_2)t + \arg H_2(\omega_1,-\omega_2)\big). \tag{4.9}$$

Higher even orders add corrections of relative size $O(a_1^2 + a_2^2)$.

Proof. Write x as a sum of four exponentials $\tfrac{1}{2}a_j \mathrm{e}^{\pm\mathrm{i}\omega_j t}$. Substituted into the second-order term, each ordered pair of exponentials with frequencies α and β contributes $\tfrac{1}{4}aa'\, H_2(\alpha,\beta)\, \mathrm{e}^{\mathrm{i}(\alpha+\beta)t}$. The pairs $(\omega_1,-\omega_2)$ and $(-\omega_2,\omega_1)$ both land at $\omega_1-\omega_2$ and, by symmetry, contribute equally, giving

$$\tfrac{1}{2} a_1 a_2\, H_2(\omega_1,-\omega_2)\, \mathrm{e}^{\mathrm{i}(\omega_1-\omega_2)t}.$$

Adding the complex conjugate term at $-(\omega_1-\omega_2)$ yields (4.9). For the squarer, $H_2 \equiv 1$ and the result agrees with (4.2). □

To lowest order, then, a receiver makes the difference frequency of two tones explicit exactly when $H_2(\omega_1,-\omega_2) \neq 0$. The second-order kernel is a precise and measurable description of which pairwise frequency relations a receiver

can convert into signals. It can be estimated from input and output records, for instance by driving the system with noise and computing input–output correlations, which is the basis of Wiener's white-noise method of system identification (Schetzen 1980). Estimating a reader's kernels before and after an intervention on the reader is one of the experimental designs of Chapter 58.

Example 4.6 (A filter before the squarer). Let the reader be a linear filter with impulse response g followed by a squarer. Then $h_2(\tau_1, \tau_2) = g(\tau_1)g(\tau_2)$ and $H_2(\omega_1, -\omega_2) = G(\omega_1)\,G(-\omega_2) = G(\omega_1)\,\overline{G(\omega_2)}$ for real g. The difference term has amplitude $a_1 a_2 |G(\omega_1)||G(\omega_2)|$. If the filter passes ω_1 and blocks ω_2, the difference term vanishes even though the squarer is perfectly capable of producing it. This is the exact form of the statement in Section 4.8: a pair of tones can beat only inside a channel that passes both.

4.11 Effective Observables

The chapter's results can now be collected into two definitions.

Definition 4.7 (Implicit relation)**.** A relational variable r is *implicit* in an input y if $r = \Phi[y]$ for some operation Φ, so that the value of r is fixed by the input and could in principle be recovered from it.

Definition 4.8 (Dynamically explicit variable)**.** A relational variable r is *dynamically explicit* at the output of a physical system D if some component of that output, available to the processes coupled to D, varies with r and, over the range of inputs considered, with nothing else beyond a stated tolerance.

A difference frequency is implicit in the sum of two tones and explicit at the output of a square-law detector followed by a low-pass filter. The moiré frequency is explicit in the light transmitted by two transparencies and merely implicit in the light from two superposed projectors. The results of this chapter state exactly when the passage from the first status to the second is possible:

> *Which relational variables become dynamically explicit depends jointly on the physical mixture and on the transduction performed by the receiver.*

The mixture fixes what is implicit. The receiver's symmetry, its filters, its time variation, and any reference it supplies fix what becomes explicit, and Proposition 4.5 expresses the dependence in a measurable quantity. None of this requires a view about whether relations "really exist" before they are read. The envelope of a beat is a property of the waveform; the component at the difference frequency is a property of the output of certain receivers; and both statements are about physics.

A third status lies beyond these two. A variable can be explicit at a receiver's output and still drive nothing of consequence, like the beat signal

of a square-law detector whose output is connected to nothing. When the explicit variable is *used*, when it feeds a controller that acts on the world or on the receiver itself, the reader becomes part of a loop, and the questions change. Part IV takes them up, and Part V follows a system in which all three statuses can be traced: the sign of a frequency difference is implicit in the fields at the fish's skin, made explicit by neurons in its midbrain, and used by the pacemaker that sets the fish's own frequency.

4.12 Audit

Canonical phenomenon

A square-law detector driven by two tones produces a signal at their difference frequency; a linear amplifier driven by the same tones does not.

Minimal model

Theorem 4.1 for linear time-invariant systems; the polynomial and symmetry results (Propositions 4.2 and 4.3); the second-order Volterra term (Proposition 4.5).

What is relational here?

The combination frequencies $k_1\omega_1 + k_2\omega_2$ of the input tones, above all the detuning $\omega_1 - \omega_2$, and the relative phases that fix the phases of those combinations.

What physically reads it?

A nonlinearity, a multiplication, or a time-varying gain, followed by a filter; in the heterodyne receiver and the lock-in amplifier, a nonlinearity or multiplication fed with a reference the reader supplies. The second-order kernel H_2 says which pairwise relations a given reader can convert into signals.

What existing field explains it?

Signal processing and nonlinear system theory (Schetzen 1980; Oppenheim and Schafer 2010); radio engineering (E. Armstrong 1921; Horowitz and Hill 2015); measurement science (Dicke 1946; Scofield 1994); cochlear mechanics and psychoacoustics (Robles and Ruggero 2001; Goldstein 1967). These fields explain everything in the chapter.

What remains unexplained?

Nothing in the engineering. The chapter leaves two questions for later parts: what changes when the receiver produces one of the relata itself and has no separate reference (Chapters 20 and 21), and what it means for an explicit variable to be read *for* the receiver rather than for the engineer who connects it to something (Chapters 25 and 26).

A discriminating experiment

Drive a receiver with two tones and measure its output at $\omega_1 - \omega_2$ and at $2\omega_1 - \omega_2$. Then replace its nonlinearity with an odd-symmetric one (symmetric clipping) while keeping its filters fixed. Proposition 4.3 predicts that the component at $\omega_1 - \omega_2$ disappears while the one at $2\omega_1 - \omega_2$ persists. For a biological receiver whose nonlinearity cannot be replaced, estimate H_2 from noise-driven recordings and predict which difference frequencies it will produce before testing them.

4.13 Notes and References

Schetzen (1980) is the classic treatment of Volterra and Wiener series and of the identification of nonlinear systems from data. Mixers, rectifiers, envelope detectors, and heterodyne receivers are covered in Horowitz and Hill (2015); E. Armstrong (1921) is the original description of the superheterodyne. For lock-in detection, see Scofield (1994), and for its use in radiometry, Dicke (1946). Linear systems, the convolution theorem, and quadrature signals are treated in Oppenheim and Schafer (2010). The auditory side of the chapter draws on Goldstein (1967) for combination tones, Kemp (1978) for the discovery that the ear emits sound, and Robles and Ruggero (2001) for cochlear nonlinearity. The electroreceptor classes of the electric fish are described in Scheich et al. (1973); Part V gives the full account.

4.14 Exercises

Exercise 4.1. Derive (4.2) and (4.3). Tabulate every frequency with its amplitude, and check the table against Proposition 4.2.

Exercise 4.2. Apply the argument of Proposition 4.3 to the half-wave rectifier $\max(x, 0)$ by writing it as the sum of an even part and an odd part. Which combination frequencies does each part produce? Compute numerically the amplitude of the component at $\omega_1 - \omega_2$ when $a_1 = 1$ and $a_2 = 0.5$.

Exercise 4.3. Verify that the average of $\max(A\cos\phi, 0)$ over a cycle of ϕ is $|A|/\pi$. For the equal-amplitude beating pair, compute the Fourier coefficient of $|\cos(\Delta\omega\, t/2)|$ at frequency $|\Delta\omega|$ and hence the amplitude of the beat signal at the output of a full-wave envelope detector.

Exercise 4.4. A receiver with an intermediate frequency of 10.7 MHz and its local oscillator above the signal is tuned to 100.0 MHz. Find the local oscillator frequency and the image frequency. Show that quadrature mixing assigns opposite signs to the signal and to the image, and write down the operation that selects one of them.

Exercise 4.5. Derive (4.7). A disturbance lies 1 Hz away from the reference frequency. Using a first-order low-pass filter with time constant T, find the

attenuation of the disturbance at the lock-in output as a function of T, and choose T for an attenuation of a factor of 100.

Exercise 4.6. For the cascade of Example 4.6 with $g(\tau) = \tau_0^{-1} \mathrm{e}^{-\tau/\tau_0}$ for $\tau \geq 0$, compute $|H_2(\omega_1, -\omega_2)|$ and plot the amplitude of the difference term against ω_2 for fixed $\omega_1 = 1/\tau_0$. Repeat with a narrow band-pass filter centered on ω_1 and explain the difference.

Exercise 4.7. Simulate a threshold element with $\vartheta = 1$ driven by $\cos \omega_1 t + 0.5 \cos \omega_2 t$ with $\omega_2/\omega_1 = 1.02$, emitting at most one pulse per carrier cycle at the upward crossing. Plot the fraction of cycles with a pulse and the pulse timing relative to the carrier, both against the beat phase ψ. Show that neither pulse probability nor pulse timing, recorded as a time series on its own, determines the sign of $\omega_2 - \omega_1$, and describe how the two together could.

You can now separate what belongs to components from what belongs to their configuration. Moiré fringes and beat envelopes are the two cleanest cases: a structure present in neither layer, neither tone, alone. You have also met the first version of the book's question. A linear sum of two tones contains no component at their difference frequency; a square-law detector puts one there. Which relational variables become explicit depends on the mixture and on the operation that receives it. Part II asks whether the "special" relations that make such structures dramatic are one phenomenon or several.

Part II

Special Relations Are Not One Thing

Which special relation is this: balance, resonance, or commensurability?

5

Magnitude Balance and Crossover

Popular accounts of complex systems often say that something interesting happens at "the right relation" between two things: order and disorder, coupling and diversity, stability and flexibility. Part II separates three different phenomena that the phrase runs together. This chapter treats the first and simplest, the balance of two magnitudes. Chapter 6 treats the match of a driving frequency to a natural one, Chapter 7 the rational ratio of two frequencies, and Chapter 8 shows all three distinctions at work in one material.

The thesis of the chapter is that a magnitude balance is a ratio of two quantities with the same dimensions, and that when this ratio passes through values of order one, a system crosses from a regime dominated by one effect to a regime dominated by the other. The fact is real and general, and it is fully owned by dimensional analysis and the theory of bifurcations and phase transitions. The popular slogan built on it, that the optimum lies in the middle, is too weak to be a law unless the two magnitudes, their ratio, the location of the crossing, and its signature are all named.

5.1 Dimensionless Numbers

A dimensionless number is a ratio of two quantities with the same physical dimensions, such as a ratio of two forces, two energies, two lengths, or two rates:

$$\Pi = \frac{A}{B}. \tag{5.1}$$

It is the same in every system of units, so it is the fifth kind of relational observable listed in Chapter 1, the scale ratio. Its value says which of the two effects dominates.

The number of such ratios needed to describe a system follows from a theorem.

Theorem 5.1 (Buckingham's Π theorem)**.** *Suppose a physical relation among n quantities $q_1, \dots, q_n$ is dimensionally homogeneous, and the quantities involve k independent fundamental dimensions. Then the relation can be written as $F(\Pi_1, \dots, \Pi_{n-k}) = 0$, where $\Pi_1, \dots, \Pi_{n-k}$ are independent dimensionless products of powers of the q_i.*

Sketch. Represent each quantity by the vector of exponents of its dimensions. The dimension matrix has rank k, so the space of exponent vectors $(e_1, \dots, e_n)$ for which $\prod q_i^{e_i}$ is dimensionless has dimension $n-k$; a basis of that space gives the Π_j. A dimensionally homogeneous relation is unchanged by any rescaling of units, and this invariance forces it to depend on the q_i only through the Π_j (Buckingham 1914; Barenblatt 1996). □

Example 5.2 (The pendulum). The period T of a pendulum might depend on its length ℓ, the gravitational acceleration g, and the mass m of the bob. These four quantities involve three dimensions (mass, length, time), so there is one dimensionless product, $T\sqrt{g/\ell}$, and the theorem says it must be a constant, at least for small swings. Mass cannot enter, because no other quantity carries the dimension of mass to cancel it. The small-amplitude result $T = 2\pi\sqrt{\ell/g}$ supplies the constant; dimensional analysis alone had already fixed everything else.

The pendulum shows the power of the theorem and also its limit. Dimensional analysis says which ratios can matter. It does not say what the function F is, and it does not say where anything interesting happens.

5.2 Order-One Transitions

When $\Pi \ll 1$, effect B dominates and A can be treated as a small correction; when $\Pi \gg 1$, the roles reverse. Somewhere in between the system changes character.

Definition 5.3 (Crossover)**.** A *crossover* is a change in a system's qualitative behavior as a dimensionless ratio Π passes from values much smaller than one to values much larger than one, such that behavior in each limit is governed by a different dominant effect.

Two kinds of crossover must be kept apart. In a *smooth* crossover, the behavior changes gradually and no particular value of Π is singled out; the statement "the crossover occurs at $\Pi \approx 1$" is only an order-of-magnitude statement, because the numerical factors in the definition of Π are conventional (a radius or a diameter, a frequency or an angular frequency). In a *sharp* transition, a bifurcation in a dynamical system or a phase transition in a large system, some quantity changes non-analytically at a definite critical value Π^*, and that value is a genuine prediction, including its factors of order one. Sharp

transitions are rounded into crossovers in finite systems and in the presence of noise, but their location in the idealized limit is well defined.

The distinction matters because the two kinds of claim have different content. A claim that some behavior appears "when the two effects balance" is a claim about a crossover, true almost by definition for a smooth one. A claim that it appears at a stated Π^* with stated signatures, and moves in a stated way when other parameters change, is a claim about a transition, and it can fail.

5.3 The Reynolds Number as Paradigm

The model of a productive magnitude balance is the Reynolds number of a flow,

$$\mathrm{Re} = \frac{\rho U L}{\mu} = \frac{U L}{\nu}, \tag{5.2}$$

where U is a characteristic speed, L a characteristic length, ρ the density, μ the dynamic viscosity, and $\nu = \mu/\rho$ the kinematic viscosity. It estimates the ratio of inertial to viscous forces. At low Reynolds number, viscosity damps disturbances and flow is smooth and orderly; at high Reynolds number, inertia carries disturbances along and flow becomes turbulent.

Reynolds (1883) established this with experiments in which a thin stream of dye was introduced into water flowing through glass pipes. At low speeds the dye stayed in a straight thread; above a critical condition it broke into eddies and filled the pipe. He showed that the condition depended on the combination of speed, pipe diameter, and viscosity in (5.2) and not on any one of them, and that the critical value depended on how undisturbed the water entering the pipe was: the more carefully the inlet was calmed, the higher the speed at which the flow stayed orderly. In pipes, the transition is found at Reynolds numbers of a few thousand.

The Reynolds number earns its place for three reasons that generalize. It predicts *similarity*: two flows with the same geometry and the same Reynolds number have the same dimensionless velocity field, which is why a scaled model in a wind tunnel can stand in for a full-sized aircraft (Barenblatt 1996). It locates a *transition* whose position, though sensitive to disturbance, is reproducible under stated conditions. And its value *outside* the transition region tells which approximations to use: at low Reynolds number the equations of motion can be linearized in the viscous limit, at high Reynolds number the viscous term matters only in thin boundary layers. A magnitude balance that does all three things is a scientific variable; one that only says "two effects compete" is a slogan.

5.4 Competing Energy Scales

Many transitions in physics are balances between two energies. A collection of magnetic moments orders when the energy of the interaction that aligns neighboring moments exceeds the thermal energy that scrambles them; the ordering temperature is set by the interaction energy, up to a factor of order one that depends on the lattice. Electrons in a solid form a metal when their kinetic energy, measured by the width of the band they occupy, is large compared with the Coulomb energy of their mutual repulsion, and they can freeze into correlated states when the balance reverses.

Chapter 8 tells the most striking recent instance. In twisted bilayer graphene, the twist between the layers sets a moiré length (Chapter 2), and near a twist of about one degree the lowest electronic bands become nearly flat: the electrons' kinetic energy becomes small compared with their interaction energy, and correlated insulating and superconducting states appear (Cao, Fatemi, Fang, et al. 2018; Cao, Fatemi, Demir, et al. 2018; Bistritzer and MacDonald 2011). Three relations are in play there, and they are of different kinds: the twist angle is a geometric relation, the moiré length is a relationally generated scale, and the ratio of interaction to kinetic energy is a magnitude balance. Keeping them apart is what makes the physics predictive.

5.5 Coupling Versus Heterogeneity

The balance that matters most for the rest of this book is the one between coupling and diversity in a population of oscillators.

Start with two. Let two phase oscillators with natural frequencies ω_1 and ω_2 interact through a coupling of strength K, as in the Kuramoto model of Chapter 1 with $N = 2$. The phase difference $\phi = \theta_1 - \theta_2$ obeys

$$\dot{\phi} = \Delta\omega - K \sin\phi, \qquad \Delta\omega = \omega_1 - \omega_2, \tag{5.3}$$

an equation often named after Adler, who derived it for an oscillator locked to an external signal (Pikovsky et al. 2001). If $|\Delta\omega| \le K$, the equation has fixed points with $\sin\phi^* = \Delta\omega/K$, one of them stable: the oscillators lock, running at a common frequency with a fixed phase difference. If $|\Delta\omega| > K$, there is no fixed point, and the phase difference drifts, slipping through a full cycle in a time

$$T_{\text{slip}} = \int_0^{2\pi} \frac{\mathrm{d}\phi}{\Delta\omega - K\sin\phi} = \frac{2\pi}{\sqrt{\Delta\omega^2 - K^2}}. \tag{5.4}$$

An observer listening to the two oscillators hears them beat at $\Omega = \sqrt{\Delta\omega^2 - K^2}$, slower than the difference $|\Delta\omega|$ of their natural frequencies: coupling pulls the frequencies together before it locks them (Figure 5.1a). The relevant balance is the ratio $K/|\Delta\omega|$, and the transition at $K/|\Delta\omega| = 1$ is sharp: it is a saddle-node bifurcation of (5.3).

For a large population, let the natural frequencies be drawn from a distribution $g(\omega)$, symmetric and single-peaked about its center, and couple every oscillator to every other with strength K/N. Kuramoto (1984) showed that in the limit of many oscillators, incoherence gives way to partial synchronization at the critical coupling

$$K_c = \frac{2}{\pi\, g(\omega_0)}, \tag{5.5}$$

where ω_0 is the center of the distribution. For the Lorentzian distribution with half-width γ, $g(\omega_0) = 1/(\pi\gamma)$, so $K_c = 2\gamma$, and above threshold the order parameter, which measures how strongly the population's phases are aligned ($r = 0$ for incoherence, $r = 1$ for perfect alignment), grows as

$$r = \sqrt{1 - \frac{K_c}{K}} \qquad (K \geq K_c) \tag{5.6}$$

(Kuramoto 1984; Strogatz 2000; Acebrón et al. 2005). The balance is between the coupling K and the spread γ of natural frequencies (Figure 5.1b). Chapter 9 develops the theory, and Appendix E gives the derivation of (5.5).

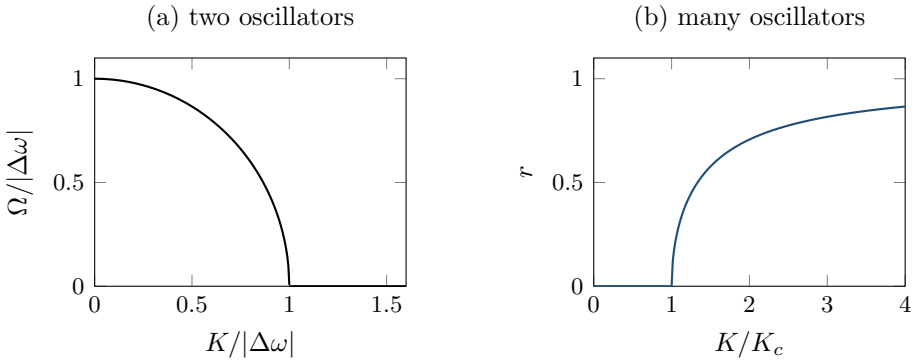


Figure 5.1: Two magnitude balances with sharp transitions. (a) The observed beat frequency Ω of two coupled oscillators falls as coupling grows and vanishes at locking, $K = |\Delta\omega|$. (b) The Kuramoto order parameter for a Lorentzian population becomes nonzero at $K = K_c = 2\gamma$ and follows (5.6) above it.

Two features of this balance matter later. First, the transition is sharp only in an idealized limit (two noiseless oscillators, or infinitely many), and it is rounded into a crossover by noise and by finite population size (Acebrón et al. 2005). Second, the quantity that changes at the transition is itself relational: the phase difference ϕ for two oscillators, the degree of alignment r for many. The magnitude balance decides *whether* a relational variable settles, and the relational variable is what settles.

The electric fish of Part V carries both kinds of relation at once. At each patch of its skin, the ratio $m = A_e/A_s$ of a neighbor's field to its own is a magnitude balance, and it sets how deep the amplitude and phase modulations

of the beat are. The frequency difference $\Delta f = f_e - f_s$ is a detuning, and its *sign* decides in which direction the fish shifts its own frequency (Chapter 20). The two relations do different jobs in the same animal, and no single notion of "the right relation" covers both.

5.6 Why "the Optimum Is in the Middle" Is Usually Too Weak

Claims of the form "the system works best at a balance between order and disorder" or "the interesting behavior lies between too much coupling and too little" are common in writing about brains, organisms, and societies. They are usually true and usually empty, for four reasons.

The first is that nearly every smooth crossover has an intermediate regime. If one effect dominates at small Π and another at large Π, there is a range in between where neither does, and any property that depends on both will look different there. Finding such a regime confirms almost nothing.

The second is that without naming the two magnitudes, the claim has no content: "order" and "disorder" are not quantities with dimensions, and their ratio is not defined.

The third is that without a predicted location, the claim cannot fail. A statement that a property peaks somewhere between two extremes is consistent with a peak anywhere.

The fourth is that intermediate regimes are not optima of anything in particular. A system in a crossover region is doing something different from what it does in either limit, and whether that something is better for any purpose is a separate question requiring its own evidence.

Principle 5.4 (Minimum content of a balance claim)**.** A claim that some behavior appears at a balance between two effects has empirical content only if it states:

1. the two magnitudes and their common dimension;
2. the dimensionless ratio Π they form, with its numerical conventions;
3. the predicted critical value Π^*, or the claim that the crossover is smooth, together with how Π^* depends on the system's other parameters;
4. the signature by which the crossing is detected, such as a bifurcation, critical slowing down, a divergent response, or hysteresis;
5. what the behavior is on each side, and which of the two sides the claimed benefit belongs to.

The Reynolds number satisfies all five. So does the locking of two oscillators: the magnitudes are the coupling and the detuning, both frequencies; the ratio is $K/|\Delta\omega|$; the critical value is one; the signature is the vanishing of the beat frequency (5.4); and the behavior on each side is locked or drifting. Chapter 8 shows a claim about twisted graphene that satisfies them and the popular gloss

on it that does not. Chapter 54 examines the "edge of chaos," a proposal that computation is optimized at a balance between order and chaos, and what became of it when its predictions were tested.

5.7 Audit

Canonical phenomenon

Water in a pipe flows smoothly at low speed and turbulently at high speed, and the change is governed by a single ratio; two coupled oscillators beat when their frequencies differ by more than their coupling and lock when they differ by less.

Minimal model

Buckingham's theorem (Theorem 5.1); the Reynolds number (5.2); the two-oscillator phase equation (5.3) with its slip frequency (5.4); the Kuramoto threshold (5.5).

What is relational here?

A ratio of two magnitudes with the same dimension: inertial to viscous force, coupling to detuning, coupling to frequency spread. In the oscillator examples, the ratio decides whether a second relational variable, the phase difference, settles.

What physically reads it?

Nothing reads the ratio. The dynamics depend on it because the equations of motion contain both magnitudes, and an experimenter reads it by measuring each and dividing. The phase difference that settles above the locking threshold is available to anything coupled to both oscillators.

What existing field explains it?

Dimensional analysis and similarity (Buckingham 1914; Barenblatt 1996); fluid mechanics (Reynolds 1883); bifurcation theory and synchronization (Pikovsky et al. 2001; Kuramoto 1984; Strogatz 2000). The chapter contains nothing these fields do not.

What remains unexplained?

None. The chapter's contribution is Principle 5.4, a test that later chapters apply to claims about brains and minds.

A discriminating experiment

For any published claim that some neural or behavioral property is maximized at a balance between two tendencies, extract the ratio Π and the predicted Π^*, then vary one of the two magnitudes while holding the other fixed. The claim survives if the property tracks Π and peaks at the predicted value; it fails if the property tracks one magnitude alone, or peaks at a value the claim did not predict.

5.8 Notes and References

Buckingham (1914) is the original statement of the Π theorem; Barenblatt (1996) is a deep modern treatment of dimensional analysis, similarity, and the cases in which it fails. Reynolds (1883) is still worth reading as an example of an experiment designed around a dimensionless ratio. For the two-oscillator phase equation and its bifurcation, see Pikovsky et al. (2001); for the Kuramoto model, Kuramoto (1984), the review by Strogatz (2000), and the survey by Acebrón et al. (2005). The magic-angle physics mentioned here is developed, with references, in Chapter 8.

5.9 Exercises

Exercise 5.1. The drag force F on a sphere of diameter D moving at speed U through a fluid of density ρ and viscosity μ depends on these four quantities. Use Theorem 5.1 to show that $F/(\rho U^2 D^2)$ is a function of the Reynolds number alone. In the limit of very low Reynolds number the drag cannot depend on ρ; what form must the function then take?

Exercise 5.2. Taking the kinematic viscosity of water as $1.0 \times 10^{-6}\,\mathrm{m^2/s}$, compute the Reynolds number for water flowing at $0.1\,\mathrm{m/s}$ through a pipe of diameter $1\,\mathrm{cm}$, and the speed at which the same pipe reaches a Reynolds number of 3000. Recompute both using the pipe's radius as the characteristic length, and explain what the factor of two implies about statements that a crossover occurs "at $\Pi \approx 1$."

Exercise 5.3. For (5.3), find the fixed points when $|\Delta\omega| < K$ and determine which is stable. Derive (5.4) and plot the observed beat frequency Ω against K for fixed $\Delta\omega$. Near the locking threshold, how does the slip period diverge?

Exercise 5.4. Simulate the Kuramoto model with $N = 2000$ oscillators whose natural frequencies are drawn from a Lorentzian of half-width $\gamma = 1$. Measure the long-time order parameter for K from 0 to 6 and compare it with (5.6). Repeat with $N = 50$ and describe how the transition is rounded.

Exercise 5.5. In the two-oscillator model, the ratio $K/|\Delta\omega|$ and the phase difference ϕ are both relational. Classify each as a magnitude balance, a detuning, or a relative phase. State which of them a device receiving only the two oscillators' output signals can read directly, and what it would have to measure to infer the other.

Exercise 5.6. Find a published claim, in any field you know, that some property is maximized between two extremes. Apply Principle 5.4 item by item, state which conditions the claim meets, and design the measurement that would test it.

6

Resonance

Push a child on a swing in time with the swing and small pushes build a large motion; push at a different rhythm and the pushes work against each other. Resonance is the second of the special relations that Part II separates. It relates the frequency of a drive to a frequency that belongs to the driven system.

The thesis of the chapter has three parts. Resonance is a peak in a system's response when a drive's frequency comes within a damping-set width of a natural frequency; it varies smoothly with the detuning, and it occurs in linear, passive systems. Every resonator is therefore a reader of a relation, the relation between an input frequency and its own, and the sharpness of its reading is bought with slowness. And resonance must be kept apart from commensurability, which concerns rational ratios between frequencies and produces locking only in nonlinear systems. The two meet in nonlinear and parametric resonance, and the chapter ends by drawing the line between them.

6.1 The Driven Oscillator

A mass m on a spring of stiffness k, with viscous damping b and a periodic force $F_0 \cos \omega t$, obeys

$$\ddot{x} + \gamma \dot{x} + \omega_0^2 x = \frac{F_0}{m} \cos \omega t, \qquad \gamma = \frac{b}{m}, \quad \omega_0^2 = \frac{k}{m}. \tag{6.1}$$

The equation is linear, so its steady-state response is a sinusoid at the driving frequency. Writing $x = \mathrm{Re}\{\tilde{X} \mathrm{e}^{\mathrm{i}\omega t}\}$ and substituting gives

$$\tilde{X}(\omega) = \frac{F_0/m}{\omega_0^2 - \omega^2 + \mathrm{i}\gamma\omega}, \tag{6.2}$$

so that $x(t) = X(\omega)\cos\big(\omega t - \delta(\omega)\big)$ with

$$X(\omega) = \frac{F_0/m}{\sqrt{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}}, \qquad \tan\delta(\omega) = \frac{\gamma\omega}{\omega_0^2 - \omega^2}. \tag{6.3}$$

Any transient set off by the initial conditions decays as $e^{-\gamma t/2}$ and leaves this steady state behind (Landau and Lifshitz 1976).

6.2 Natural Frequency

Without drive and damping, the oscillator swings at its natural frequency $\omega_0 = \sqrt{k/m}$. With damping and no drive, it rings down at the slightly lower frequency $\omega_d = \sqrt{\omega_0^2 - \gamma^2/4}$. Both frequencies belong to the oscillator alone: they are fixed by its mass, stiffness, and damping, and they are there whether or not anything drives it.

This is the feature that makes resonance a relation of a particular kind. Of the two frequencies it relates, one belongs to the outside world and one belongs to the system that responds. The resonator brings its own relatum.

6.3 The Resonance Curve

The amplitude (6.3) is small for slow drives, where it approaches the static deflection $X(0) = F_0/k$, and small again for fast drives, where inertia prevents the mass from following. In between it peaks. Setting the derivative of the denominator to zero places the peak at

$$\omega_r = \sqrt{\omega_0^2 - \gamma^2/2}, \tag{6.4}$$

which exists when $\gamma < \sqrt{2}\,\omega_0$ and approaches ω_0 as damping becomes light. The phase lag rises from zero through $\pi/2$, reached exactly at $\omega = \omega_0$, toward π: below resonance the mass moves with the force, above it the mass moves against it. Figure 6.1 shows the curves for three values of damping.

6.4 Damping

The steady state is a balance between the work done by the drive and the energy dissipated by damping. The time-averaged power absorbed from the drive is

$$P(\omega) = \frac{F_0^2}{2m}\,\frac{\gamma\omega^2}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}. \tag{6.5}$$

Near resonance, where $\omega_0^2 - \omega^2 \approx 2\omega_0(\omega_0 - \omega)$, this becomes a Lorentzian,

$$P(\omega) \approx \frac{F_0^2}{8m}\,\frac{\gamma}{(\omega - \omega_0)^2 + (\gamma/2)^2}, \tag{6.6}$$

whose full width at half maximum is γ. Damping therefore sets the width of the frequency window within which the drive counts as "matching" the natural frequency. The relation that resonance responds to is the detuning $\omega - \omega_0$ *measured in units of* the damping rate: a ratio of two frequencies of the kind treated in Chapter 5.

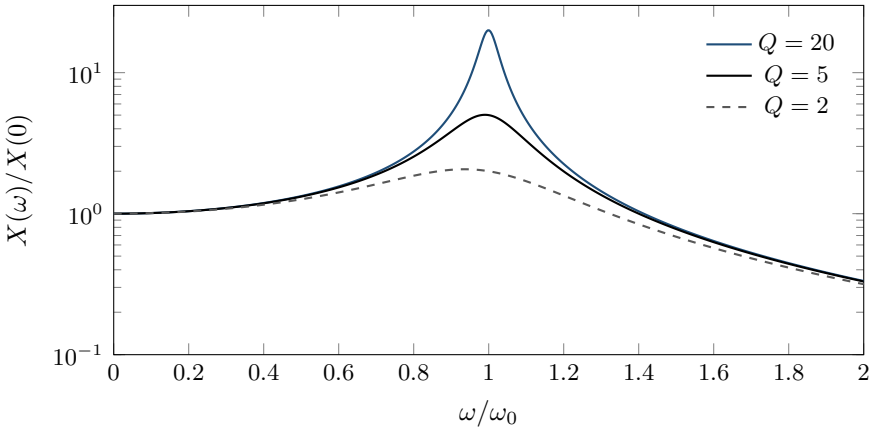


Figure 6.1: Resonance curves of the driven damped oscillator (6.3), normalized by the static deflection, for three quality factors. The peak height is close to Q and its width close to ω_0/Q.

6.5 The Quality Factor

The sharpness of a resonance is summarized by the quality factor.

Definition 6.1 (Quality factor)**.** The *quality factor* of a damped oscillator is $Q = \omega_0/\gamma$.

Three properties follow from (6.3) and (6.6). At $\omega = \omega_0$ the amplitude is $X(\omega_0) = F_0/(m\gamma\omega_0)$, so the response is amplified relative to the static deflection by exactly $X(\omega_0)/X(0) = Q$. The half-power bandwidth is $\gamma = \omega_0/Q$. And the stored energy of the free oscillator decays as $\mathrm{e}^{-\gamma t}$, so that for light damping Q equals 2π times the ratio of the energy stored to the energy lost per cycle, and the amplitude falls by a factor of e in about Q/π cycles.

The last two properties together express a trade-off that constrains every resonant reader. A resonator with a narrow bandwidth distinguishes frequencies finely, and it takes a long time to build up its response or to let an old one die away. A resonator that responds quickly has a broad bandwidth. The product of the bandwidth ω_0/Q and the response time $2Q/\omega_0$ is fixed at two. This is the same trade-off that decided, in Chapter 3, whether a listener hears two tones or one beating tone, and it is a theorem about all linear filters (Appendix B). A resonator cannot be both fast and selective.

6.6 Nonlinear Resonance

Real springs stiffen or soften at large displacements. Adding a cubic term gives the Duffing equation,

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x + \beta x^3 = \frac{F_0}{m}\cos\omega t. \tag{6.7}$$

Its free oscillations have a frequency that depends on their amplitude A; to first order in β,

$$\omega(A) \approx \omega_0 + \frac{3\beta A^2}{8\omega_0}, \tag{6.8}$$

rising with amplitude for a stiffening spring ($\beta > 0$) and falling for a softening one (Landau and Lifshitz 1976; Nayfeh and Mook 1979). Seeking a driven response of the form $x \approx A\cos(\omega t - \delta)$ and keeping only terms at the driving frequency (the method of harmonic balance) gives the frequency-response relation

$$\left[\left(\omega_0^2 - \omega^2 + \tfrac{3}{4}\beta A^2\right)^2 + \gamma^2\omega^2\right]A^2 = \left(\frac{F_0}{m}\right)^2. \tag{6.9}$$

The resonance peak bends toward the direction in which the natural frequency moves with amplitude. When it bends far enough, (6.9) has three solutions for A over a range of driving frequencies, two of them stable. Sweeping the drive frequency upward and then downward traces different branches, and the amplitude jumps between them at different frequencies on the way up and on the way down (Nayfeh and Mook 1979). The response now depends on history.

The cubic term also creates resonances at other driving frequencies. Because the nonlinearity generates harmonics (Chapter 4), a drive near $\omega_0/3$ has a third harmonic near ω_0 and can excite the oscillator, a *superharmonic* resonance; and a drive near $3\omega_0$ can excite a response at a third of its frequency, a *subharmonic* resonance (Nayfeh and Mook 1979). At these secondary resonances the drive and the response stand in a rational ratio, 1:3 or 3:1. Resonance and commensurability overlap here, and Chapter 7 approaches the overlap from the other side.

6.7 Parametric Resonance

A swing can also be pumped without pushing. A child who stands at the ends of the arc and crouches at the bottom changes the effective length of the pendulum twice in every swing, and the swing grows. The drive enters as a periodic change in a parameter of the oscillator rather than as an external force. The standard model is the Mathieu equation,

$$\ddot{x} + \gamma\dot{x} + \omega_0^2\left[1 + h\cos(\omega_p t)\right]x = 0, \tag{6.10}$$

with a modulation of relative depth h at the pump frequency ω_p.

Without damping, the rest state becomes unstable in zones around $\omega_p = 2\omega_0/n$ for $n = 1, 2, \ldots$, and the widest zone, around $\omega_p = 2\omega_0$, extends to first order in h over $|\omega_p - 2\omega_0| < h\omega_0/2$ (Landau and Lifshitz 1976). Inside it the oscillation grows exponentially at *half* the pump frequency. Damping raises a threshold: averaging (6.10) over the fast oscillation at $\omega_p = 2\omega_0$ gives a growth rate $h\omega_0/4 - \gamma/2$, so the modulation must exceed $h_c = 2/Q$ before anything grows.

Example 6.2 (Faraday waves). Faraday (1831) observed the most famous instance of parametric resonance. A layer of liquid on a plate vibrated vertically develops a pattern of standing ridges on its surface, and the ridges oscillate at half the frequency of the plate. The vertical vibration modulates the effective gravity felt by the surface waves, which is a parametric drive, and the half-frequency response is its signature.

In parametric resonance the relation that matters is the ratio of pump frequency to natural frequency, and the zones of instability sit at rational values of it, $2/n$. Plotted in the plane of pump frequency and modulation depth, the zones are tongues that widen with h, a shape that returns in Chapter 7 as the Arnold tongues of mode locking.

6.8 Neural Resonance

Nerve cells resonate. Hutcheon and Yarom (2000) showed that many neurons respond to oscillating input current with a voltage response that peaks at a preferred frequency, and traced the peak to a combination of two filters: the passive membrane, whose capacitance and resistance attenuate fast inputs, and slow voltage-dependent currents that oppose changes in membrane voltage and attenuate slow inputs. Low-pass and high-pass filtering together produce a band-pass response. They also emphasized a distinction that matters here: a neuron can resonate, preferring inputs at a certain frequency, without oscillating on its own.

Izhikevich (2001) drew the consequence for how such a neuron reads its input. A neuron at rest near the threshold of oscillation responds to a brief input with a damped subthreshold oscillation. A second input arriving one oscillation period later adds to the first; one arriving half a period later cancels it. A resonator neuron therefore responds selectively to pairs of inputs whose *interval* matches its intrinsic period, where an integrator neuron responds to the *number* of inputs arriving within a time window. Izhikevich showed that in such a neuron even an inhibitory pulse, suitably timed, can trigger a spike. The linear core of this behavior can be written down exactly.

Proposition 6.3 (A resonator reads intervals)**.** *Let a linear oscillator* (6.1) *with* $\gamma < 2\omega_0$ *and no periodic force receive two equal impulses separated by a time* Δt*. After the second impulse, the oscillation amplitude is proportional to*

$$\left|1 + \mathrm{e}^{-\gamma \Delta t/2}\, \mathrm{e}^{\mathrm{i}\omega_d \Delta t}\right|, \tag{6.11}$$

which is largest when Δt *is a whole number of periods* $2\pi/\omega_d$ *and smallest when it is an odd number of half periods.*

Proof. The response to a unit impulse at time zero is $\mathrm{Im}\{\mathrm{e}^{(\mathrm{i}\omega_d - \gamma/2)t}\}/\omega_d$ for $t \geq 0$. By linearity the response to the pair is the sum of this function and its

copy delayed by Δt. At a time $t' \geq 0$ after the second impulse the sum is

$$\frac{1}{\omega_d}\,\mathrm{Im}\Big\{\mathrm{e}^{(\mathrm{i}\omega_d-\gamma/2)t'}\big(1+\mathrm{e}^{(\mathrm{i}\omega_d-\gamma/2)\Delta t}\big)\Big\},$$

a damped oscillation whose amplitude is proportional to $|1+\mathrm{e}^{(\mathrm{i}\omega_d-\gamma/2)\Delta t}|$, which is (6.11). □

A resonant neuron reads a temporal relation between its inputs, their interval measured against its own period. The same neuron with its slow currents blocked, and its band-pass response gone, would read their sum. Which variable of its input the cell makes explicit is set by its membrane, in the sense of Chapter 4.

6.9 The Resonator as a Reader

A bank of resonators with different natural frequencies is a spectrum analyzer: each member responds to the part of the input near its own frequency, and the pattern of their responses reports the input's spectrum. Helmholtz (1885) proposed that the ear works this way, as a set of resonators each responding to its own tone. The modern picture of the cochlea is richer, an active and nonlinear filter bank whose tuning depends on the motion of its own cells (Robles and Ruggero 2001), but the principle survives: a tuned element reports how its input relates to its own frequency.

The resonator joins the local oscillator and the lock-in reference of Chapter 4 and the reference grating of Chapter 2 as a reader that supplies one relatum and reads the relation of the input to it. It differs from them in needing no power source and no clock. Its reference is built into its mass and stiffness, or into the capacitance and currents of a membrane. The electric fish of Part V does something related and more active: it generates its own reference signal, its electric discharge, and reads its surroundings through the way they alter it (Chapter 19).

6.10 Why Resonance Is Not Commensurability

Resonance and commensurability both concern frequencies, both produce striking effects when frequencies stand in special relations, and both are called "resonance" in some literatures. Celestial mechanics, for instance, speaks of orbital resonances when the periods of two bodies stand in a ratio of small integers, which in the terms of this book is a commensurability. The overlap in vocabulary hides a difference in structure, set out in Table 6.1.

The difference shows up in two limiting cases. A linear oscillator driven at exactly $\frac{3}{2}\omega_0$ shows no special response at all, even though the ratio of drive to natural frequency is a simple rational number: commensurability

	Magnitude balance	**Resonance**	**Commensurability**
Relation	$A/B \sim 1$	$\omega \approx \omega_0$	$\omega_1/\omega_2 = p/q$
What is compared	Two magnitudes of the same dimension	A drive frequency and a natural frequency of the responder	Two frequencies, either of which may be intrinsic
Minimal system	Any system with two competing effects	A linear damped oscillator	A nonlinear self-sustained oscillator, forced or coupled
Signature	Crossover or bifurcation at Π^*	Response peak of width ω_0/Q	Locking plateaus, Arnold tongues, devil's staircase
Dependence on detuning	Through the ratio	Smooth, Lorentzian near the peak	Different at rational and irrational ratios
Examples	Reynolds number; coupling against frequency spread	Swing, tuning fork, cochlear filter	Huygens's clocks; periodically stimulated heart cells

Table 6.1: Three special relations, often run together as "the right relation." Chapters 5 and 7 treat the first and third columns.

without resonance. A linear oscillator driven at a frequency whose ratio to ω_0 is irrational but lies within ω_0/Q of one responds almost as strongly as one driven at exactly ω_0: resonance without commensurability. A linear system is blind to whether a frequency ratio is rational. Only nonlinear systems, and in particular self-sustained oscillators that can adjust their own phase, respond to rationality as such, by locking; Chapter 7 develops the theory.

Where the two meet, in the secondary resonances of the Duffing oscillator and the instability tongues of parametric resonance, a nonlinearity or a parametric drive is always present, and the response at the rational ratio is a resonance of one frequency with a harmonic or subharmonic of the other. Keeping the two concepts apart is what allows the combined cases to be analyzed at all.

6.11 Audit

Canonical phenomenon

A swing grows when pushed in time with its natural motion; a vibrated liquid surface forms ridges that oscillate at half the drive frequency; some neurons fire only when inputs arrive at intervals matching their intrinsic period.

Minimal model

The driven damped oscillator (6.1) with its amplitude and phase (6.3); the quality factor; the Duffing (6.7) and Mathieu (6.10) equations; the two-impulse response (6.11).

What is relational here?

The detuning $\omega - \omega_0$ measured in units of the damping rate; for parametric resonance, the ratio of pump to natural frequency; for the resonant neuron, the interval between inputs measured against the intrinsic period.

What physically reads it?

The resonator itself. It supplies one relatum, its natural frequency, and its amplitude makes explicit how the input frequency relates to it. Its selectivity and speed are traded against each other through Q.

What existing field explains it?

Classical mechanics and the theory of oscillations (Landau and Lifshitz 1976; Nayfeh and Mook 1979); Faraday's experiments (Faraday 1831); auditory physiology (Helmholtz 1885; Robles and Ruggero 2001); cellular neurophysiology (Hutcheon and Yarom 2000; Izhikevich 2001). Nothing in the chapter goes beyond them.

What remains unexplained?

None. The chapter's contribution is the separation in Table 6.1, which later chapters rely on when relational claims about neural rhythms are examined (Chapter 46).

A discriminating experiment

Drive a lightly damped mechanical or electrical oscillator at a sequence of frequencies with rational and irrational ratios to its natural frequency, all within its bandwidth. The analysis predicts a smooth response curve with no feature at the rational ratios. Then add a cubic stiffness and repeat near $\omega_0/3$ and $3\omega_0$: secondary resonances should appear at those ratios and nowhere else in the swept range.

6.12 Notes and References

Landau and Lifshitz (1976) treats forced oscillations, anharmonic oscillations, and parametric resonance with admirable economy, including the width of the principal instability zone. Nayfeh and Mook (1979) is the standard reference for nonlinear oscillations, the Duffing equation, jump phenomena, and secondary resonances. Faraday (1831) is the original report of the half-frequency surface waves. On neural resonance, see Hutcheon and Yarom (2000) for mechanisms and the distinction between resonance and oscillation, and Izhikevich (2001) for resonate-and-fire neurons and their selectivity to input intervals. Helmholtz

(1885) gives the resonance theory of hearing; Robles and Ruggero (2001) reviews the cochlea as it is now understood.

6.13 Exercises

Exercise 6.1. Derive (6.2) and (6.3). Show that the phase lag equals $\pi/2$ at $\omega = \omega_0$ for every value of the damping.

Exercise 6.2. Derive (6.4), show that $X(\omega_0)/X(0) = Q$, and derive the Lorentzian approximation (6.6) and its half-power width. For $Q = 50$, how many cycles does the free oscillator take for its amplitude to fall by a factor of ten?

Exercise 6.3. Derive (6.9) by harmonic balance. For $\omega_0 = 1$, $\gamma = 0.05$, $\beta = 0.1$, and $F_0/m = 0.1$ (illustrative values), solve it numerically for A as a function of ω, identify the range with three solutions, and find the frequencies at which the amplitude jumps during an upward and a downward sweep. Confirm by integrating (6.7) directly.

Exercise 6.4. Write the solution of (6.10) with $\omega_p = 2\omega_0$ as $x = a(t)\cos\omega_0 t + b(t)\sin\omega_0 t$ with slowly varying a and b, average over the fast oscillation, and show that the growth rate is $h\omega_0/4 - \gamma/2$. Map the instability region in the plane of ω_p/ω_0 and h numerically for $Q = 20$ and compare its shape with the tongue described in the text.

Exercise 6.5. For $Q = 10$, plot (6.11) against $\Delta t\,\omega_d/2\pi$ from 0 to 3. A neuron modeled by this oscillator fires when its amplitude exceeds a threshold set at 1.5 times the response to a single impulse. Find the ranges of intervals that make it fire, and describe what happens to these ranges as Q decreases.

Exercise 6.6. Classify each of the following as a magnitude balance, a resonance, a commensurability, or a combination, and justify the classification using Table 6.1: a swing pushed on every second swing; two metronomes on a movable board falling into step; a radio's tuned input circuit; a wine glass shattered by a sustained sung note; a flexible cylinder in a steady current whose vortex-shedding frequency follows the cylinder's own vibration over a range of flow speeds.

7

Commensurability

Two periodic processes are *commensurate* when their frequencies stand in a ratio of whole numbers. Taken as a fact about two numbers, commensurability is fragile: the rational numbers have measure zero, and an arbitrarily small change in either frequency destroys the relation. Taken as a fact about coupled oscillators, it is robust: a self-sustained oscillator driven near a rational multiple of its own frequency adjusts its rhythm and holds the ratio fixed over a whole interval of detunings. The thesis of this chapter is that locking to a rational ratio is a dynamical phenomenon with a precise and general theory, the theory of circle maps, in which the relational observable is the rotation number and the regions of locking are the Arnold tongues. That theory separates commensurability from resonance and from magnitude balance, and it supplies the specific signatures against which any claim about commensurate or incommensurate rhythms, in physics or in the brain, has to be tested.

7.1 Rational Ratios and the Torus

Measure the phase of each of two oscillators in cycles, so that $\theta_i \in [0, 1)$, and suppose they run independently at frequencies f_1 and f_2:

$$\dot{\theta}_1 = f_1, \qquad \dot{\theta}_2 = f_2.$$

The pair (θ_1, θ_2) moves on a torus, and its trajectory is a straight line of slope f_2/f_1 wrapped around it.

Definition 7.1 (Commensurability)**.** Two frequencies $f_1, f_2 > 0$ are *commensurate* if $f_1/f_2 = p/q$ for coprime positive integers p and q, and *incommensurate* otherwise.

Theorem 7.2 (Closed and dense orbits)**.** *If $f_1/f_2 = p/q$ in lowest terms, every trajectory of the uncoupled flow is closed and returns to its starting point after time $p/f_1 = q/f_2$, having wound p times around one cycle of the torus and q times around the other. If f_1/f_2 is irrational, every trajectory is dense on the torus.*

The theorem is classical (Arnold 1983), and its lesson for us is statistical. Commensurability of two independent oscillators is a coincidence of measure zero. A pair of clocks built independently will not keep a rational ratio, however carefully each is made, because the smallest drift in either frequency moves the ratio off any rational value. When a rational ratio persists in nature, coupling is doing work. The rest of the chapter is about that work: the dynamics that capture a ratio and hold it.

7.2 Locking: The Adler Equation

The simplest case is a 1:1 ratio. Let an oscillator with natural angular frequency ω_0 be driven weakly at angular frequency ω, and let ϕ be the phase of the oscillator relative to the drive, in radians. For weak coupling the relative phase obeys

$$\dot{\phi} = \nu - K \sin \phi, \qquad \nu = \omega_0 - \omega, \tag{7.1}$$

where ν is the detuning and K measures the strength of the drive's influence. Adler derived this equation in 1946 for electronic oscillators injected with an external signal (R. Adler 1946); the same equation governs two weakly coupled phase oscillators, with ϕ their phase difference (Pikovsky et al. 2001, ch. 7).

Proposition 7.3 (Locking range and phase slips)**.** *If $|\nu| \le K$, equation (7.1) has a stable fixed point $\phi^* = \arcsin(\nu/K)$, and the oscillator runs at exactly the drive frequency. If $|\nu| > K$, the relative phase increases or decreases without bound, one full slip taking time*

$$T_{\text{slip}} = \int_0^{2\pi} \frac{\mathrm{d}\phi}{\nu - K \sin \phi} = \frac{2\pi}{\sqrt{\nu^2 - K^2}}, \tag{7.2}$$

so that the observed frequency difference is $\operatorname{sgn}(\nu)\sqrt{\nu^2 - K^2}$.

The locking region $|\nu| \le K$ is a wedge in the plane of detuning and coupling, the simplest of the Arnold tongues introduced below. Outside it, coupling still acts: the observed frequency difference is smaller than the natural detuning, and it approaches zero as the square root of the distance to the edge of the wedge (Figure 7.1). Near the edge the relative phase lingers for a long time close to $\phi = \pi/2$, where the fixed point used to be, and then slips quickly through a full cycle. That lingering is the elementary mechanism of metastability, and Chapter 11 builds on it.

7.3 The Circle Map

Locking at other ratios is easiest to see in a discrete-time model. Suppose an oscillator of natural period T_0 receives a brief stimulus every T seconds, and each stimulus shifts its phase by an amount $\Delta(\theta)$ that depends on the

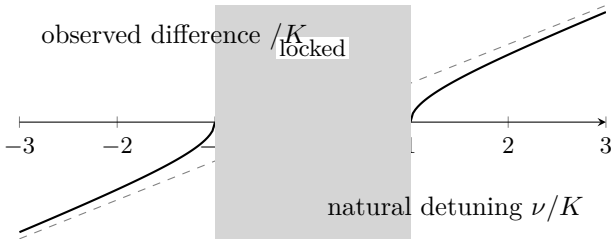


Figure 7.1: Frequency locking in the Adler equation (7.1). Inside the band $|\nu| \le K$ the oscillator runs at the drive frequency; outside it the observed frequency difference (solid) is pulled below the natural detuning (dashed) and vanishes as a square root at the edge.

phase θ at which the stimulus arrives. The function Δ is the oscillator's *phase response curve*. Measured in cycles, the phase just before successive stimuli obeys $\theta_{n+1} = \theta_n + \Omega + \Delta(\theta_n) \pmod 1$, with $\Omega = T/T_0$ the number of natural cycles that would elapse between stimuli in the absence of coupling (Glass and Mackey 1988). The standard model takes a sinusoidal phase response.

Model 7.4 (Sine circle map)**.** For $\Omega \in \mathbb{R}$ and $K \ge 0$,

$$\theta_{n+1} = \theta_n + \Omega - \frac{K}{2\pi}\sin(2\pi\theta_n) \pmod 1. \tag{7.3}$$

Its lift to the real line is $F(x) = x + \Omega - (K/2\pi)\sin 2\pi x$, which satisfies $F(x+1) = F(x) + 1$.

For $0 \le K < 1$ the lift is strictly increasing, since $F'(x) = 1 - K\cos 2\pi x > 0$, so the map is an orientation-preserving diffeomorphism of the circle. At $K = 1$ it remains a homeomorphism with a single inflection point where $F' = 0$; for $K > 1$ it folds back on itself and is no longer invertible. The variable that summarizes the long-run relation between oscillator and stimulus is the average advance of phase per stimulus.

Definition 7.5 (Rotation number)**.** The *rotation number* of the lift F is

$$\rho(\Omega, K) = \lim_{n\to\infty} \frac{F^n(x) - x}{n}. \tag{7.4}$$

Theorem 7.6 (Poincaré)**.** *For an orientation-preserving homeomorphism of the circle, the limit* (7.4) *exists and is independent of the initial point x. It is a rational number p/q if and only if the map has a periodic orbit of type p/q, that is, a point x with $F^q(x) = x + p$.*

The rotation number is a relational observable in the sense of Chapter 1. It is defined on the pair (oscillator, stimulus), it does not depend on where the origin of phase is placed, and it is the frequency ratio the pair actually realizes, as opposed to the ratio Ω of natural frequencies. Without coupling

the two coincide: $\rho(\Omega, 0) = \Omega$. With coupling they come apart, and the way they come apart is the whole theory of locking.

Proposition 7.7 (Plateaus)**.** *For $0 \le K \le 1$, $\rho(\Omega, K)$ is a continuous, nondecreasing function of Ω. For $0 < K \le 1$, each rational value p/q is taken on a closed interval of Ω of positive length, and each irrational value at a single point.*

The proposition is standard for families of circle diffeomorphisms of this kind (Arnold 1983; Pikovsky et al. 2001). Its content is that rational ratios are no longer fragile once the oscillators interact. A rational ratio of natural frequencies, a set of measure zero, becomes an interval of natural frequencies that all produce the same realized ratio.

7.4 Arnold Tongues

Plot the locking intervals in the plane of natural ratio Ω and coupling K.

Definition 7.8 (Arnold tongue)**.** The *Arnold tongue* of the rational p/q is the set $\{(\Omega, K) : \rho(\Omega, K) = p/q\}$.

Each tongue grows from the point $(\Omega, K) = (p/q, 0)$ and widens with coupling (Figure 7.2). The widths follow a strict hierarchy. The $0/1$ tongue is exactly $|\Omega| \le K/2\pi$, because a fixed point of (7.3) requires $\Omega = (K/2\pi)\sin 2\pi\theta$ for some θ; its width is K/π. The $1/2$ tongue has width close to $K^2/4\pi$ for small K, and in general the width of the p/q tongue shrinks like K^q as $K \to 0$ (Appendix F derives the first cases and gives a numerical recipe). Low denominators dominate. Between two tongues p/q and p'/q' with $|pq' - p'q| = 1$, the widest tongue lies at the mediant $(p + p')/(q + q')$, so the tongues are ordered by the Farey tree (Pikovsky et al. 2001).

As K approaches 1 the tongues widen until, at $K = 1$, they fill the whole line of Ω except for a set of measure zero. Jensen et al. (1983) showed that the complement of the locking intervals at this critical coupling is a Cantor set of fractal dimension $D \approx 0.87$, and that the exponent is the same for a large class of maps. For $K > 1$ the map is no longer invertible, the tongues overlap, a given Ω can support more than one stable ratio, and chaotic motion becomes possible (Glass and Mackey 1988).

7.5 The Devil's Staircase

At fixed coupling, the realized ratio as a function of the natural ratio is a staircase (Figure 7.3). Every rational has its own step, the steps with low denominators are the widest, and at $K = 1$ the steps occupy the whole interval while the function still climbs continuously from 0 to 1. Such a function is called a devil's staircase, and at $K = 1$ it is *complete*.

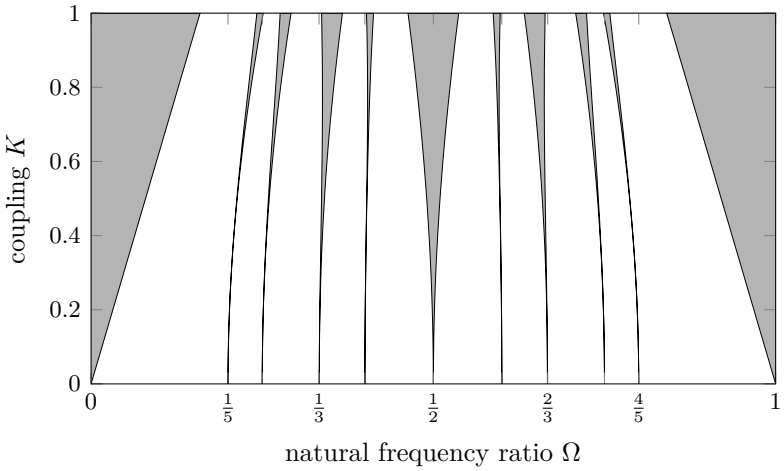


Figure 7.2: Arnold tongues of the sine circle map (7.3) for all ratios p/q with $q \le 5$, computed by locating the edges of each locking interval at 26 values of K. Each tongue is the set of natural ratios Ω that produce the realized ratio p/q. Unlabeled ticks mark $\frac{1}{4}$, $\frac{2}{5}$, $\frac{3}{5}$, and $\frac{3}{4}$.

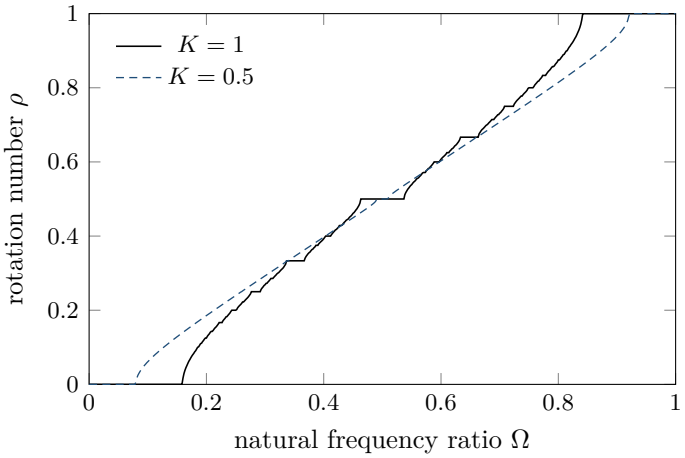


Figure 7.3: The devil's staircase: rotation number of the sine circle map against the natural frequency ratio, computed at 601 values of Ω for $K = 1$ (solid) and $K = 0.5$ (dashed). At $K = 1$ the plateaus fill the interval up to a set of measure zero.

Staircases of this kind are common in condensed matter, where the competing periods belong to a crystal lattice and to a modulation of positions, spins, or charges (Bak 1982; Bak 1986). In the Frenkel–Kontorova chain, for example, particles joined by springs sit in a periodic substrate potential, and the natural spacing of the chain competes with the period of the substrate. As a control parameter varies, the wavevector of the modulated structure locks onto rational fractions of the lattice period over finite intervals and passes through incommensurate values only at isolated points: a devil's staircase

in a material. The phases on the steps are called commensurate, the rest incommensurate, and the transitions between them are among the best-studied relational phenomena in physics.

7.6 Locking in Living Tissue

Biological oscillators lock in the same way.

Example 7.9 (Heart cells driven by a stimulator). Guevara et al. (1981) perturbed spontaneously beating aggregates of embryonic chick heart cells, first with single current pulses and then with periodic trains of pulses. From the responses to single pulses, delivered at different phases of the beat, they predicted the regular and irregular dynamics that periodic stimulation would produce, including period-doubling bifurcations, and observed them. The analysis used exactly the construction of Section 7.3: a map from the phase at one stimulus to the phase at the next, built from the measured phase response curve, whose rotation number organizes the locked rhythms (Glass and Mackey 1988). The prediction of the driven behavior came from the oscillator's response to single perturbations, which is the relational information the circle map needs and nothing more.

The same logic has been carried to human brain rhythms. Notbohm et al. (2016) stimulated participants with flickering light at a range of frequencies and intensities around each participant's intrinsic alpha frequency. Phase coupling between the flicker and the recorded oscillation grew with stimulation intensity and was stronger at frequencies close to the intrinsic frequency, the pattern an Arnold tongue predicts; only inside the tongue did rhythmic stimulation differ significantly from jittered stimulation, and single recordings showed intermittent phase locking. Rhythmic transcranial magnetic stimulation has likewise been reported to entrain local oscillations at their natural frequency (Thut et al. 2011). In each case the thing measured is a relational observable, the realized relation between an oscillation and its drive, and the evidence for locking is the shape of the region in which that relation is held fixed.

7.7 Three Special Relations

Part II distinguishes three relations that are often run together as "the right ratio." With the circle map in hand the distinction can be made sharp.

A damped linear oscillator driven at frequency ω responds at ω, whatever its natural frequency. Its natural frequency sets the amplitude of the response, which peaks near resonance, and never the ratio. Locking requires an oscillator with a rhythm of its own that coupling can capture, and it can occur at ratios such as 1:3 far from any resonance peak. Magnitude balance involves no frequencies at all: a Reynolds number near its critical value marks a change of

Relation	Condition	Signature	What it requires
Magnitude balance	$A/B \sim 1$	Crossover between regimes	Two competing scales
Resonance	$f \approx f_0$	Peak in response amplitude, width set by damping	A responder with a natural frequency; a linear system suffices
Commensurability (locking)	$f_1/f_2 = p/q$ held over an interval	Plateaus, tongues, staircase	A self-sustained oscillator and nonlinear coupling

Table 7.1: Three special relations. Chapter 5 treats the first, Chapter 6 the second, and this chapter the third.

regime in a flow that need not oscillate (Chapter 5). The twisted bilayer of Chapter 8 will turn out to belong to the first row, although its moiré pattern invites a reading in terms of the third. A theory that says a system works best "at the right relation" without saying which row it means makes no prediction that could fail.

7.8 Irrational Ratios and the Golden Mean

An irrational rotation number means quasiperiodic motion: the realized ratio never repeats, and the orbit on the circle is dense. Among irrationals, the golden mean $g = (\sqrt{5}-1)/2 \approx 0.618$ has a special place. Its continued fraction has every partial quotient equal to 1, its best rational approximations are ratios of successive Fibonacci numbers (1/2, 2/3, 3/5, 5/8, ...), and it is, in a precise number-theoretic sense, the irrational number hardest to approximate by rationals. A pair of uncoupled oscillators whose frequency ratio is the golden mean therefore stays as far as possible from the low-denominator ratios at which locking is easiest.

The qualification "uncoupled" matters. Once the oscillators interact, the realized ratio ρ departs from the natural ratio Ω, and Proposition 7.7 says that a natural ratio lying inside a tongue locks whatever its arithmetic. For the sine circle map at $K = 0.6$, a natural ratio equal to the golden mean produces a realized ratio of about 0.623 (our computation from (7.4)); at $K = 1$ the locked plateaus occupy the whole interval up to a set of measure zero, so almost every natural ratio, golden or not, ends up locked to some rational (Jensen et al. 1983). Whether two coupled rhythms lock depends on where the pair sits in the tongue diagram, which depends on the coupling and on the phase response, and not on the arithmetic of the natural ratio alone.

7.9 The Golden-Ratio Hypothesis for Brain Rhythms

Pletzer et al. (2010) proposed that the classical frequency bands of the human electroencephalogram (EEG) form a geometric series whose neighboring center frequencies stand in the ratio 1.618, the reciprocal of g. They showed that the excitatory phases of two oscillations whose frequencies are in the golden ratio never coincide in the mathematical sense, and that coincidences within a physiological tolerance are rarest and most irregular at that ratio; they interpreted the golden ratio as providing the most desynchronized, resting state of the brain, together with the possibility of irregular transient coupling. Klimesch (2018) developed a related architecture in which the center frequencies of traditional bands form a binary hierarchy, with phase coupling between a frequency and its harmonic multiples, and proposed that body oscillations such as heart rate and breathing can be predicted from brain oscillations.

We treat these proposals as hypotheses, and the theory of this chapter says what a test would need. First, the argument from coincidences of excitatory phases is kinematic: it treats the rhythms as uncoupled. The circle map shows that coupled rhythms lock or fail to lock according to their coupling and phase response, and that coupling moves the realized ratio away from the natural one (Section 7.8). Second, the ratio in question relates the centers of frequency bands, which are summaries across people and partly a matter of convention; individual peak frequencies vary, and the aperiodic component of the EEG spectrum shifts where peaks appear to lie unless it is modeled separately (Donoghue et al. 2020). Third, a ratio observed in data needs a null model: how often would ratios near 1.618 appear by chance, given the measured spread of peak frequencies? Appendix J describes how to build such nulls.

A decisive test would estimate individual peak frequencies after removing the aperiodic component, compare the distribution of their ratios with a surrogate null, and, most directly, measure the tongues: drive one rhythm at ratios near the golden ratio and near nearby rationals, at several stimulation strengths, and ask whether locking near 1.618 is weaker than the tongue geometry of the measured phase response predicts. The sources cited here propose the hypothesis and argue for it; they do not report that test, and until it is done the golden ratio in the EEG remains a conjecture about a relation, not an established relational law. Chapter 46 returns to relational measures of neural rhythms and the standard they must meet.

7.10 A Relation the Fish Avoids

The weakly electric fish of Part V offers a contrast. When two fish discharge at nearby frequencies, their fields beat slowly, and slow beats interfere with

each fish's electrolocation. The jamming avoidance response moves each fish's frequency so as to increase the magnitude of the frequency difference Δf: the fish raises its frequency if the neighbor's is lower and lowers it if the neighbor's is higher (Metzner 1993b; Heiligenberg 1991a). The variable the fish reads and acts on is the detuning and its sign, and the behavior drives the pair out of the region of slow beats rather than into a locked ratio. Chapters 20 and 21 follow that relation from the waveform to the pacemaker.

7.11 Audit

Canonical phenomenon

Spontaneously beating aggregates of chick heart cells, stimulated periodically, lock their beat to rational ratios of the stimulus rate over finite ranges of stimulus period (Guevara et al. 1981).

Minimal model

The Adler equation (7.1) for 1:1 locking; the sine circle map (7.3), its rotation number (7.4), Arnold tongues, and the devil's staircase.

What is relational here?

The rotation number ρ, the frequency ratio realized by the pair (oscillator, drive), as distinct from the ratio Ω of natural frequencies; within a locked state, the relative phase at which the drive arrives.

What physically reads it?

The oscillator itself, through its phase response curve: each stimulus shifts the phase by an amount that depends on the relative phase at which it arrives, so the relative phase is transduced into a phase correction. No further reader uses the rotation number.

What existing field explains it?

Nonlinear dynamics and the theory of circle maps (Arnold 1983; Jensen et al. 1983), the physics of commensurate and incommensurate phases (Bak 1982), cardiac and biological rhythm theory (Glass and Mackey 1988), and synchronization theory (Pikovsky et al. 2001). They explain everything in the chapter's physics and physiology.

What remains unexplained?

Nothing in the dynamics. The open question is empirical and concerns neural rhythms: whether brain rhythms sit inside or outside locking tongues, and whether the arithmetic of their ratios matters beyond what their coupling predicts.

A discriminating experiment

Drive a neural rhythm at frequency ratios near the golden ratio and near neighboring rationals, at several strengths, and measure the locking regions.

The circle-map account predicts tongue widths from the measured phase response alone; the golden-ratio hypothesis predicts weaker locking near 1.618 than that geometry implies. If the measured tongues match the phase-response prediction, the arithmetic of the ratio adds nothing.

7.12 Notes and References

Arnold (1983) treats circle maps, rotation numbers, and the tongues that bear his name. Glass and Mackey (1988) is the classic account of locking in biological rhythms, with the chick heart experiments of Guevara et al. (1981) as a central case; Pikovsky et al. (2001) covers phase locking, Farey ordering, and entrainment across physics and biology, and is the best single reference for this chapter. R. Adler (1946) derived the locking equation for injected electronic oscillators. The complete devil's staircase and its universal fractal dimension are due to Jensen et al. (1983); Bak (1982) reviews commensurate and incommensurate phases in condensed matter, and Bak (1986) gives a short account for a general scientific audience. The golden-ratio hypothesis for EEG bands is stated by Pletzer et al. (2010) and extended by Klimesch (2018); entrainment of human alpha rhythms within an Arnold tongue is reported by Notbohm et al. (2016), and entrainment by rhythmic magnetic stimulation by Thut et al. (2011). The separation of periodic and aperiodic spectral components, which any test of frequency ratios requires, is described by Donoghue et al. (2020).

7.13 Exercises

Exercise 7.1. For the Adler equation (7.1), find both fixed points when $|\nu| < K$ and show that only one is stable. Evaluate the integral in (7.2) (substitute $u = \tan(\phi/2)$) and confirm the slip period. How long does the phase linger near $\phi = \pi/2$ when $\nu = 1.01\,K$?

Exercise 7.2. An oscillator receives a stimulus every T seconds, and each stimulus shifts its phase (in cycles) by $\Delta(\theta) = -\varepsilon \sin 2\pi\theta$. Show that the phase before successive stimuli obeys the sine circle map, and express Ω and K in terms of T, the natural period T_0, and ε.

Exercise 7.3. Show that the $0/1$ tongue of (7.3) is exactly $|\Omega| \le K/2\pi$. Compute F^2 to second order in K near $\Omega = 1/2$ and show that the $1/2$ tongue has width approximately $K^2/4\pi$.

Exercise 7.4. Write a program that computes $\rho(\Omega, K)$ by iterating the lift for n steps after a transient. For $K = 0.5$, 0.9, and 1, estimate the total length of the plateaus with denominators $q \le 10$ and plot it against K. How large must n be to resolve a plateau of width 10^{-4}?

Exercise 7.5. List, in increasing order, the rationals between $1/3$ and $1/2$ with denominator at most 8. Predict from the Farey ordering which of them has the widest tongue at $K = 0.8$, and check the prediction numerically.

Exercise 7.6. Compute the first eight continued-fraction convergents of $g = (\sqrt{5}-1)/2$ and the error of each. Then compute $\rho(g, K)$ for K from 0 to 1 in steps of 0.1 and plot the difference $\rho - g$. What does the plot imply for an argument about brain rhythms that relies on the natural ratio alone?

Exercise 7.7. You are given individual alpha and theta peak frequencies from 60 participants. Design a test of the claim that their ratios cluster near 1.618: state the test statistic, the surrogate null, how the aperiodic spectral component is handled, and what result would count against the claim.

8

Magic Angles Without Magic

Stack two sheets of graphene and turn one of them by about one degree. The electrons near the Fermi level, which in a single sheet move as fast as any electrons in a solid, slow almost to a halt; the lowest band flattens; the material becomes an insulator when a particular number of electrons is added, and a superconductor when a few more are (Cao, Fatemi, Fang, et al. 2018; Cao, Fatemi, Demir, et al. 2018). Nothing about either sheet has changed. Only their relation has: a twist of one degree between two otherwise identical lattices.

The case is the most dramatic example of relational structure in contemporary physics, and it is easy to misread. The misreading says that at the magic angle two lattices are "misaligned by just the right amount," as if new behavior emerged from an optimum of disorder. The thesis of this chapter is that the magic angle is a quantitative prediction of a specific model. The twist generates a long length scale, and with it a kinetic energy scale; when the energy of tunneling between the layers becomes comparable to that kinetic scale, interference between tunneling paths drives the electrons' velocity through zero. The prediction preceded the experiment by seven years. What transfers from graphene to other relational questions is the structure of that prediction, and the chapter closes by stating it as a checklist.

8.1 Two Sheets and a Twist

Graphene is a single layer of carbon atoms on a honeycomb lattice, with carbon–carbon distance $a_{cc} \approx 0.142\,\text{nm}$ and lattice constant $a = \sqrt{3}\,a_{cc} \approx 0.246\,\text{nm}$. Near the corners of its hexagonal Brillouin zone, the points conventionally called K and K', the electronic bands form cones: the energy rises linearly with momentum measured from the corner, $E = \pm\hbar v_F |\boldsymbol{k}|$, with a Fermi velocity v_F of about $10^6\,\text{m/s}$ (Castro Neto et al. 2009). Electrons near these Dirac points behave as massless particles moving at that speed.

Place a second sheet on the first and rotate it by an angle θ. The two honeycombs drift in and out of registry across the sample, producing a moiré pattern of regions where atoms sit on top of atoms (AA stacking) and regions

of the two Bernal stackings (AB and BA). The moiré period is

$$L_{\mathrm{M}} = \frac{a}{2\sin(\theta/2)} \approx \frac{a}{\theta}, \tag{8.1}$$

the same relation derived for two rotated gratings in Chapter 2. At $\theta = 1.1^\circ$ it gives $L_{\mathrm{M}} \approx 12.8\,\mathrm{nm}$, about fifty lattice constants, and each moiré cell then contains roughly $4(L_{\mathrm{M}}/a)^2 \approx 11{,}000$ carbon atoms (our arithmetic from (8.1)). A misorientation of one degree has created a structure fifty times larger than either of the lattices that produce it.

For almost every angle the combined structure is incommensurate: no finite supercell repeats exactly, and the atomic arrangement is aperiodic (S. Carr et al. 2017). The moiré pattern is nevertheless periodic at long wavelengths, because the local stacking varies smoothly with position with period L_{M}. That long-wavelength periodicity is what the continuum theory of the next sections uses, and it is why the theory works at any angle.

8.2 The Moiré Brillouin Zone

Rotating a lattice rotates its reciprocal lattice. The Dirac points of the two layers, each at distance

$$k_{\mathrm{D}} = \frac{4\pi}{3a} \tag{8.2}$$

from the center of its layer's Brillouin zone, are rotated relative to each other by θ, and so they are separated by

$$k_\theta = 2k_{\mathrm{D}}\sin(\theta/2) \tag{8.3}$$

(Figure 8.1). The momentum k_θ is the reciprocal of the moiré length up to a factor of order one. It is the relational momentum scale of the problem: it exists only because there are two layers at an angle, and it vanishes when the angle does.

An electron moving from one layer to the other near a Dirac point must change its momentum by a vector of length k_θ. There are three such vectors, related by rotations of 120°, corresponding to the three equivalent corners of the Brillouin zone through which tunneling can proceed (Bistritzer and MacDonald 2011). Repeated tunneling builds a lattice in momentum space whose unit is set by k_θ: the reciprocal lattice of the moiré superlattice.

8.3 The Bistritzer–MacDonald Model

Bistritzer and MacDonald (2011) wrote down a continuum model for the twisted bilayer that keeps only what the long-wavelength physics needs: a Dirac Hamiltonian for each layer and a tunneling term that varies periodically in space with the moiré period.

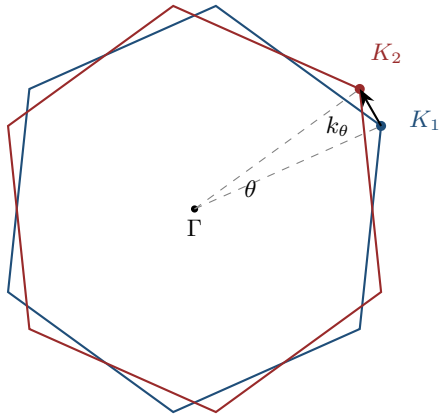


Figure 8.1: Brillouin zones of two graphene layers rotated by θ (exaggerated here to 12°). Corresponding Dirac points K_1 and K_2, each at distance k_D from the zone center Γ, are separated by $k_\theta = 2k_\mathrm{D}\sin(\theta/2)$, which sets the size of the moiré Brillouin zone.

Model 8.1 (Continuum model of the twisted bilayer)**.** Near one Dirac point, the electron wave function has a component in each layer, $\Psi = (\psi_1, \psi_2)$, each a two-component spinor on the sublattices. The Hamiltonian is

$$H = \begin{pmatrix} h_{\theta/2}(\boldsymbol{k}) & T(\boldsymbol{r}) \\ T^\dagger(\boldsymbol{r}) & h_{-\theta/2}(\boldsymbol{k}) \end{pmatrix}, \qquad T(\boldsymbol{r}) = \sum_{j=1}^{3} T_j \, \mathrm{e}^{-\mathrm{i}\boldsymbol{q}_j\cdot\boldsymbol{r}}, \tag{8.4}$$

where $h_{\pm\theta/2}$ is the Dirac Hamiltonian of a layer rotated by $\pm\theta/2$, with velocity v_F, the three momentum transfers $\boldsymbol{q}_j$ have length k_θ, and the 2×2 matrices T_j carry a single tunneling energy w. Comparison with the known electronic structure of Bernal-stacked bilayers fixes $w \approx 110\,\mathrm{meV}$.

Up to an overall energy scale, the bands of the model depend on a single dimensionless number (Bistritzer and MacDonald 2011):

$$\alpha = \frac{w}{\hbar v_\mathrm{F} k_\theta}. \tag{8.5}$$

The numerator is the energy an electron gains by tunneling between layers. The denominator is the kinetic energy of an electron that has traveled a distance in momentum space equal to the separation of the two layers' Dirac points. Because k_θ shrinks with the angle, α grows roughly as $1/\theta$. At large angles, α is small, the layers barely notice each other near the Dirac points, and the low-energy spectrum is that of two independent graphene sheets with a slightly reduced velocity (Lopes dos Santos et al. 2007). As the angle decreases, α grows and the layers couple strongly.

The key result can be obtained by truncating the momentum-space lattice after its first shell, keeping one Dirac spinor in one layer and three in the other.

Proposition 8.2 (Renormalized Dirac velocity)**.** *In the four-spinor truncation of Model 8.1, the Hamiltonian has two zero-energy states at the Dirac point, and the velocity of the Dirac cone formed from them is*

$$\frac{v^\star}{v_\mathrm{F}} = \frac{1 - 3\alpha^2}{1 + 6\alpha^2}, \tag{8.6}$$

which vanishes at $\alpha = 1/\sqrt{3}$.

For small α the formula reduces to $v^\star/v_\mathrm{F} \approx 1 - 9\alpha^2$, the reduction first found by Lopes dos Santos et al. (2007). The denominator in (8.6) records how much of the zero-energy state lives in the second layer, and the numerator records that the second layer's components carry velocity of the opposite sign (Bistritzer and MacDonald 2011). The velocity vanishes because the state at the Dirac point is a coherent combination of pieces moving in opposite directions, and at one value of α they cancel. The full numerical solution of the model places the first zero of the velocity at $\theta \approx 1.05°$ for $w \approx 110\,\mathrm{meV}$, with further zeros at smaller angles, near 0.5°, 0.35°, 0.24°, and 0.2°; at the first of them the whole lowest band becomes very flat (Bistritzer and MacDonald 2011). These are the magic angles.

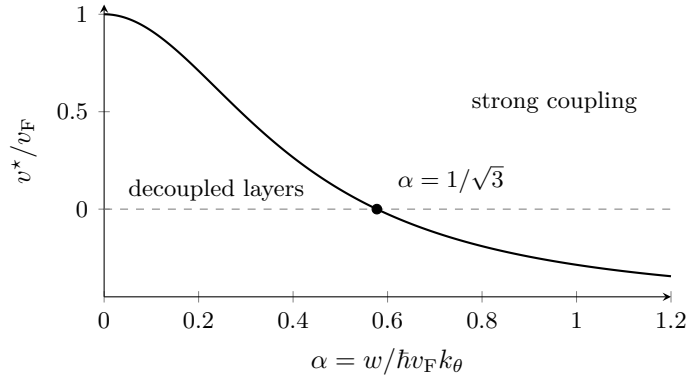


Figure 8.2: Renormalized Dirac velocity in the four-spinor truncation of the continuum model, equation (8.6). The velocity passes through zero, changing sign, at $\alpha = 1/\sqrt{3}$. The full model gives the first zero at $\theta \approx 1.05°$ for the tunneling energy $w \approx 110\,\mathrm{meV}$ (Bistritzer and MacDonald 2011).

Example 8.3 (How sensitive is the angle?). With $v_\mathrm{F} = 10^6\,\mathrm{m/s}$ and $a = 0.246\,\mathrm{nm}$, equations (8.2)–(8.5) give $\alpha \approx 0.54$ at $\theta = 1.05°$, and the truncated formula (8.6) would put the zero near $\theta \approx 0.97°$ (our arithmetic). The full model with the parameters of Bistritzer and MacDonald (2011) gives 1.05°; tight-binding calculations with other interlayer couplings found flat bands near 1.5° (Suárez Morell et al. 2010). The magic angle is a derived quantity. It moves when the tunneling energy, the Fermi velocity, or the relaxation of the lattice changes, because what is fixed is a value of α, a ratio of two energies, and not a value of θ.

8.4 What Happens on a Flat Band

A flat band is a band whose electrons have almost no kinetic energy to lose or gain. When the kinetic energy scale collapses, the Coulomb repulsion between electrons, which in ordinary graphene is a small correction, becomes the largest energy in the problem. The single-particle model of Section 8.3 predicts the flat band; it does not predict what the interacting electrons will do on it.

The experiments answered that question in 2018. In bilayers twisted by about 1.1°, Cao, Fatemi, Demir, et al. (2018) found insulating states at half filling of the flat bands that no single-particle theory predicts, the signature of strong correlations. Doping the same devices electrostatically away from these insulating states, Cao, Fatemi, Fang, et al. (2018) found zero-resistance states with critical temperatures up to 1.7 K, in domes on the phase diagram of temperature against carrier density that resemble the phase diagrams of copper-oxide superconductors. Since then magic-angle bilayers have been found to host magnetism, electronic nematicity, resistivity linear in temperature, and quantized anomalous Hall states, with properties that are extremely sensitive to carrier density, to nearby gates, and to variation of the twist angle across a device (Andrei and MacDonald 2020). The mechanism of the superconductivity is still debated.

For this book the division of labor is the point. The relation between the layers sets a length (L_{M}), a momentum (k_θ), and an energy ratio (α), and those fix where the band is flat. Everything that follows once the band is flat depends on interactions among electrons, which the relational story does not supply.

8.5 Turning the Other Knob

If the control parameter is α and not θ, it should be possible to reach the magic condition by changing the numerator instead of the denominator. Yankowitz et al. (2019) did so. Hydrostatic pressure pushes the layers closer together and increases the coupling between them; applying it to a bilayer twisted by more than 1.1°, where correlated phases are otherwise absent, they induced superconductivity. Pressure shifted the superconducting transition to larger angles and somewhat higher temperatures.

The experiment is an intervention on a relation in the sense of Chapter 1. The twist angle stayed fixed; the interlayer tunneling energy changed; the behavior tracked the ratio of tunneling energy to the kinetic scale set by the angle. Two different manipulations reached the same state because they moved the same dimensionless number.

8.6 The Chiral Limit and the Pattern of Magic Angles

The truncated model explains the first magic angle but not the whole flat band or the pattern of smaller magic angles, which Bistritzer and MacDonald (2011) found only numerically. Tarnopolsky et al. (2019) studied a limit of the continuum model in which tunneling between atoms of the same sublattice (the AA regions) is switched off. In this chiral limit the lowest band is not merely slow at the Dirac point; it is exactly flat throughout the moiré Brillouin zone at a discrete set of values of α. Parametrized by α, which scales as $1/\theta$, the magic values recur with an almost constant spacing, $\Delta\alpha \approx 3/2$. The flat-band wave functions can be written in closed form using ratios of theta functions, like quantum Hall wave functions on a torus, and the first magic angle turns out to be unusually robust; lattice relaxation, which the authors discuss, helps justify the chiral parameters.

The irregular sequence of magic angles in θ is therefore a regular sequence in the relational variable α. Chosen correctly, the relational coordinate turns an apparent numerology into a law.

8.7 Not the Perfect Amount of Disorder

Three features of the physics refute the reading of the magic angle as an optimum of misalignment.

First, the twisted bilayer at the magic angle is ordered. At the scale that matters its structure is a perfectly periodic moiré superlattice, and the flat band arises from coherent interference between three tunneling paths. Disorder in the ordinary sense, variation of the twist angle from place to place, degrades the effect, and the experiments that resolve the phase diagram depend on devices with little of it (Yankowitz et al. 2019; Andrei and MacDonald 2020).

Second, the magic angle requires no commensurability. Bistritzer and MacDonald (2011) emphasize that their model can be solved at any twist angle, whether or not the two lattices share a common supercell, and the first magic angle is not a special rational angle. In the vocabulary of Chapter 7, it does not belong in the row of locking.

Third, the condition is a statement about two energies of order each other. The magic condition fixes α, the ratio of the interlayer tunneling energy to the kinetic energy at the moiré momentum scale, at a particular value of order one where the velocity crosses zero. That places the phenomenon in the row of magnitude balance (Chapter 5), sharpened by an interference zero: a crossover between decoupled and strongly coupled layers, with an exact cancellation at a predicted point inside it.

8.8 A Moiré Where Commensurability Does Matter

Moiré superlattices also host a phenomenon that does belong to the row of commensurability. Electrons in a periodic potential and a magnetic field develop a fractal energy spectrum, Hofstadter's butterfly, whose structure depends on the rational ratio of magnetic flux through each unit cell to the flux quantum. With atomic lattices the required fields are unreachably large, and with engineered lattices of hundreds of nanometers the fields are too small to overcome disorder. Dean et al. (2013) used the moiré superlattice formed between bilayer graphene and hexagonal boron nitride, with periods of the order of ten nanometers, to put the commensurability condition within reach and observe the fractal spectrum.

The same kind of object, a moiré superlattice, served two different relations. In the twisted bilayer the relation that matters is a balance of energies; in the Hofstadter experiment it is a rational ratio of two lengths, the moiré period and the magnetic length. The pattern does not determine the physics. The Hamiltonian does.

8.9 Why Graphene Is Not a Theory of Consciousness

The magic angle has been tempting as a metaphor for mind: consciousness as what appears when many internal models of the world are misaligned by just the right amount. The physics shows how much the metaphor leaves out. The magic angle came with a model that named its control parameter, its intrinsic scales, the scale generated by the relation, and the transition predicted at a definite value of a dimensionless ratio. The prediction was published in 2011 (Bistritzer and MacDonald 2011); the correlated insulator and the superconductor were reported in 2018 (Cao, Fatemi, Fang, et al. 2018; Cao, Fatemi, Demir, et al. 2018); a second manipulation that moved the same ratio reproduced the effect (Yankowitz et al. 2019). None of those elements comes with the metaphor.

Principle 8.4 (What a relational explanation must supply)**.** A claim that a phenomenon arises from a relation between components earns explanatory status only when it identifies:

1. a *control parameter* that sets the relation and can be varied (the twist angle; the interlayer spacing under pressure);
2. the *intrinsic scales* of the components (a, v_{F}, w);
3. the *relationally generated scale* that exists only because of the relation (L_{M}, k_θ, and the kinetic energy $\hbar v_{\mathrm{F}} k_\theta$);

4. a *predicted transition* at a stated value of a dimensionless combination of these ($v^\star = 0$ at a definite α).

Part V applies the checklist to the electric fish, where the control parameter is the frequency difference Δf between two fish and the relationally generated scale is the beat period $1/|\Delta f|$, a time that exists only because there are two discharges. Part XIII applies it to relational proposals about neural rhythms and consciousness, most of which do not yet supply the fourth item.

8.10 Audit

Canonical phenomenon

Two graphene sheets twisted by about 1.1° develop flat bands, correlated insulating states at half filling, and superconductivity up to 1.7 K (Cao, Fatemi, Fang, et al. 2018; Cao, Fatemi, Demir, et al. 2018).

Minimal model

The continuum model (8.4) with the single parameter $\alpha = w/\hbar v_\mathrm{F} k_\theta$; in its simplest truncation, $v^\star/v_\mathrm{F} = (1 - 3\alpha^2)/(1 + 6\alpha^2)$.

What is relational here?

The twist angle between the layers and the quantities it generates: the moiré period L_M, the momentum k_θ separating the layers' Dirac points, and the ratio α of tunneling to kinetic energy.

What physically reads it?

The electrons, in the sense that their band structure depends on the relation: tunneling between layers transduces the relative orientation into interference between momentum-space paths. No component of the system uses the relation to regulate anything.

What existing field explains it?

Condensed-matter physics explains the flat band (Bistritzer and MacDonald 2011; Tarnopolsky et al. 2019); the correlated phases are an active research area within the same field (Andrei and MacDonald 2020). Nothing in the chapter requires a new discipline.

What remains unexplained?

None that bears on this book. The open problems (the mechanism of superconductivity, the nature of the correlated states) belong to many-body physics and are not relational in any sense that physics lacks the tools to handle.

A discriminating experiment

The pressure experiment already realized it: change the tunneling energy at fixed angle and check that the correlated phases follow α rather than θ (Yankowitz et al. 2019). For a relational theory in another domain,

the analog is to find two independent manipulations that move the same dimensionless ratio and to check that the behavior depends on the ratio alone.

8.11 Notes and References

Castro Neto et al. (2009) is the standard review of graphene's electronic structure. The continuum theory of the twisted bilayer begins with Lopes dos Santos et al. (2007), who found the velocity reduction at small angles, and Bistritzer and MacDonald (2011), who predicted the magic angles; tight-binding calculations by Suárez Morell et al. (2010) found flat bands at small twists independently. Tarnopolsky et al. (2019) explain the exact flatness and the periodicity of the magic angles in the chiral limit. The experimental discovery of correlated insulators and superconductivity is reported by Cao, Fatemi, Demir, et al. (2018) and Cao, Fatemi, Fang, et al. (2018); pressure tuning by Yankowitz et al. (2019). Andrei and MacDonald (2020) review the field and its open questions, and S. Carr et al. (2017) survey the control of layered materials by twist angle, which they call twistronics. Dean et al. (2013) report the Hofstadter butterfly in graphene moiré superlattices. The geometry of moiré periods is developed in Chapter 2, and the classification of special relations used here in Chapters 5 to 7.

8.12 Exercises

Exercise 8.1. Compute the moiré period (8.1) and the number of carbon atoms per moiré cell for $\theta = 5^\circ$, 1.1°, and 0.5°. By what factor does the number of atoms per cell grow when the angle is halved, and why?

Exercise 8.2. With $w = 110\,\mathrm{meV}$ and $a = 0.246\,\mathrm{nm}$, compute α as a function of θ for $v_\mathrm{F} = 0.9 \times 10^6$, 1.0×10^6, and $1.1 \times 10^6\,\mathrm{m/s}$. For each, find the angle at which the truncated formula (8.6) gives $v^\star = 0$. What spread in the predicted magic angle does a ten percent uncertainty in v_F produce?

Exercise 8.3. Suppose pressure increases w by fifteen percent. Using the truncated formula, by how much does the magic angle move, and in which direction? Explain why this is consistent with superconductivity appearing under pressure at angles larger than 1.1°.

Exercise 8.4. Expand (8.6) to second order in α and recover $v^\star/v_\mathrm{F} \approx 1 - 9\alpha^2$. At what angle does this approximation differ from the full truncated formula by ten percent?

Exercise 8.5. A moiré superlattice has a triangular unit cell of area $(\sqrt{3}/2)L_\mathrm{M}^2$. Compute the magnetic field that threads one flux quantum h/e through each cell for $L_\mathrm{M} = 13\,\mathrm{nm}$, and compare it with the field needed for an atomic lattice

with $L_{\mathrm{M}} = 0.25\,\mathrm{nm}$. Which of the three special relations distinguished in Chapter 7 is at work?

Exercise 8.6. Choose a published proposal that explains a biological or cognitive phenomenon by a relation between two processes. Apply Principle 8.4: identify, or show to be missing, the control parameter, the intrinsic scales, the relationally generated scale, and the predicted transition.

Three different things hide behind the phrase "the right relation." A ratio of magnitudes near one marks a crossover between regimes. A frequency near a natural frequency produces a response peak. A rational ratio of frequencies produces locking. Magic-angle graphene, the Reynolds number, a driven pendulum, and a devil's staircase each belong to a different column, and a theory that runs them together explains none of them. Part III follows relations that stop being descriptions and start behaving as variables with their own dynamics.

Part III

Collective Variables

When does a relation become a variable the system obeys?

9

Synchronization

A population of oscillators with slightly different natural frequencies, each weakly influenced by all the others, can do one of two things. It can stay incoherent, each member keeping its own time, or part of it can fall into step and run at a common frequency. Winfree (1967) saw that the change between these behaviors is abrupt, a threshold in the strength of coupling relative to the spread of frequencies, and compared it to a phase transition; Kuramoto (1975) found a model simple enough to solve exactly. The thesis of this chapter is that synchronization in a population is carried by a relational variable of the whole population, the order parameter, which every oscillator both helps to produce and responds to. Its onset is a magnitude balance between coupling and heterogeneity, and above onset the population divides into a locked core and drifting stragglers. The medium that couples the oscillators is also what makes the collective relation physically available to each of them, and that observation connects the physics of this chapter to the reader problem of Part IV.

9.1 From Clocks to Populations

Huygens's two pendulum clocks settled into opposition because the beam that carried them transmitted each clock's motion to the other (Chapter 1). Populations synchronize the same way, with more members and more varied natural frequencies. Males of several firefly species flash in synchrony, coordinating visually; Buck (1988) reviews the phenomenon, which is widespread geographically and taxonomically, rests on several different neural mechanisms of flash control, and is bound up with courtship in ways that remain only partly understood. Chemical oscillators, cardiac pacemaker cells, and electronic circuits provide further cases (Pikovsky et al. 2001).

The common structure is a large set of self-sustained oscillators, each with its own natural frequency, coupled through a shared medium or through mutual influence. The question is when, and to what extent, the population keeps a common time.

9.2 Phase Oscillators

A self-sustained oscillator returns to a stable cycle after a perturbation. When perturbations are weak, the only variable that can change appreciably and stay changed is the oscillator's position on its cycle, its phase: deviations off the cycle decay, while shifts along the cycle persist because nothing restores them. Weakly coupled oscillators can therefore be described by their phases alone, and averaging over a cycle makes the coupling depend on phase differences (Appendix D). This is the invariance of Chapter 1 at work: the dynamics does not change when all phases are shifted together, so only the relations among phases enter.

Winfree's own formulation gives each oscillator a phase, an *influence function* $X(\theta)$ describing the signal it emits at each point of its cycle, and a *sensitivity function* $Z(\theta)$ describing how strongly it responds at each point:

$$\dot{\theta}_i = \omega_i + Z(\theta_i)\,\frac{1}{N}\sum_{j=1}^{N} X(\theta_j). \tag{9.1}$$

Every oscillator contributes to a single mean influence and every oscillator feels it (Winfree 1967; Strogatz 2000). Kuramoto's simplification replaces the product $Z(\theta_i)X(\theta_j)$ by a function of the phase difference $\theta_j - \theta_i$ alone, which is what makes the model solvable.

Model 9.1 (Kuramoto model)**.** N oscillators with phases θ_i and natural frequencies ω_i, drawn independently from a density $g(\omega)$, evolve as

$$\dot{\theta}_i = \omega_i + \frac{K}{N}\sum_{j=1}^{N}\sin(\theta_j - \theta_i), \qquad i = 1, \dots, N, \tag{9.2}$$

with coupling strength $K \geq 0$.

The model keeps the simplest coupling function that depends only on phase differences, pulls each oscillator toward the others, and treats every pair alike. Its value lies in what can be computed from it exactly.

9.3 The Order Parameter

Definition 9.2 (Order parameter)**.** The *complex order parameter* of the population is

$$r\,\mathrm{e}^{\mathrm{i}\psi} = \frac{1}{N}\sum_{j=1}^{N}\mathrm{e}^{\mathrm{i}\theta_j}. \tag{9.3}$$

Its modulus $r \in [0, 1]$ measures phase coherence; its argument ψ is the collective phase.

The modulus r is a relational observable on all N components. It is unchanged when every phase is shifted by the same amount, so by Chapter 1 it is a quantity no convention can alter, and it is 1 when all oscillators share a phase and near 0 when phases are spread evenly around the circle.

Multiplying (9.3) by $e^{-i\theta_i}$ and taking the imaginary part gives $r\sin(\psi-\theta_i) = N^{-1}\sum_j \sin(\theta_j-\theta_i)$, so the model can be rewritten exactly as

$$\dot{\theta}_i = \omega_i + Kr\sin(\psi-\theta_i). \tag{9.4}$$

Each oscillator feels the rest of the population only through the order parameter. It is pulled toward the collective phase ψ with a strength Kr proportional to the coherence of the population, and the coherence is in turn produced by the oscillators' phases. The order parameter is both an effect of the population and the cause of each member's motion, which is the feedback that makes synchronization self-reinforcing once it begins.

9.4 The Self-Consistency Argument

Kuramoto found the synchronized states by assuming that, in the limit $N\to\infty$, the population settles into a state with constant coherence r and a collective phase rotating at a constant rate. Take g to be symmetric about its center and unimodal, and work in the frame rotating at the center frequency, so that the center is 0 and ψ can be set to 0.

The oscillators split into two groups. Those with $|\omega_i| \le Kr$ reach a fixed point of (9.4), with

$$\sin\theta_i = \frac{\omega_i}{Kr}, \qquad |\theta_i| \le \frac{\pi}{2}. \tag{9.5}$$

These are the *locked* oscillators: each is an Adler equation (Chapter 7) with effective coupling Kr, and each sits inside its own locking range. Those with $|\omega_i| > Kr$ cannot lock; they drift around the circle, lingering where their velocity $\omega_i - Kr\sin\theta$ is smallest. In the stationary state their phases are distributed with density inversely proportional to that velocity, and because g is symmetric, the drifting oscillators at ω and $-\omega$ contribute equal and opposite amounts to the imaginary part of the order parameter and nothing on average to its real part (Kuramoto 1984; Strogatz 2000). Only the locked oscillators contribute, and the requirement that they reproduce the r they respond to gives

$$r = Kr\int_{-\pi/2}^{\pi/2}\cos^2\theta\; g(Kr\sin\theta)\,d\theta. \tag{9.6}$$

Proposition 9.3 (Kuramoto threshold)**.** *For symmetric unimodal g, equation (9.6) always has the incoherent solution $r=0$. A branch of partially synchronized solutions with $r>0$ bifurcates from it at*

$$K_c = \frac{2}{\pi g(0)}. \tag{9.7}$$

For the Lorentzian density $g(\omega) = \gamma/[\pi(\gamma^2 + \omega^2)]$, $K_c = 2\gamma$ *and the branch is*

$$r = \sqrt{1 - \frac{K_c}{K}}, \qquad K > K_c. \tag{9.8}$$

The threshold follows from dividing (9.6) by r and letting $r \to 0$: the integral tends to $g(0)\,\pi/2$, so a nonzero solution first appears when $Kg(0)\pi/2 = 1$. Near threshold the coherence grows as $(K - K_c)^{1/2}$ whenever $g''(0) < 0$, the signature of a continuous transition; Appendix E carries out the expansion, gives the exact Lorentzian result through the reduction of Ott and Antonsen (2008), and treats the finite-size fluctuations that keep r of order $N^{-1/2}$ below threshold.

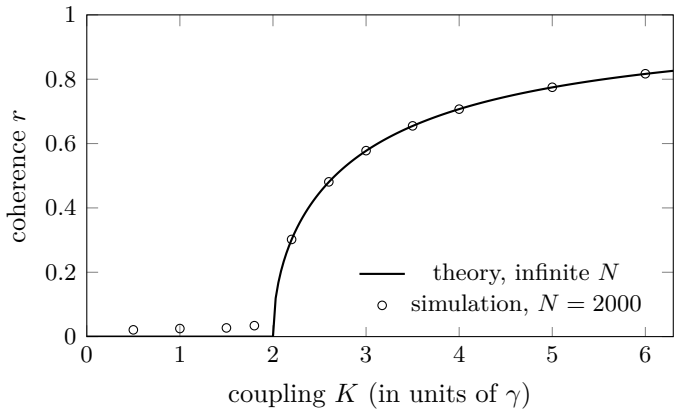


Figure 9.1: Onset of synchronization in the Kuramoto model with Lorentzian natural frequencies of half-width $\gamma = 1$. Solid line: the exact result (9.8), with threshold $K_c = 2\gamma$. Circles: time-averaged coherence in simulations of $N = 2000$ oscillators (our computation; fourth-order Runge–Kutta with step 0.02, averaged over the second half of runs of length 150; frequencies set at the Lorentzian quantiles and clipped at $|\omega| = 60$). Below threshold the simulated r is of order $N^{-1/2}$.

Figure 9.1 compares the exact curve with a simulation of two thousand oscillators. The agreement above threshold is to the third decimal place; below threshold the finite population keeps a residual coherence of about 0.02, the size expected from random phases.

9.5 The Threshold as a Balance

For the Lorentzian density the threshold is $K_c = 2\gamma$, twice the half-width of the frequency distribution; for a Gaussian of standard deviation σ it is $K_c = \sigma\sqrt{8/\pi} \approx 1.6\,\sigma$. In either case the control parameter is a dimensionless ratio, coupling divided by spread, and synchronization sets in when that ratio passes a value of order one. In the classification of Part II, the onset of collective synchrony is a case of magnitude balance (Chapter 5), while the capture of

each individual oscillator is a case of locking at the ratio 1:1 (Chapter 7): the locked oscillators are exactly those whose detuning from the center lies inside the tongue of width $2Kr$. The population combines the two relations, and the self-consistency condition (9.6) is the equation that ties them together.

The form of the distribution matters beyond its width. For a uniform distribution of frequencies the transition becomes discontinuous: at threshold the coherence jumps from zero to a finite value (Exercise 9.5). Heterogeneity is described by a whole function, and the balance involves its shape.

9.6 Partial Synchrony and Clusters

Above threshold the synchronized state is partial. The fraction of oscillators that lock is the probability mass of g within $|\omega| \leq Kr$, and it grows with coupling; the rest drift, each with a time-averaged frequency pulled toward the center but distinct from it. A single number r therefore summarizes a population that contains two dynamically different kinds of member.

In networks where each oscillator is coupled only to some others, the coupling term becomes $N^{-1}\sum_j A_{ij}\sin(\theta_j - \theta_i)$ with an adjacency matrix A_{ij}, and synchronization can organize into clusters: groups of oscillators that share a phase while different groups do not. Pecora et al. (2014) showed that the possible patterns of synchronized clusters are determined by the symmetries of the network, the permutations of nodes that leave the adjacency matrix unchanged, and confirmed the connection in an electro-optic network with real heterogeneity and noise. They also predicted and observed isolated desynchronization, in which one cluster loses synchrony while the others keep it. Cluster synchronization is a relation among relations: which oscillators share a phase depends on which nodes are equivalent under the network's symmetries, a property of the whole graph.

Example 9.4 (Clusters on a ring). Four identical oscillators on a ring, each coupled to its two neighbors, have the symmetries of a square. Among the states consistent with those symmetries are full synchrony, a state in which opposite nodes share a phase while neighbors differ, and a traveling wave in which each node leads its neighbor by a quarter cycle. Which of these patterns is stable depends on the coupling function; which patterns the symmetries of the graph make invariant is fixed by the graph alone (Pecora et al. 2014).

9.7 Measuring Synchrony in Data

In a model the phases are given. In an experiment they have to be estimated, and the estimate is itself a reading of the signal. The usual route treats a recorded signal $x(t)$ as the real part of a complex analytic signal whose argument defines an instantaneous phase (Appendix B); for narrow-band signals this phase behaves like the phase of an oscillator. Given phases $\theta_1(t)$ and $\theta_2(t)$

for two recording sites, the analogue of the order parameter for their relative phase is

$$R_{12} = \left| \frac{1}{T} \int_0^T \mathrm{e}^{\mathrm{i}[\theta_1(t) - \theta_2(t)]} \, \mathrm{d}t \right|, \tag{9.9}$$

the mean phase coherence of Mormann et al. (2000); its across-trial version, computed at a fixed latency over repeated trials, is the phase-locking value of Lachaux et al. (1999). Both equal 1 when the phase difference is constant and fall toward 0 when it wanders uniformly.

The measures are relational observables in the sense of Chapter 1, and they inherit the caution of that chapter. A high value shows that the phase difference is concentrated; it does not show that the two sites are coupled to each other. Two oscillators driven by a common input will show a stable phase relation with no interaction between them, and two electrodes that record the same source through the conducting tissue of the head will show zero-lag coherence with no second source at all (Appendix K). Finite recordings produce nonzero values by chance, so every estimate needs a null distribution built from surrogates that preserve each signal's spectrum while destroying the relation (Appendix J). The Kuramoto model supplies a sharper test than any coherence value: if two populations are coupled, their locked fraction should grow with coupling strength in a way the self-consistency condition predicts, and interventions that change the coupling should move the threshold. Coherence is where the evidence for synchronization begins, not where it ends, and Chapter 46 returns to the standard that relational measures of neural data must meet.

9.8 What Reads the Order Parameter?

No oscillator in the Kuramoto model computes r. Each responds, through its own phase response, to the sum of the influences it receives, and in the mean-field form (9.4) that sum is the order parameter. In physical realizations the sum is a physical quantity. In Huygens's clocks it is the motion of the beam. In globally coupled electrochemical or electronic oscillators it is a shared current or voltage. Among fireflies it is the pattern of light that each insect sees. The coupling medium performs, in the vocabulary of Chapter 1, the mixing M: it adds the contributions of all members and delivers the sum to each. Each member's phase response then performs a reading D, turning the relation between its own phase and the collective phase into a correction of its timing.

The relation is made physically explicit by the medium, and it drives each unit. That is a real reading in the thin sense of Chapter 15: a physical operation whose output depends systematically on a relational variable. It is not yet a reader that the system builds, maintains, or uses to regulate anything beyond its own timing, and nothing in the population misreads the order parameter or

could. The firefly case sits one step further up the ladder, since the coupling passes through a visual system that detects flashes; the weakly electric fish of Part V sits further still. Two fish with nearby discharge frequencies are also coupled through a shared medium, the water, but their reading loop does the opposite of Kuramoto coupling: each fish reads the sign of the frequency difference and moves its own frequency away from its neighbor's (Chapters 20 and 21).

9.9 Audit

Canonical phenomenon

Large populations of self-sustained oscillators with different natural frequencies (fireflies, clocks on a common support, chemical and electronic oscillators) fall partly into step when their coupling exceeds a threshold set by the spread of their frequencies.

Minimal model

The Kuramoto model (9.2), its mean-field form (9.4), and the self-consistency condition (9.6) with threshold $K_c = 2/\pi g(0)$.

What is relational here?

The order parameter $r\,\mathrm{e}^{\mathrm{i}\psi}$, a function of all phases that is invariant under a common shift; in networks, the pattern of synchronized clusters fixed by the graph's symmetries.

What physically reads it?

The coupling medium sums the members' contributions and delivers the collective phase to each; each member's phase response transduces the relation between its phase and the collective phase into a timing correction. No component uses the order parameter for anything else.

What existing field explains it?

Nonlinear dynamics and statistical physics (Winfree 1967; Kuramoto 1984; Strogatz 2000; Acebrón et al. 2005), with exact low-dimensional reductions (Ott and Antonsen 2008) and network extensions (Pecora et al. 2014); synchronization theory (Pikovsky et al. 2001). The chapter adds nothing to them.

What remains unexplained?

None in the physics. The chapter leaves a question for the reader problem: the difference between a medium that delivers a collective relation to every unit and a unit that builds its own means of reading a relation and acts on what it reads.

A discriminating experiment

In a population whose coupling can be set externally (electrochemical or

electronic oscillators), hold the coupling medium fixed and alter each unit's phase response, for instance by changing its waveform. The Kuramoto account predicts that the threshold moves with the first Fourier coefficient of the coupling function. A result in which the collective relation depended on details the phase reduction discards would show that the relational description is incomplete for that system.

9.10 Notes and References

Winfree (1967) introduced the population view of biological rhythms and the idea of a synchronization threshold; his book (Winfree 2001) is the classic treatment of biological timing. Kuramoto (1975) presented the model, and Kuramoto (1984) developed its analysis together with the reduction of oscillators to phases. Strogatz (2000) is an engaging history of the model's mathematics, and Acebrón et al. (2005) a comprehensive review. Ott and Antonsen (2008) found the exact low-dimensional description that makes the Lorentzian case solvable in time as well as at equilibrium. Pikovsky et al. (2001) cover synchronization across physics, chemistry, and biology. Cluster synchronization and its dependence on network symmetry are due to Pecora et al. (2014); synchronous fireflies are reviewed by Buck (1988).

9.11 Exercises

Exercise 9.1. Derive the mean-field form (9.4) from (9.2) and (9.3). Why is the substitution exact for any N, and why does the self-consistency argument nevertheless require $N \to \infty$?

Exercise 9.2. For $N = 2$, rewrite the model as a single equation for $\phi = \theta_1 - \theta_2$. Show that it is the Adler equation of Chapter 7 with coupling K, and state the locking condition in terms of $\omega_1 - \omega_2$.

Exercise 9.3. For a drifting oscillator with $|\omega| > Kr$, show that the stationary density of its phase is proportional to $1/|\omega - Kr\sin\theta|$ and compute its normalization. Show that the drifting oscillators contribute nothing to the real part of the order parameter when g is symmetric.

Exercise 9.4. Compute K_c for a Gaussian density of standard deviation σ. Then expand (9.6) to second order in Kr and show that $r^2 \approx 16(K - K_c)/[\pi K_c^4 |g''(0)|]$ near threshold.

Exercise 9.5. Let g be uniform on $[-\gamma, \gamma]$. Show that (9.6) is satisfied at $K = 4\gamma/\pi$ by every r with $Kr \le \gamma$, and conclude that at threshold the coherence jumps from 0 to $\pi/4$.

Exercise 9.6. Simulate the model for $N = 500$ oscillators with Lorentzian frequencies and reproduce Figure 9.1. How does the residual coherence below

threshold scale with N? Estimate, from your simulations, how close to K_c one must be before finite-size effects dominate for $N = 500$.

Exercise 9.7. For a population of fireflies, identify the mixing M and the reading D in the sense of Section 9.8. What would you need to show in order to say that an individual firefly reads a relation between its own rhythm and the group's, rather than responding to each flash it sees?

10

Coordination Dynamics

Move your two index fingers back and forth in opposition, one flexing while the other extends, and speed up gradually. At some frequency the pattern breaks and the fingers fall into moving together, in phase. Slow down again and they stay in phase. J. A. Kelso (1984) measured this switch, and Haken et al. (1985) wrote down a model of it whose only variable is the relative phase of the two fingers. The thesis of this chapter is that relative phase in human coordination is an order parameter with its own dynamical law: the law can be written down, derived from a model of the components, and tested by predictions about stability that were confirmed before the transition was explained in any other way. Coordination dynamics is the clearest case in biology of a relational variable that obeys a law of its own. It also shows exactly what such a law leaves open: which physical processes in the nervous system carry the relation that the law describes.

10.1 The Bimanual Experiment

In Kelso's experiments participants oscillated their hands or fingers rhythmically at movement frequencies that were increased during the trial. Starting in the anti-phase pattern, in which homologous muscles of the two hands contract alternately, participants switched abruptly to the in-phase pattern, in which homologous muscles contract together, once the movement frequency passed a critical value (J. A. Kelso 1984). The boundary between the two patterns was indexed by a dimensionless critical number that stayed constant whether the hands moved freely or against resistive loads. Kelso described the switch as a bifurcation: continuous scaling of the movement frequency renders the existing pattern unstable, and a new stable pattern takes over, much as an animal changes gait when it runs faster (J. A. Kelso 1984; Haken et al. 1985).

A further observation constrains any explanation. The switch runs one way: a participant who starts in phase and speeds up stays in phase, and one who has switched to in-phase does not return to anti-phase when the frequency is lowered again (Schöner and J. A. S. Kelso 1988; J. A. S. Kelso

1995). Whatever mechanism is at work acts on the relation between the hands and has memory of which pattern the hands are in.

10.2 Relative Phase as Order Parameter

The variable that distinguishes the two patterns is the relative phase of the two hands, $\phi = \theta_{\mathrm{L}} - \theta_{\mathrm{R}}$, near 0 in the in-phase pattern and near π in the anti-phase pattern. It is a relational observable in the sense of Chapter 1, and it earns a special role.

Definition 10.1 (Order and control parameters)**.** An *order parameter* is a collective variable that distinguishes the qualitatively different patterns a system can adopt, changes abruptly when the system passes from one pattern to another, and evolves more slowly than the variables of the components. A *control parameter* is a variable that moves the system through its patterns without specifying which pattern it should adopt.

The relative phase satisfies all three conditions: it separates in-phase from anti-phase, it jumps from near π to near 0 at the transition, and it changes over many movement cycles while each finger's position changes within one. Movement frequency is the control parameter: it is not a prescription for either pattern, yet changing it changes which patterns are stable. The vocabulary comes from synergetics (Chapter 13), where order parameters govern the macroscopic behavior of lasers and fluids (Haken 1983b).

10.3 The HKB Model

Haken et al. (1985) proposed the simplest dynamics for the relative phase consistent with the observations.

Model 10.2 (Haken–Kelso–Bunz)**.** The relative phase ϕ descends a potential,

$$\dot{\phi} = -\frac{\mathrm{d}V}{\mathrm{d}\phi} = -a\sin\phi - 2b\sin 2\phi, \qquad V(\phi) = -a\cos\phi - b\cos 2\phi, \tag{10.1}$$

with $a, b > 0$. The ratio b/a decreases as movement frequency increases.

The form of the potential follows from two requirements: the dynamics must be periodic in ϕ and symmetric under exchanging the hands, $\phi \mapsto -\phi$. The two lowest Fourier terms that satisfy both are $\cos\phi$ and $\cos 2\phi$, and the model keeps exactly these.

Proposition 10.3 (Stability of the two patterns)**.** *The in-phase pattern $\phi = 0$ is a stable fixed point for all $a, b > 0$, with $V''(0) = a + 4b$. The anti-phase pattern $\phi = \pi$ has $V''(\pi) = 4b - a$ and is stable if and only if $b/a > 1/4$. For $b/a > 1/4$ two further fixed points at $\cos\phi^* = -a/4b$ are maxima of V*

separating the two basins; as b/a decreases to $1/4$ they converge on π and destroy its stability.

The proof is a direct computation: the fixed points satisfy $\sin\phi\,(a + 4b\cos\phi) = 0$, and the curvature of V at each gives its stability. Figure 10.1 shows the landscape. At high values of b/a, slow movements, the potential has two wells. As movement speeds up the anti-phase well becomes shallower and vanishes at $b/a = 1/4$, and the relative phase rolls into the in-phase well. Because the in-phase well never disappears, reversing the change of frequency leaves the system in phase, which is the hysteresis of the experiments.

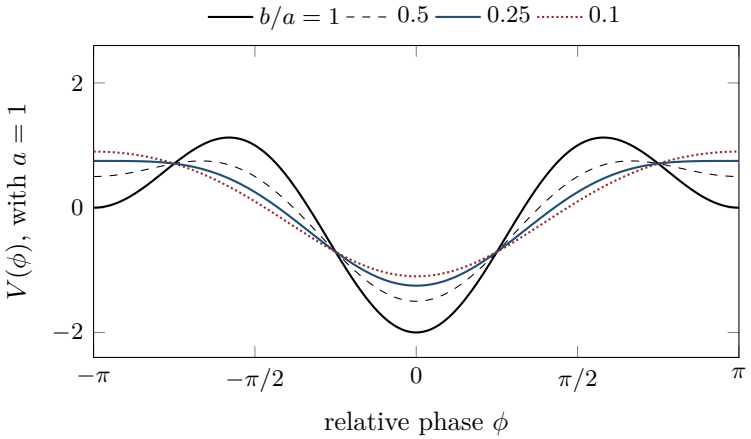


Figure 10.1: The HKB potential (10.1) for decreasing b/a, the direction of increasing movement frequency. The anti-phase minimum at $\phi = \pm\pi$ flattens and disappears at $b/a = 1/4$; the in-phase minimum at $\phi = 0$ persists throughout.

The model has two free parameters, and one combination of them, b/a, carries the whole transition. That is the Part II pattern again: a ratio of two magnitudes passing a critical value of order one (Chapter 5). The relation among the components (relative phase) is governed by a relation among the parameters (b/a).

The law's empirical content lies in its stability claims. The values of a and b are not measured directly; they are inferred from how the relative phase behaves, and any potential with the right two wells would fit the switch itself. What distinguishes the HKB model from a redescription of the switch is that it fixes how stability changes on the way to the switch, and those changes can be measured before anything switches.

10.4 From the Components Up

A law for the relative phase could be a description with nothing underneath it. Haken et al. (1985) showed that it need not be: an equation of the HKB

form follows from a model in which each hand is a self-sustained nonlinear oscillator and the two oscillators are coupled nonlinearly. Averaging the component dynamics over a cycle yields the relative-phase equation, with a and b depending on the coupling constants and on the amplitude of oscillation, which in the model falls as frequency rises. The collective law is therefore derivable from a lower level of description, which is the structure Schöner and J. A. S. Kelso (1988) made the basis of a general strategy: find the order parameter, write its low-dimensional dynamics, then derive that dynamics from the components and test the link.

The derivation also shows where the relational law is silent. The component model specifies that the oscillators are coupled and how the coupling depends on their states. It does not specify the physical channel of the coupling, which in a human being could be mechanical, proprioceptive, visual, or central. Different channels could produce the same values of a and b, and the relative-phase law would not tell them apart.

10.5 Fluctuations and Critical Slowing Down

The model's most important predictions concern what happens before the transition. Add noise to (10.1), as every real movement has:

$$\mathrm{d}\phi = -V'(\phi)\,\mathrm{d}t + \sqrt{Q}\,\mathrm{d}W_t, \tag{10.2}$$

where W_t is a Wiener process and Q sets the noise intensity.

Proposition 10.4 (Critical slowing and critical fluctuations)**.** *Linearized about the anti-phase pattern, with $\lambda = V''(\pi) = 4b - a > 0$, the relative phase relaxes after a perturbation with time constant*

$$\tau_{\mathrm{rel}} = \frac{1}{\lambda} = \frac{1}{4b - a}, \tag{10.3}$$

and fluctuates about π with stationary variance $Q/2\lambda$. Both diverge as $b/a \to 1/4$.

The proposition is the standard result for a linear stochastic equation with restoring rate λ (Haken 1983b). Its meaning for the experiment is sharp. Long before the switch, while the anti-phase pattern is still stable, two measurable quantities should grow as the movement frequency increases: the time the relative phase takes to recover from a small perturbation, and the size of its spontaneous fluctuations. Neither prediction follows from a switch mechanism that simply triggers the in-phase pattern at a threshold, because such a mechanism says nothing about the stability of the pattern it replaces. Both were confirmed. J. A. S. Kelso et al. (1986) reported enhanced fluctuations of relative phase approaching the transition, and Scholz et al. (1987) reported critical slowing down of the recovery from perturbations, together

with measurements of switching time. Schöner and J. A. S. Kelso (1988) placed these results in the framework of nonequilibrium phase transitions, with order-parameter dynamics, stability, fluctuations, and time scales as the observables.

Example 10.5 (Reading stability from noise). Suppose Q is fixed and measurements at three movement frequencies give $b/a = 1$, 0.5, and 0.3. With $a = 1$, equation (10.3) gives relaxation times of 1/3, 1, and 5 time units, and the standard deviation of the relative phase grows in proportion to $1/\sqrt{\lambda}$, by a factor of $\sqrt{15} \approx 3.9$ between the first and last condition. The illustrative numbers show the form of the prediction: the stability of the anti-phase pattern can be read from its fluctuations, and it degrades continuously before the pattern is lost.

10.6 Breaking the Symmetry

Two hands are nearly identical oscillators. A hand and a foot are not, nor are two people swinging pendulums of different lengths. The HKB model extends to such cases by adding a detuning $\delta\omega$, the difference between the components' preferred frequencies:

$$\dot{\phi} = \delta\omega - a \sin\phi - 2b \sin 2\phi. \tag{10.4}$$

The symmetry $\phi \mapsto -\phi$ is broken. The stable patterns shift away from exactly 0 and π, the anti-phase pattern loses stability sooner, and when $|\delta\omega|$ exceeds the largest value of $a \sin\phi + 2b \sin 2\phi$ there are no fixed points left (J. A. S. Kelso 1995; Tognoli and J. A. S. Kelso 2014). The relative phase then drifts, but not uniformly: it lingers near the phases where the fixed points used to be and passes quickly through the rest. That regime, coordination without stable patterns, is the subject of Chapter 11.

10.7 Coordination Between People

The bimanual transition could have been a fact about one nervous system, the product of crosstalk between the two hemispheres' motor commands. Schmidt et al. (1990) removed that possibility. Two seated participants swung their lower legs while watching each other, keeping a common tempo in either the symmetric pattern (relative phase 0°) or the alternating pattern (180°). The only coupling between them was visual. As the frequency of oscillation increased, the alternating pattern became less stable, and participants who had been instructed to stay in it showed evidence of switching to the symmetric pattern. The transition met the criteria of a bifurcation that the HKB model predicts, namely hysteresis, critical fluctuations, and divergence, and matched the transitions observed within a single person.

The experiment establishes two things. First, the relational law is not tied to a particular coupling channel: the same order parameter, with the same qualitative dynamics, appears when the two oscillators are limbs of one body and when they belong to two bodies joined only by sight. Second, in the interpersonal case the channel is known by construction. Each person's visual system must register the relation between the other's leg and their own, since nothing else links the two, and that registration must feed the motor system that adjusts the swing. The reader is not located in the brain by this experiment, but its sensory channel is fixed.

10.8 What Reads the Relative Phase?

The HKB law says how the relative phase evolves. For it to hold, the nervous system must somehow make the relative phase of the two hands available to the processes that change their movements, since the coupling terms depend on ϕ. The law itself is agnostic about how. At the level of the component model the dependence enters as coupling terms between two oscillators; at the level of the body it could run through proprioceptive feedback from the moving limbs, through vision of the fingers, or through crosstalk between the motor commands, and the relative-phase equation does not distinguish these possibilities.

The comparison with the electric fish is instructive. In the jamming avoidance response the relational variable is the frequency difference between the fish and a neighbor, and its reader is known down to identified populations of neurons (Chapter 21). The behavior also has a low-dimensional law at the level of the relation: with the stimulus frequency clamped to follow the fish's own, the dynamics of the frequency difference turns out to be locally unstable and globally nonlinear, with a snap-through bifurcation as a function of the difference (Madhav et al. 2013). Both cases are closed loops in which the relata are the animal's own rhythms and the relation feeds back on them. In bimanual coordination the loop is described at the level of the relation and its reader is unlocated; between two people the reader's sensory channel is fixed but its neural implementation is not; in the fish the loop is described at both levels. That is the difference Part V exploits.

Coordination dynamics also shows that a relational law can hold without fixing its implementation. If proprioceptive and visual coupling can produce the same potential, then the order parameter and its law are multiply realizable in the sense examined in Chapter 24, and evidence for the law is not evidence for any particular reader.

10.9 Audit

Canonical phenomenon

People oscillating two fingers in anti-phase switch abruptly to in-phase when movement frequency passes a critical value, and do not switch back when it is lowered (J. A. Kelso 1984).

Minimal model

The HKB equation $\dot{\phi} = -a \sin \phi - 2b \sin 2\phi$, with anti-phase stable only for $b/a > 1/4$; its stochastic version (10.2), which predicts critical slowing and critical fluctuations.

What is relational here?

The relative phase ϕ of the two limbs, an order parameter; the ratio b/a of the model's two coupling magnitudes, which carries the transition.

What physically reads it?

Some process in the sensorimotor system must transduce the relative phase into corrections of each limb's timing; the law requires such a reader but does not say whether it is proprioceptive, visual, or central.

What existing field explains it?

Synergetics and the theory of nonequilibrium phase transitions (Haken 1983b; Haken et al. 1985; Schöner and J. A. S. Kelso 1988), developed for movement science as coordination dynamics (J. A. S. Kelso 1995) and extended to coordination between people (Schmidt et al. 1990). They explain everything the chapter reports.

What remains unexplained?

The location and nature of the reader. The relational law is established; the physical process that makes relative phase available to the motor system is not identified by it.

A discriminating experiment

Measure a and b (through relaxation times and fluctuations) while selectively removing candidate channels: vision of the hands, proprioceptive feedback degraded by tendon vibration, or crosstalk tested by moving one limb passively. If the critical ratio and the fluctuation growth change only when one channel is removed, that channel is the reader; if they survive the removal of each, the coupling is redundant or central.

10.10 Notes and References

The bimanual phase transition is reported by J. A. Kelso (1984), and its model by Haken et al. (1985). Critical fluctuations and critical slowing down of relative phase were measured by J. A. S. Kelso et al. (1986) and Scholz et al.

(1987); Schöner and J. A. S. Kelso (1988) set out the general framework of dynamic patterns in behavioral and neural systems. Schmidt et al. (1990) found the same transition between two people coupled only by vision. J. A. S. Kelso (1995) is the book-length account of coordination dynamics, including symmetry breaking and extensions beyond the two hands; Tognoli and J. A. S. Kelso (2014) connect the detuned model to metastable brain dynamics. The underlying mathematics of order parameters is Haken (1983b). The relation-level dynamics of the jamming avoidance response under closed-loop control is analyzed by Madhav et al. (2013).

10.11 Exercises

Exercise 10.1. Find all fixed points of (10.1) for $b/a = 1$ and for $b/a = 0.2$, classify their stability, and sketch the potential in each case. What kind of bifurcation occurs at $b/a = 1/4$?

Exercise 10.2. Suppose b/a decreases linearly from 1 to 0.1 over 200 time units and then increases back to 1, with $a = 1$ and small noise. Describe, without simulation, the trajectory of ϕ starting at π. Then simulate (10.2) with $Q = 0.02$ and compare.

Exercise 10.3. Linearize (10.2) about $\phi = \pi$ and derive the relaxation time and stationary variance of Proposition 10.4. Express the standard deviation of ϕ as a function of b/a and plot it for b/a between 0.26 and 1.

Exercise 10.4. In simulations of (10.2) with b/a held just above $1/4$, noise can carry the system out of the anti-phase well before the well disappears. Estimate how the mean escape time depends on the barrier height ΔV and on Q, and explain why switching in experiments can occur before the deterministic critical point.

Exercise 10.5. For (10.4) with $a = 1$ and $b = 0.5$, find numerically the largest $|\delta\omega|$ for which a stable fixed point near π exists, and the largest for which any fixed point exists. Plot the region of $(\delta\omega, b/a)$ in which the anti-phase pattern survives.

Exercise 10.6. In the interpersonal experiment of Section 10.7, the coupling runs through vision. Predict how the critical frequency and the growth of fluctuations should change if each participant sees the other's leg only intermittently, for instance with a strobe at a fixed rate. What result would show that the relevant variable is the perceived relative phase rather than the other person's movement as such?

Exercise 10.7. Design the experiment of the audit in detail: the manipulations, the measurements of relaxation time and fluctuation, the sample size needed to detect a change of 0.05 in the critical value of b/a, and the result that would locate the reader in proprioception.

11

Metastability

Brain regions come into and out of coordination. Recorded over seconds, pairs of regions show intervals in which their rhythms keep a steady phase relation, separated by intervals in which the relation drifts, and at no time is the whole system either locked together or fully independent. The word *metastability* names this condition, and it is used in two ways. In its precise sense it names a dynamical regime with no stable states in which a relation nevertheless lingers near certain values, because the remnants of vanished fixed points or a chain of saddles slow it down there. In its loose sense it names a balance between integration and segregation. The thesis of this chapter is that the precise sense makes quantitative predictions about how long a relation lingers and how it moves on, that these predictions distinguish metastability from its look-alikes, and that empirical claims about a metastable brain have to be held to them.

11.1 Integration Without Permanent Synchrony

Tognoli and J. A. S. Kelso (2014) state the problem from the side of neuroscience. Neural ensembles oscillate across a wide range of frequencies and are transiently coupled when people attend, perceive, think, and act; parts of the brain engage and disengage in time. Coordination dynamics proposes that this is metastability: tendencies of regions to express their own autonomy and specialized function (segregation) coexist with tendencies to couple and coordinate (integration), and neither wins permanently. Tognoli and Kelso illustrate the idea in continuous neural and behavioral recordings and note that it had yet to be demonstrated within a fully spatiotemporal treatment.

Two things make the proposal attractive. A system locked into global synchrony carries little information, because every part says the same thing; a system of independent parts integrates nothing. And the brain's functional repertoire appears to change on the timescale of perception and action, faster than any change in its anatomy. Both points are qualitative. The dynamical theory of the next sections makes them quantitative, beginning with the simplest system that shows the phenomenon.

11.2 The Ghost of a Fixed Point

Chapter 7 derived the Adler equation for the relative phase of an oscillator and its drive,

$$\dot{\phi} = \nu - K \sin \phi. \tag{11.1}$$

For $|\nu| \le K$ the relative phase locks. At $\nu = K$ the stable and unstable fixed points meet at $\phi = \pi/2$ and annihilate, in a saddle-node bifurcation. Just beyond it, for ν slightly larger than K, there are no fixed points, but the velocity $\nu - K \sin \phi$ is still very small near $\pi/2$. The relative phase spends a long time crawling through that region and then slips rapidly around the rest of the circle (Figure 11.1). The vanished fixed point still shapes the motion. Dynamicists call such a region a *ghost* or a *bottleneck* (Strogatz 2015).

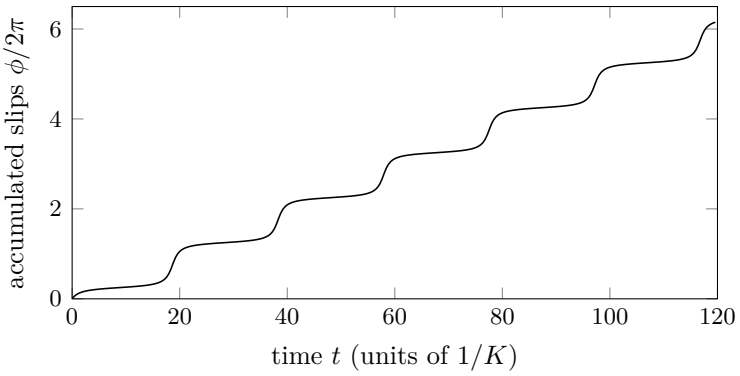


Figure 11.1: Metastable coordination in the Adler equation (11.1) just beyond the locking range, with $\nu = 1.05\,K$ (our computation). The relative phase dwells for long intervals near the ghost of the vanished fixed point and then slips through a full cycle quickly. The interval between slips, about $19.6/K$, matches (11.2).

Proposition 11.1 (Dwell time near a saddle-node ghost)**.** *For $\nu > K > 0$, the relative phase of* (11.1) *slips with period*

$$T = \frac{2\pi}{\sqrt{\nu^2 - K^2}}, \tag{11.2}$$

almost all of which is spent near $\phi = \pi/2$. As $\nu \to K^+$, $T \approx 2\pi/\sqrt{2K(\nu - K)}$, so the dwell time diverges as the inverse square root of the distance past the bifurcation.

The square-root law is general. Near any saddle-node bifurcation the dynamics reduces to the normal form $\dot{x} = \varepsilon + c\,x^2$, and the time to pass through the bottleneck is $\pi/\sqrt{c\,\varepsilon}$ (Strogatz 2015). The law is the first quantitative signature of metastability in the precise sense.

Definition 11.2 (Metastable coordination)**.** A system of interacting oscillators is in a regime of *metastable coordination* when its relative phases have no stable fixed points, yet each relative phase spends long intervals near a small set of values and passes quickly between them, the lingering being produced by the remnants of fixed points or by saddles of the dynamics.

The definition makes metastability a property of the relations, not of the oscillators. Each oscillator can run steadily throughout; what dwells and switches is the phase relation between them.

11.3 Coordination Without Stable Patterns

The detuned HKB model of Chapter 10,

$$\dot{\phi} = \delta\omega - a\sin\phi - 2b\sin 2\phi, \tag{11.3}$$

shows the same structure with two ghosts instead of one. When the detuning $\delta\omega$ exceeds the largest value of $a\sin\phi + 2b\sin 2\phi$, both the in-phase and the anti-phase patterns have disappeared. The relative phase wraps continually, but slowly near the phases where the two patterns used to be and quickly elsewhere, so a histogram of ϕ over a long recording shows two peaks even though no pattern is stable (J. A. S. Kelso 1995; Tognoli and J. A. S. Kelso 2014). Tognoli and Kelso take this regime as the dynamical core of the metastable brain: coordination tendencies without coordination states.

The model makes a prediction that the loose sense of the word does not. If the detuning is controlled experimentally, for instance by changing the frequency of one of two coupled rhythms, the dwell times near each former pattern should scale as the inverse square root of the distance past its own saddle-node bifurcation, and the dwell near the weaker pattern (anti-phase) should vanish first. A system that merely balances integration against segregation, with no ghosts, predicts neither the scaling nor the order.

11.4 Two Ways to Dwell

Intermittent dwelling and switching has a second common cause, and the two must be told apart.

In the first, described above, the dynamics has no stable states, and the dwelling is deterministic: slowing near a ghost. In the second, the dynamics has two or more stable states, and noise carries the system from one to another over a barrier. The HKB model with its two wells, driven by noise at a movement frequency where both patterns are stable, is an example (Chapter 10). Noise-driven transitions of this kind, called noise-induced phase slips when the variable is a phase, are studied throughout synchronization theory (Pikovsky et al. 2001).

Proposition 11.3 (Signatures of the two mechanisms)**.** *The two mechanisms predict different dwell-time statistics. Near a saddle-node ghost with weak noise, dwell times are nearly regular, and their mean scales as* $\varepsilon^{-1/2}$ *in the distance* ε *past the bifurcation. For noise-driven escape over a barrier of height* ΔV *with noise intensity* D*, dwell times are approximately exponentially distributed, and their mean grows exponentially, as* $\exp(\Delta V/D)$*, when the barrier grows or the noise falls.*

The second statement is the standard escape-rate result for noise-activated transitions, and its consequence for data is direct. A histogram of dwell times that is peaked away from zero, with a mean that follows a square-root law as a control parameter moves, indicates ghosts. A histogram that is exponential, with a mean that changes steeply with noise level, indicates barrier crossing between genuinely stable states. Both produce a relative phase that lingers and switches, and a time series that "looks metastable" is compatible with either until its dwell statistics have been measured against a control parameter.

11.5 Heteroclinic Channels

Ghosts give lingering near a few values. Neural activity often shows something richer: reproducible sequences of transient states, visited in a fixed order. Rabinovich et al. (2008) argued that neural computation relies on such transients rather than on convergence to attractors, and proposed that their dynamical basis is a *stable heteroclinic channel*: a chain of saddle states, each connected to the next by its unstable direction, such that trajectories starting near the chain follow it from saddle to saddle. Each saddle attracts along most directions and repels along one, so the system approaches a state, lingers, and leaves in a definite direction toward the next.

Model 11.4 (Winnerless competition)**.** A population of n competing units with activities $x_i \geq 0$ obeys the generalized Lotka–Volterra equations

$$\dot{x}_i = x_i\Big(\sigma_i - \sum_{j=1}^{n} \rho_{ij} x_j\Big) + \eta_i(t), \tag{11.4}$$

with growth rates $\sigma_i > 0$, inhibition strengths $\rho_{ij} > 0$, and small noise η_i. When the inhibition is asymmetric, so that each unit is suppressed most strongly by the one that follows it in a cycle, the states in which one unit dominates are saddles connected in sequence, and the activity passes from one dominant unit to the next without any unit winning permanently.

Without noise the passage around an exact heteroclinic cycle slows down on every visit, because the trajectory approaches the saddles ever more closely. With small noise the dwell times settle to a finite, reproducible value set by the noise and the saddles' eigenvalues, and the sequence itself stays fixed. The

signature of a heteroclinic channel is therefore a reproducible order of states with variable timing, which differs from the signature of ghosts (a few preferred values, square-root scaling) and from that of noisy multistability (exponential dwell times, no fixed order).

11.6 Metastability in Brain Models

K. J. Friston (1997) studied simulated neural systems composed of populations coupled by extrinsic connections and found that the richest behavior arose when those connections were sparse. The activity then showed intermittency, the recurrent and self-limiting expression of stereotyped transients. The dynamics conformed to a single complex attractor, yet the transients gave the appearance of an attractor that changed over time; Friston called this metastability and measured it with an index based on the entropy of the signal's spectral density. Deco and Kringelbach (2016) connected metastability to the communication-through-coherence hypothesis, according to which two neural groups communicate most effectively when their fluctuations in excitability are coordinated in time. Drawing on whole-brain computational models of neuroimaging data, they linked such coherence to metastability, understood as the exploration of a repertoire of functional configurations made possible by the brain's structural connectivity.

Example 11.5 (Quantifying variability of synchrony). Given phases $\theta_j(t)$ for N regions, the time-resolved order parameter $r(t) = |N^{-1} \sum_j \mathrm{e}^{\mathrm{i}\theta_j(t)}|$ (Chapter 9) measures global coherence at each moment. Its temporal standard deviation is a natural index of how much coherence rises and falls. The index has to be read with care. A finite population of oscillators produces fluctuations of $r(t)$ of order $N^{-1/2}$ with no metastable dynamics at all, and estimated phases add fluctuations of their own (Appendix K), so the index needs a null distribution from surrogate data that preserve each signal's spectrum (Appendix J). Even when it exceeds that null, a large index does not say which of the mechanisms of Sections 11.4 and 11.5 produced it.

11.7 Two Senses of the Word

The loose sense of metastability, a brain balanced between integration and segregation, is compatible with every one of the mechanisms above and with others besides. It describes; it does not predict. In the vocabulary of Part II it is a claim that the brain operates "in the middle," which Chapter 5 warns is usually too weak to be a law.

The precise senses predict. Ghosts predict square-root scaling of dwell times with the distance past a bifurcation and the order in which patterns lose their dwell. Heteroclinic channels predict reproducible sequences with variable timing and a specific dependence of dwell times on noise. Noisy multistability

predicts exponential dwell distributions. An empirical claim that a brain, a behavior, or a pair of rhythms is metastable earns content by naming which mechanism it means and showing the corresponding signature, ideally under a manipulation of the control parameter. The chapter's audit turns this into an experiment.

The same discipline applies to the intuition with which this book began: that experience might depend on difference rather than on integration alone, a conscious state being the pattern of discrepancies among partially independent processes that never quite coincide. Stated loosely, the intuition is compatible with every mechanism above. Stated precisely, it becomes a claim about which mechanism produces the discrepancies, which process reads them, and what changes when they change; Chapter 45 examines the neural evidence for claims of this form.

11.8 What Reads a Metastable Relation?

In the communication-through-coherence hypothesis the relation between two rhythms matters because something reads it (Fries 2015; Deco and Kringelbach 2016). A receiving population is more excitable at some phases of its own rhythm than at others, so input arriving from a sending population has more effect when it arrives at the excitable phase. The receiver's excitability cycle is a physical reader of the timing relation between sender and receiver: it transduces a phase difference into a difference in effective influence. If the phase relation dwells and switches, the channels of communication open and close with it.

That reading is first-order: it responds to the current phase relation. Whether any process reads the pattern of dwelling and switching itself, a relation between relations over time, is a separate question, and no result cited in this chapter answers it. It is not an idle question. Part V shows an animal that reads a second-order relation: the electric fish obtains the sign of a frequency difference from the relative timing of two modulations, each of which is itself a relation between its own field and a neighbor's (Chapter 21). So second-order relations can be read by physical systems, and the evidence required to show it is known. For metastable brain dynamics that evidence would be a process whose activity depends on dwell times or switching order, and whose disruption changes behavior while the first-order phase relations are left intact.

11.9 Audit

Canonical phenomenon

Neural and behavioral recordings in which pairs of rhythms repeatedly

approach a steady phase relation, linger, and drift away, without settling into any fixed pattern (Tognoli and J. A. S. Kelso 2014).

Minimal model

The Adler and detuned HKB equations beyond their bifurcations, with dwell times $T \propto \varepsilon^{-1/2}$; stable heteroclinic channels in winnerless competition (11.4); noise-driven escape between stable states as the alternative.

What is relational here?

The relative phase between rhythms, and the second-order pattern of its dwelling and switching over time.

What physically reads it?

For first-order relations, the excitability cycle of a receiving population, which converts a phase relation into effective influence (Fries 2015). For the second-order pattern, no reader has been identified.

What existing field explains it?

Nonlinear dynamics (saddle-node bottlenecks, heteroclinic cycles, noise-induced transitions) (Strogatz 2015; Pikovsky et al. 2001; Rabinovich et al. 2008); coordination dynamics (J. A. S. Kelso 1995; Tognoli and J. A. S. Kelso 2014); computational neuroscience of large-scale brain models (K. J. Friston 1997; Deco and Kringelbach 2016). They cover the chapter.

What remains unexplained?

Whether the brain's metastability is of the ghost type, the heteroclinic type, or noisy multistability, and whether any neural process reads the dwell-and-switch pattern itself. Both are empirical questions with known tests.

A discriminating experiment

Drive one of two coupled neural rhythms (by rhythmic sensory or transcranial stimulation) at a controlled detuning just beyond the locking range, and measure dwell times of the relative phase as the detuning is stepped. Square-root scaling identifies ghosts; exponential dwell distributions whose mean depends steeply on noise identify barrier crossing; a reproducible order of states across several regions with variable timing identifies a heteroclinic channel.

11.10 Notes and References

Tognoli and J. A. S. Kelso (2014) give the coordination-dynamics account of the metastable brain; J. A. S. Kelso (1995) develops its theoretical basis in the detuned relative-phase equation. Saddle-node bottlenecks and their square-root scaling are treated in Strogatz (2015); noise-induced phase slips in Pikovsky et al. (2001). Rabinovich et al. (2008) argue for transient dynamics and stable

heteroclinic channels as the basis of neural processing. K. J. Friston (1997) introduced metastability into the analysis of simulated neuronal dynamics. Deco and Kringelbach (2016) connect metastability in whole-brain models to communication through coherence, whose general statement is Fries (2015). Estimation of phases and synchrony from recordings is covered in Appendices B and K, and surrogate nulls in Appendix J.

11.11 Exercises

Exercise 11.1. Derive (11.2) by integrating $\mathrm{d}t = \mathrm{d}\phi/(\nu - K\sin\phi)$ over one cycle. Show that for $\nu \to K^{+}$ the fraction of the period spent within ± 0.2 radians of $\pi/2$ tends to 1.

Exercise 11.2. For the normal form $\dot{x} = \varepsilon + x^{2}$, compute the time to travel from $x = -\infty$ to $x = +\infty$ and confirm the $\varepsilon^{-1/2}$ law. Show that the Adler equation near $\phi = \pi/2$ reduces to this form, and identify ε and the coefficient c in terms of ν and K.

Exercise 11.3. Simulate (11.3) with $a = 1$, $b = 0.5$, weak noise, and $\delta\omega$ just above the largest value of $a\sin\phi + 2b\sin 2\phi$. Plot a histogram of ϕ and of the dwell times near each peak. Repeat for three values of $\delta\omega$ and check the scaling of the mean dwell near the in-phase ghost.

Exercise 11.4. Simulate the noisy HKB model with $b/a = 0.6$, where both patterns are stable, for several noise intensities. Plot the distribution of dwell times in the anti-phase well and its mean against $1/Q$. Which signature of Proposition 11.3 do you find?

Exercise 11.5. Simulate (11.4) for $n = 3$ with $\sigma_i = 1$, $\rho_{ii} = 1$, and a cyclic asymmetry in the off-diagonal ρ_{ij} of your choice that produces a heteroclinic cycle. Show that without noise the dwell times grow on each passage, and that with noise of amplitude 10^{-6} they settle to a stable value while the order of states is preserved.

Exercise 11.6. For $N = 20$ Kuramoto oscillators below the synchronization threshold, compute the temporal standard deviation of $r(t)$. How large is it, and how does it scale with N? Explain why a metastability index computed from twenty brain regions needs a surrogate null before it can support any claim.

Exercise 11.7. Design an experiment that tests whether a receiving neural population reads the phase relation to a sender: stimulate the sender at controlled phases of the receiver's rhythm and state the result that would show phase-dependent influence. How would you extend the design to test whether any process reads the dwell-and-switch pattern of that relation?

12

Chimera States and Frustration

Take a ring of identical oscillators, each coupled to the others by the same rule, and start them from a random state. Two outcomes look natural: the ring locks into a common rhythm, or every oscillator keeps drifting against every other. In 2002 Kuramoto and Battogtokh reported a third. Part of the ring locked into a coherent block while the rest drifted incoherently, although every oscillator and every connection was the same (Kuramoto and Battogtokh 2002). Abrams and Strogatz solved the state exactly two years later and named it a *chimera*, after the mythical animal assembled from incongruous parts (Abrams and Strogatz 2004).

The thesis of this chapter is that chimeras and their close relatives, frustrated states, are patterns of relations that the equations permit without displaying them. In a chimera the symmetry of the model is broken by the configuration of phase relations and by nothing in the parts. In a frustrated system the relations that each coupling prefers cannot all be realized at once, and the system settles into a compromise whose structure is again relational. Both belong squarely to nonlinear dynamics and statistical physics. Both have also been borrowed by neuroscience more freely than the evidence warrants, and the chapter ends by stating what the borrowing would have to show.

12.1 Identical Parts, Split Behavior

The model of Abrams and Strogatz (2004) places phase oscillators on a ring, with position $x \in [-\pi, \pi)$ and phase $\theta(x, t)$, and couples them nonlocally:

$$\frac{\partial \theta}{\partial t}(x,t) = \omega - \int_{-\pi}^{\pi} G(x - x') \, \sin\bigl(\theta(x,t) - \theta(x',t) + \alpha\bigr) \, \mathrm{d}x'. \tag{12.1}$$

Every oscillator has the same natural frequency ω. The kernel G depends only on distance along the ring, so the equations are unchanged by any rotation of the ring; Abrams and Strogatz (2004) take $G(x) = (1 + A\cos x)/2\pi$ with

$0 \le A \le 1$, so that near neighbors are coupled more strongly than distant ones. The phase lag α shifts the preferred phase relation between coupled oscillators away from zero. Kuramoto and Battogtokh (2002) used an exponentially decaying kernel on a line segment and found the same phenomenon; in both models chimeras appear when α is a little below $\pi/2$.

Definition 12.1 (Chimera state)**.** Consider a network of identical oscillators whose coupling is invariant under a group of permutations that acts transitively on the oscillators. A *chimera state* is a stable or long-lived state in which the oscillators split into a *coherent* subset, phase-locked and rotating at a common frequency, and an *incoherent* subset whose phases drift relative to the coherent one.

The definition carries the whole point. Transitivity means that every oscillator is equivalent to every other under a symmetry of the equations. As in Chapter 1, where invariance under a transitive group forced every frame-free quantity to be relational, nothing about an individual oscillator can explain why it ends up coherent or incoherent. Which oscillators do which is fixed by the initial configuration of phases, and a relabeling of the ring moves the coherent block elsewhere without changing the equations. The partition belongs to the configuration.

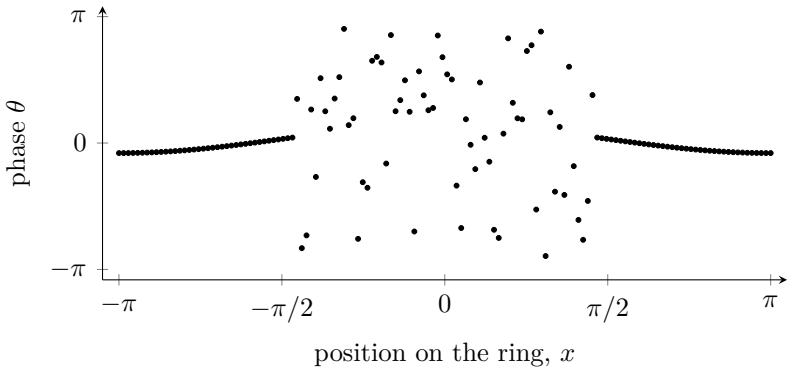


Figure 12.1: Schematic snapshot of a chimera on a ring (constructed for illustration, not simulated). Near $x = \pm\pi$ the phases vary smoothly: that block is coherent. Near $x = 0$ the phases are scattered and drift: that block is incoherent. Every oscillator obeys the same equation (12.1).

The mechanism is easiest to see through a relational variable that the coupling itself computes. Define the *local order parameter*

$$R(x,t)\,\mathrm{e}^{\mathrm{i}\Theta(x,t)} = \int_{-\pi}^{\pi} G(x-x')\,\mathrm{e}^{\mathrm{i}\theta(x',t)}\,\mathrm{d}x', \tag{12.2}$$

a weighted average of the phasors of the oscillators near x. Since $\int G(x -$

$x')\sin(\theta-\theta'+\alpha)\,\mathrm{d}x' = \mathrm{Im}\big[\mathrm{e}^{\mathrm{i}(\theta+\alpha)}\int G(x-x')\mathrm{e}^{-\mathrm{i}\theta'}\,\mathrm{d}x'\big]$, equation (12.1) becomes

$$\frac{\partial\theta}{\partial t} = \omega - R(x,t)\,\sin\big(\theta - \Theta(x,t) + \alpha\big). \tag{12.3}$$

Each oscillator is driven by its phase relative to the local field, and by nothing else. In a steadily rotating chimera, with pattern frequency Ω, write $\phi = \theta - \Omega t$ and suppose R and $\Theta - \Omega t$ are stationary. Then $\dot\phi = (\omega-\Omega) - R(x)\sin(\phi - \Theta(x) + \alpha)$, which has fixed points where $R(x) \geq |\omega - \Omega|$ and none where $R(x) < |\omega - \Omega|$. Where the local field is strong, oscillators lock to it; where it is weak, they drift. The field is in turn produced by the oscillators, so a chimera is a self-consistent solution in which a spatially varying $R(x)$ sustains the very split that produces it. Abrams and Strogatz (2004) solved the self-consistency equations exactly and showed that the stable chimera is born from a spatially modulated drift state and destroyed in a saddle-node bifurcation with an unstable chimera.

Remark 12.2. In finite rings the picture needs a qualification. Wolfrum and Omel'chenko (2011) showed that chimera states in finite systems of this type are chaotic transients that eventually collapse to full synchrony, with lifetimes that grow rapidly with the number of oscillators. The chimera of the continuum model is therefore the limit of increasingly long-lived transients. For most physical purposes the distinction does not matter; for claims about biological systems of modest size, it does.

12.2 The Minimal Chimera

Nonlocal coupling is not required. Abrams, Mirollo, et al. (2008) found chimeras in the simplest network that can support them: two equal populations of identical oscillators, coupled more strongly within each population than between them. With $\sigma \in \{1,2\}$ labeling the populations and N oscillators in each,

$$\begin{aligned} \dot\theta_i^\sigma &= \omega + \sum_{\sigma'=1}^{2} \frac{K_{\sigma\sigma'}}{N} \sum_{j=1}^{N} \sin(\theta_j^{\sigma'} - \theta_i^\sigma - \alpha), \\ K_{11} &= K_{22} = \mu, \qquad K_{12} = K_{21} = \nu, \end{aligned} \tag{12.4}$$

with $\mu > \nu \geq 0$ and the normalization $\mu + \nu = 1$. In terms of the population order parameters $r_\sigma \mathrm{e}^{\mathrm{i}\Phi_\sigma} = N^{-1}\sum_j \mathrm{e}^{\mathrm{i}\theta_j^\sigma}$, each oscillator feels

$$\dot\theta_i^\sigma = \omega + \mathrm{Im}\Big[\mathrm{e}^{-\mathrm{i}(\theta_i^\sigma+\alpha)}\big(\mu\, r_\sigma \mathrm{e}^{\mathrm{i}\Phi_\sigma} + \nu\, r_{\sigma'}\mathrm{e}^{\mathrm{i}\Phi_{\sigma'}}\big)\Big], \tag{12.5}$$

where σ' is the other population. For phase lags a little below $\pi/2$ and a range of the coupling contrast $A = \mu - \nu$, the system has chimera states in which one population is perfectly synchronized ($r_1 = 1$) and the other is only partially so ($r_2 < 1$), in stationary and in breathing forms (Abrams, Mirollo, et al. 2008).

In the limit of large N the dynamics of the order parameters reduce exactly to a small set of ordinary differential equations by the method of Ott and Antonsen (2008), which is why the model is called solvable (Appendix E).

Example 12.3 (Chimeras in the laboratory). Chimeras were first produced experimentally a decade after their discovery in simulation. Tinsley et al. (2012) coupled populations of photosensitive Belousov–Zhabotinsky oscillators, catalyst-loaded particles whose phases could be manipulated by controlling each particle's exposure to light, and observed coexisting synchronized and unsynchronized subpopulations although every oscillator was coupled in an equivalent way. Martens et al. (2013) built the two-population architecture from metronomes: two swings, each carrying fifteen metronomes, joined by a spring. Within each swing the metronomes are coupled through the swing's motion; the spring couples the swings. For a range of spring stiffness the symmetry broke: the metronomes on one swing ticked in lockstep while those on the other moved erratically, although all were set to the same tempo. Their model traces the chimera to a competition between two antagonistic synchronization patterns.

12.3 Frustration

A chimera breaks a symmetry the equations respect. Frustration is the complementary situation: the equations ask for relations that no configuration can supply.

Definition 12.4 (Frustration)**.** A system of coupled components is *frustrated* when the relation preferred by each coupling, taken alone, cannot be realized simultaneously for all couplings by any configuration.

The term comes from the physics of magnets. In an antiferromagnet each bond prefers the two spins it joins to point in opposite directions. On a square lattice every bond can be satisfied at once. On a triangle it cannot: once two spins are antiparallel, the third agrees with one of them, and at most two of the three bonds can be satisfied. Wannier (1950) showed that the Ising antiferromagnet on the triangular lattice keeps a finite entropy down to zero temperature because of the many equally good compromises, and frustration became a central concept in the theory of spin glasses (Binder and A. P. Young 1986).

The oscillator version shows the relational structure directly. Three identical phase oscillators with repulsive global coupling obey

$$\dot{\theta}_i = \omega - \frac{K}{3}\sum_{j=1}^{3}\sin(\theta_j - \theta_i), \qquad K > 0. \tag{12.6}$$

The dynamics of the phase differences are a gradient flow, $\dot{\theta}_i = \omega - \partial U/\partial\theta_i$, on the potential

$$U = \frac{K}{6}\sum_{i,j}\cos(\theta_j - \theta_i) = \frac{K}{6}\Big|\sum_j \mathrm{e}^{\mathrm{i}\theta_j}\Big|^2. \tag{12.7}$$

Each pair, taken alone, would minimize its share of U by sitting in anti-phase, a phase difference of π. The three phase differences around the triangle are not independent, however:

$$(\theta_1 - \theta_2) + (\theta_2 - \theta_3) + (\theta_3 - \theta_1) = 0, \tag{12.8}$$

and three differences of π would sum to 3π, which is not a multiple of 2π. The minimum of U is reached when the three phasors cancel, which forces an equilateral arrangement with every pairwise difference equal to $\pm 2\pi/3$. Each pair is left at $\cos(2\pi/3) = -1/2$ instead of the -1 it would prefer.

Proposition 12.5 (Cycle constraint)**.** *For phase oscillators on any network, the phase differences $\theta_i - \theta_j$ taken around a closed cycle of the network sum to a multiple of 2π. Couplings whose preferred phase differences violate this constraint on some cycle cannot all be satisfied, and the network is frustrated.*

Proof. Around a cycle $i_1 \to i_2 \to \cdots \to i_k \to i_1$ the sum of differences telescopes to zero when phases are treated as real numbers, and to a multiple of 2π when each difference is reduced modulo 2π. If each coupling on the cycle prefers a difference $\delta_{i_m i_{m+1}}$ with $\sum_m \delta_{i_m i_{m+1}} \not\equiv 0 \pmod{2\pi}$, no configuration realizes all of them. □

The proposition is elementary, and it is the sharpest statement in this chapter about relations as such. Relations among the parts of a configuration are not free variables. They obey constraints that no single relation reveals and that only appear when relations are composed around cycles. The same structure reappears, in another guise, in Chapter 1's exclusive-or example, where three variables were jointly constrained although every pair was independent.

Two further sources of frustration matter for oscillators. The first is a phase lag. In the Sakaguchi–Kuramoto model, $\dot{\theta}_i = \omega_i + (K/N)\sum_j \sin(\theta_j - \theta_i - \alpha)$, each coupling prefers a nonzero phase difference, which cannot hold in both directions for every pair (Sakaguchi and Kuramoto 1986). The lag shifts the collective frequency and, near $\alpha = \pi/2$, brings the coupling close to the boundary between attraction and repulsion, where the chimeras of Section 12.1 live. Phase lags of this kind are not exotic; they arise whenever the oscillators' frequencies depend on their amplitudes (Appendix D). The second is topology. On a ring of N oscillators with nearest-neighbor coupling, $\dot{\theta}_j = \omega + K[\sin(\theta_{j+1} - \theta_j) + \sin(\theta_{j-1} - \theta_j)]$, the *twisted states* $\theta_j = 2\pi q j/N + \omega t$ are solutions for every integer q. Their *winding number*

$$q = \frac{1}{2\pi}\sum_{j=1}^{N}\big[\theta_{j+1} - \theta_j\big]_{2\pi}, \tag{12.9}$$

where $[\cdot]_{2\pi}$ reduces a difference to the interval $(-\pi, \pi]$, is a relational quantity with a topological character: it is an integer, so it cannot change under small continuous changes of the phases, and it changes only through a *phase slip*, in which some adjacent difference passes through π. Linearizing about a twisted state gives eigenvalues $\lambda_m = -4K\cos(2\pi q/N)\sin^2(\pi m/N)$, so for $K > 0$ the state is stable exactly when $|q| < N/4$ (Exercise 12.4). Twisted states with small winding numbers therefore coexist with full synchrony as stable alternatives, and Wiley et al. (2006) measured the relative sizes of their basins of attraction. A ring of identical oscillators can hold a permanent twist, a relation around the whole ring that no local measurement detects.

12.4 Chimeras in the Brain?

Brains show coexisting synchronized and desynchronized activity at many scales, and the word "chimera" has become a popular label for such coexistence. The most cited case is unihemispheric slow-wave sleep. In dolphins and seals, and in birds, one cerebral hemisphere can show the electroencephalographic signs of slow-wave sleep while the other shows those of waking (Rattenborg et al. 2000). Two hemispheres, one synchronized and one not: the resemblance to the two-population model of Section 12.2 is plain, and reviews of chimera theory list unihemispheric sleep among its motivations (Panaggio and Abrams 2015).

The resemblance is an analogy, and it stops short of a demonstration for three reasons.

First, coexistence of coherence and incoherence is generic in populations whose parts differ. In the Kuramoto model with a spread of natural frequencies, above the locking threshold the oscillators whose natural frequencies lie within Kr of the collective frequency lock and the rest drift (Kuramoto 1984; Strogatz 2000). That is a coexistence of synchrony and asynchrony produced entirely by differences among the parts. A chimera produces the same picture with no such differences. Observing the picture therefore says nothing about which mechanism produced it.

Second, the brain's hemispheres are not identical, identically driven oscillator populations. Unihemispheric sleep is accompanied by behavioral asymmetries, such as keeping open the eye connected to the waking hemisphere, and by asymmetric neuromodulatory states (Rattenborg et al. 2000). Those asymmetries break the symmetry that the chimera explanation needs to be broken by the configuration alone.

Third, the chimera is a precise dynamical object with signatures beyond coexistence: a partition set by initial conditions, a coherent group rotating at a frequency distinct from the mean drift of the incoherent group, and dependence on a phase lag near $\pi/2$. A claim that a biological pattern is a chimera should test at least one of these.

Hypothesis 12.6 (What a neural chimera would have to show)**.** If a pattern of coexisting synchronized and desynchronized neural activity is a chimera in the technical sense, then (i) the pattern persists when differences among the participating units are experimentally reduced; (ii) which units belong to the coherent group can be changed by perturbing the initial state, without changing the units; and (iii) the coherent group's frequency differs from the mean frequency of the incoherent group in the way the self-consistency analysis predicts. A pattern that fails (i) or (ii) is partial synchrony produced by heterogeneity.

None of this makes chimera theory irrelevant to biology. It makes the theory a source of precise alternatives, which is the most useful thing a dynamical model can offer an experimentalist.

12.5 What the Coupling Reads

The chapter's formal core, equation (12.3), contains the first appearance in this book of a relational variable that is physically explicit in the dynamics of a part. Each oscillator is driven by $R\sin(\theta - \Theta + \alpha)$: by its own phase relative to the local field. The local order parameter is a relation among many oscillators, and the coupling medium, whatever it is physically (a shared swing, a chemical signal, light), computes it and delivers it to every oscillator. In the language of Chapter 1, the coupling is the operation that makes the relation available to the parts.

It is available in the thinnest sense. The oscillator's rate of change depends on the relation; the oscillator does nothing with it, and nothing in the system categorizes, stores, or acts on the relation in order to change it. Chapter 15 separates this minimal sense, sensitivity, from detection, discrimination, control, and interpretation. Chimeras and frustrated states show how rich the relational structure of identical parts can be before any of those further capacities appear.

12.6 Audit

Canonical phenomenon

Two swings, each carrying fifteen identical metronomes and joined by a spring: for a range of spring stiffness the metronomes on one swing lock while those on the other drift (Martens et al. 2013). On the side of frustration, three repulsively coupled oscillators settle at mutual phase differences of $2\pi/3$ because anti-phase cannot hold around the triangle.

Minimal model

The nonlocal ring (12.1) and its mean-field form (12.3); the two-population model (12.4); the cycle constraint (Proposition 12.5) and the winding number (12.9).

What is relational here?
The local order parameter $Re^{i\Theta}$, each oscillator's phase relative to it, the partition of the population into coherent and incoherent sets, and the sums of phase differences around cycles.

What physically reads it?
The coupling medium computes the local field and delivers each oscillator's phase relative to it; the oscillators' dynamics depend on that relation. Nothing reads the partition itself, except the experimenter.

What existing field explains it?
Nonlinear dynamics and synchronization theory (Kuramoto and Battogtokh 2002; Abrams and Strogatz 2004; Abrams, Mirollo, et al. 2008; Ott and Antonsen 2008; Panaggio and Abrams 2015; Wolfrum and Omel'chenko 2011), the statistical physics of frustration (Wannier 1950; Binder and A. P. Young 1986), and the dynamics of twisted states (Wiley et al. 2006). They explain everything in the chapter.

What remains unexplained?
Nothing in the physics. An empirical question remains open: whether any biological coexistence of synchrony and asynchrony is a chimera in the technical sense.

A discriminating experiment
In a preparation that shows coexisting synchronized and desynchronized groups (cultured neuronal networks, coupled cardiac cell populations), apply Hypothesis 12.6: reduce the heterogeneity of the units, and perturb the initial state to see whether the coherent group can be moved. A partition that moves with the initial state and survives homogenization is a chimera; a partition that tracks the units' intrinsic properties is heterogeneity-driven partial synchrony.

12.7 Notes and References

The phenomenon was reported by Kuramoto and Battogtokh (2002); Abrams and Strogatz (2004) gave the exact solution for the cosine kernel and the name, and Abrams, Mirollo, et al. (2008) the solvable two-population model. Panaggio and Abrams (2015) is the standard review, and covers the many later variants (spiral-wave chimeras in two dimensions, amplitude chimeras, chimeras in networks with delays). Wolfrum and Omel'chenko (2011) is essential on the status of chimeras in finite systems. The experiments of Tinsley et al. (2012) and Martens et al. (2013) were the first laboratory realizations with chemical and mechanical oscillators. The Ott–Antonsen reduction that makes the two-population model solvable is due to Ott and Antonsen (2008).

Frustration entered physics through Wannier's analysis of the triangular Ising antiferromagnet (Wannier 1950); Binder and A. P. Young (1986) review its role in spin glasses. The phase-lag model is due to Sakaguchi and Kuramoto (1986). Twisted states and the sizes of their basins are studied by Wiley et al. (2006). For the Kuramoto model with heterogeneous frequencies, including partial locking, see Kuramoto (1984), Strogatz (2000), and Acebrón et al. (2005); for synchronization generally, Pikovsky et al. (2001). On unihemispheric sleep, Rattenborg et al. (2000) is the review we rely on.

12.8 Exercises

Exercise 12.1. Derive (12.3) from (12.1) and (12.2). For $G(x) = (1 + A\cos x)/2\pi$, show that $Re^{i\Theta}$ depends on only two complex moments of the phase distribution, $\int e^{i\theta(x')}\,dx'$ and $\int \cos x'\, e^{i\theta(x')}\,dx'$ (with a third moment in $\sin x'$), and explain why this makes an exact solution possible.

Exercise 12.2. For two identical oscillators with phase lag, $\dot\theta_1 = \omega + K\sin(\theta_2 - \theta_1 - \alpha)$ and $\dot\theta_2 = \omega + K\sin(\theta_1 - \theta_2 - \alpha)$, write the equation for $\phi = \theta_1 - \theta_2$. Find its fixed points and their stability as functions of α, and compute the common frequency of the in-phase state. State the condition on α under which the coupling is effectively repulsive.

Exercise 12.3. For the repulsive triangle (12.6), reduce the dynamics to the two phase differences $\phi_1 = \theta_1 - \theta_2$ and $\phi_2 = \theta_2 - \theta_3$. Find all fixed points, classify them, and verify that the stable ones have $\phi_1 = \phi_2 = \pm 2\pi/3$. Repeat for four oscillators on a ring with nearest-neighbor repulsive coupling and explain why that network is not frustrated.

Exercise 12.4. For the nearest-neighbor ring, verify that $\theta_j = 2\pi q j/N + \omega t$ is a solution, linearize about it, and derive the eigenvalues

$$\lambda_m = -4K\cos(2\pi q/N)\,\sin^2(\pi m/N), \qquad m = 0, 1, \dots, N-1.$$

Conclude that the state is stable for $|q| < N/4$. Simulate $N = 40$ oscillators from random initial conditions many times and record the winding number of the final state.

Exercise 12.5. In the Kuramoto model with natural frequencies drawn from a Lorentzian of half-width γ, the stationary order parameter above threshold is $r = \sqrt{1 - 2\gamma/K}$. Taking this result as given, compute the fraction of oscillators that are locked as a function of K/γ. Explain in one paragraph why a snapshot of this state and a snapshot of a chimera can look the same, and what measurement distinguishes them.

Exercise 12.6. Simulate the two-population model (12.4) with $N = 64$, $\alpha = \pi/2 - 0.1$, and several values of $A = \mu - \nu$ between 0.1 and 0.4. Plot $r_1(t)$ and $r_2(t)$. Starting from initial conditions in which population 1 is nearly

synchronized and population 2 is random, find a chimera; then swap the initial conditions and confirm that the other population becomes the coherent one.

Exercise 12.7. Design an experiment on a cultured neuronal network grown on a multi-electrode array that tests Hypothesis 12.6. Specify how you would reduce heterogeneity among the recorded units, how you would perturb the initial state, and what result would lead you to conclude that the observed pattern is not a chimera.

13

Synergetics

A laser below threshold is a lamp. Its atoms emit independently, the phases of their contributions are uncorrelated, and the light is incoherent. Raise the pump past a threshold and the output changes character: one mode of the cavity field grows, the atoms' emissions fall into step with it, and the beam becomes coherent. Nothing about any single atom changed at threshold. What changed is a relation among all of them, and one variable, the amplitude of the growing mode, came to describe the whole.

Hermann Haken built a general theory around this observation and called it *synergetics* (Haken 1983b; Haken 1983a). Its central claim is that near an instability the few modes that become unstable evolve slowly, the many modes that remain stable relax quickly and follow them, and the macroscopic behavior of the system is captured by low-dimensional equations for the slow modes. Haken called the slow amplitudes *order parameters* and the dependence of the fast modes on them the *slaving principle*. The thesis of this chapter is that the slaving principle is a theorem about timescale separation with a clear domain of validity, and that its causal vocabulary ("enslavement," "circular causality") describes that separation without adding a new kind of cause. The chapter works through a two-mode example, the laser, and the coordination of limbs, and then states where the vocabulary helps and where it misleads.

13.1 Instability and the Separation of Timescales

Consider a system $\dot{\boldsymbol{u}} = \boldsymbol{N}(\boldsymbol{u}; \lambda)$ with a steady state $\boldsymbol{u}_0$ that depends on a control parameter λ (a pump rate, a temperature difference, a movement frequency). Linearizing about the steady state, $\boldsymbol{u} = \boldsymbol{u}_0 + \sum_k \xi_k(t)\, \boldsymbol{v}_k$, decomposes small deviations into modes $\boldsymbol{v}_k$ that grow or decay independently at rates given by the eigenvalues λ_k of the Jacobian. Far from any instability, all eigenvalues have negative real parts and every mode decays. As λ approaches a critical value, the real part of one eigenvalue, or of a few, approaches zero. Those modes decay ever more slowly and, past the critical value, grow. The other modes keep decaying at rates that stay finite.

This spectral structure is the precondition for everything that follows. Near the instability there are two widely separated timescales: the slow timescale $1/|\lambda_s|$ of the critical modes, which diverges at threshold, and the fast timescale $1/\gamma$, where γ bounds the decay rates of all other modes from below.

Definition 13.1 (Order parameter)**.** Near an instability whose linearization has a spectral gap, an *order parameter* is the amplitude of a critical mode: a collective variable whose dynamics is slow compared with the relaxation of all other modes, and to which the amplitudes of those modes become asymptotically tied.

The term comes from the theory of equilibrium phase transitions, where an order parameter measures the degree of order that appears below a critical temperature, such as the magnetization of a ferromagnet. Synergetics extends it to systems driven far from equilibrium, where the transition is an instability of a steady state and not a minimum of a free energy.

An order parameter is a relational variable in the sense of Chapter 1. The amplitude of a mode $\boldsymbol{v}_k$ is a projection of the whole configuration onto a pattern, $\xi_k = \langle \boldsymbol{w}_k, \boldsymbol{u} - \boldsymbol{u}_0 \rangle$ with $\boldsymbol{w}_k$ the adjoint mode: a weighted relation among all the components at once. The amplitude of a convection roll is a correlation between local fluid velocities and a spatial pattern; the Kuramoto order parameter $r = |N^{-1} \sum_j \mathrm{e}^{\mathrm{i}\theta_j}|$ (Chapter 9) is a relation among all the phases.

Example 13.2 (Convection rolls). A horizontal layer of fluid heated from below conducts heat without moving as long as the temperature difference across it is small. The control parameter is the Rayleigh number, a dimensionless ratio of buoyancy to viscous and thermal dissipation (Chapter 5). Above a critical value the conducting state becomes unstable and the fluid organizes into parallel convection rolls. Near onset the pattern is described by the amplitude of a single roll mode; the velocity and temperature fluctuations at all other wavelengths decay quickly and are determined by that amplitude (Haken 1983b). Billions of molecules, and a continuum of fluid modes, are summarized near onset by one number: how strongly the fluid's local motions correlate with the roll pattern.

13.2 Adiabatic Elimination: A Worked Example

The simplest system that shows slaving has one slow and one fast variable:

$$\dot{x} = \lambda x - xy, \qquad \dot{y} = -\gamma y + x^2, \qquad \gamma > 0,\ |\lambda| \ll \gamma. \tag{13.1}$$

The variable x is the critical mode: its linear growth rate λ is near zero. The variable y is stable, relaxing at rate γ, and is driven by x.

Because y relaxes much faster than x changes, after a transient of duration of order $1/\gamma$ the fast variable follows the slow one. Setting $\dot{y} \approx 0$ gives

$$y \approx \frac{x^2}{\gamma}, \tag{13.2}$$

and substituting into the equation for x gives a closed equation for the order parameter alone:

$$\dot{x} = \lambda x - \frac{x^3}{\gamma}. \tag{13.3}$$

This is the normal form of a supercritical pitchfork bifurcation (Appendix C). For $\lambda < 0$ the state $x = 0$ is stable. For $\lambda > 0$ it loses stability and two new states $x^* = \pm\sqrt{\lambda\gamma}$ appear, with $y^* = \lambda$. These are also the exact fixed points of (13.1), because at a fixed point $\dot{y} = 0$ holds exactly; the approximation concerns only the transients.

The procedure is called *adiabatic elimination*. Its mathematical justification is the existence of a *slow manifold*, an invariant curve $y = h(x)$ that attracts nearby trajectories at rate of order γ and on which the dynamics is slow. Its shape follows from requiring invariance: along the curve, $\dot{y} = h'(x)\,\dot{x}$, so

$$h'(x)\big(\lambda x - x\,h(x)\big) = -\gamma\,h(x) + x^2. \tag{13.4}$$

Expanding $h(x) = ax^2 + bx^4 + \cdots$ and matching powers of x gives

$$a = \frac{1}{\gamma + 2\lambda}, \qquad b = \frac{2a^2}{\gamma + 4\lambda}, \tag{13.5}$$

so that $h(x) = x^2/\gamma + O(\lambda x^2/\gamma^2, x^4/\gamma^3)$ and the adiabatic result (13.2) is the leading term (Exercise 13.1). At $\lambda = 0$ the slow manifold is the center manifold of the bifurcation, and its existence is guaranteed by the center manifold theorem (Guckenheimer and Holmes 1983; Kuznetsov 2004).

13.3 The Slaving Principle

The general statement has the same structure as the example.

Proposition 13.3 (Slaving near an instability)**.** *Let $\dot{\boldsymbol{u}} = \boldsymbol{L}\boldsymbol{u} + \boldsymbol{N}(\boldsymbol{u})$, with $\boldsymbol{N}$ smooth and at least quadratic, and suppose the spectrum of $\boldsymbol{L}$ splits into critical eigenvalues with real parts near zero (modes $\boldsymbol{s}$) and stable eigenvalues with real parts at most $-\gamma < 0$ (modes $\boldsymbol{f}$). Then, in a neighborhood of $\boldsymbol{u} = 0$, there is an invariant manifold $\boldsymbol{f} = \boldsymbol{h}(\boldsymbol{s})$ that attracts nearby trajectories at a rate of order γ, and the long-time dynamics is governed by the reduced equation*

$$\dot{\boldsymbol{s}} = \boldsymbol{L}_s\boldsymbol{s} + \boldsymbol{N}_s\big(\boldsymbol{s}, \boldsymbol{h}(\boldsymbol{s})\big).$$

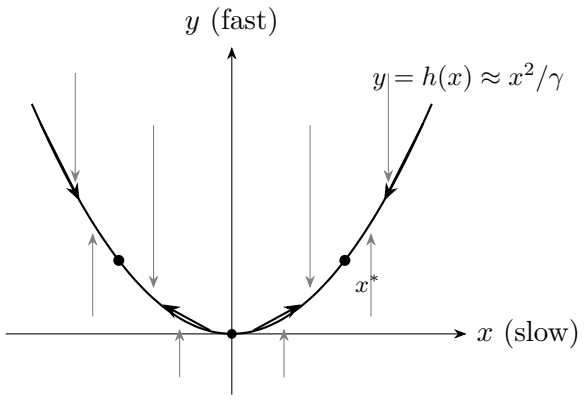


Figure 13.1: Adiabatic elimination in the two-mode system (13.1) for small $\lambda > 0$. Gray arrows: fast relaxation onto the slow manifold $y = h(x)$. Black arrows: slow motion along it, away from the unstable state at the origin and toward the stable states $x^* = \pm\sqrt{\lambda\gamma}$.

In mathematical form this is the center manifold theorem and its extensions to slow manifolds (Guckenheimer and Holmes 1983; Kuznetsov 2004); Haken's slaving principle is its physically oriented formulation, developed with the aim of applying it across physics, chemistry, and biology (Haken 1983b; Haken 1983a). The result is exact at the bifurcation point and holds as an approximation, with errors controlled by the ratio of timescales, when the control parameter is near but not at its critical value.[1]

Haken described the relation between order parameters and slaved modes as *circular causality*: the order parameters are generated by the joint dynamics of the parts, and they in turn determine the behavior of the parts (Haken 1983b). In the example, x grows because of its own linear instability, and y follows x through $y \approx x^2/\gamma$; y then feeds back on x through the term $-xy$, which saturates its growth. Both directions of the circle are visible in the equations. Neither requires anything beyond the equations.

13.4 The Laser

The laser was the case that led Haken to the general theory, and it remains the clearest. A population of pumped atoms sits in an optical cavity. Below threshold, each atom radiates spontaneously and independently, and the field in the cavity is a superposition of uncorrelated contributions. The relevant variables are the amplitudes of the cavity modes, the polarizations of the atoms, and the atomic inversion, and in many lasers the atomic variables relax much faster than the field in the cavity. Eliminating them adiabatically, as in

[1] The hypotheses matter. The statement is local (valid near the instability), requires a gap in the spectrum, and, in the presence of noise, holds only when the noise does not itself mix the timescales.

Section 13.2, leaves an equation for the amplitude E of the single mode that becomes unstable first:

$$\dot{E} = (G - \kappa)\, E - \beta\, |E|^2 E + F(t), \tag{13.6}$$

where G is the gain supplied by the pumped atoms, κ the loss of the cavity, $\beta > 0$ a saturation coefficient, and F a fluctuating force that represents spontaneous emission (Haken 1983b). The threshold $G = \kappa$ is a bifurcation of the same type as (13.3). Below it, the fluctuations of E are small and centered on zero; above it, $|E|$ settles near $\sqrt{(G - \kappa)/\beta}$ while its phase diffuses, and the light becomes coherent. The atomic polarizations are slaved to the field: each atom's dipole oscillates in step with the mode because the mode drives it.

Haken used this reduction to argue that the laser threshold is a nonequilibrium phase transition, with the field amplitude as order parameter. He also showed that the equations of the single-mode laser with its atomic variables retained are equivalent to the Lorenz equations of fluid convection, so that the same order-parameter description, and the same routes to instability beyond threshold, apply to lasers and to fluids (Haken 1975). That equivalence is the style of result synergetics aims at: a shared mathematical form across substrates, established by derivation from each substrate's own equations.

13.5 Order Parameters of Behavior

The method crossed into the study of movement through the coordination of two limbs. When people move their index fingers rhythmically in anti-phase and gradually increase the frequency, the movement switches spontaneously to in-phase at a critical frequency; the reverse switch does not occur (J. A. Kelso 1984). Haken et al. (1985) modeled the relative phase ϕ of the two fingers as the order parameter of the coordination, with dynamics

$$\dot{\phi} = -a \sin\phi - 2b \sin 2\phi = -\frac{\mathrm{d}V}{\mathrm{d}\phi}, \qquad V(\phi) = -a\cos\phi - b\cos 2\phi, \tag{13.7}$$

where the ratio b/a decreases as movement frequency increases. Anti-phase coordination ($\phi = \pi$) is stable while $b/a > 1/4$ and loses stability when the ratio falls below $1/4$, which reproduces the one-way switch. Chapter 10 develops the model and its evidence.

The behavioral case added something the laser did not need. In the laser the fast variables and their relaxation rates are known from atomic physics. In the finger experiment nobody measures the "fast modes"; their existence is inferred from the success of the order-parameter description. The model therefore made predictions that tested the slaving picture directly: if relative phase is the order parameter of an instability, its fluctuations should grow and its recovery from perturbations should slow down as the critical frequency is approached. Schöner and J. A. S. Kelso (1988) set out these predictions of

critical fluctuations and critical slowing down, and the experiments of Kelso's group confirmed them (J. A. S. Kelso 1995). That is the right standard for any claim that a biological variable is an order parameter.

13.6 What the Vocabulary Adds

Synergetics supplies three things of lasting value: a procedure (find the instability, separate the timescales, reduce to the order parameters), a family of normal forms that recur across substrates, and a set of testable signatures (critical slowing, critical fluctuations, hysteresis where the bifurcation is subcritical). Its language is more ambitious than its mathematics, and the gap should be kept in view.

"Enslavement" describes a function $\boldsymbol{f} = \boldsymbol{h}(\boldsymbol{s})$ on an invariant manifold. It says that, once transients have died out, the fast variables are determined by the slow ones. It does not say that the order parameter exerts a force on the parts that the parts' own equations do not contain. Every trajectory of the reduced system is a trajectory of the full system; the full microscopic dynamics determine everything, including the slow manifold and the order parameter's equation. What the reduction supplies is *autonomy of description*: the order parameter's future can be computed from its present without tracking the fast modes. Whether that autonomy amounts to a distinct macroscopic cause is a philosophical question with a literature of its own, taken up in Chapter 14.

The vocabulary misleads in three situations. When there is no spectral gap, as in developed turbulence, no small set of order parameters exists and the procedure does not apply. When noise is strong enough to mix timescales, the slaved variables no longer follow the slow ones, and the reduced description fails in ways the deterministic analysis does not show. And when an "order parameter" is chosen by fitting a low-dimensional model to data, without evidence of an instability or of timescale separation, the fit does not establish slaving. A good low-dimensional fit is compatible with many underlying structures; the signatures of Section 13.5 are what distinguish an order parameter from a convenient summary.

The chapter also sharpens a theme of Part III. In the laser the order parameter is physically explicit as a field that acts on each atom, just as the local mean field of Chapter 12 acts on each oscillator. The collective relation is delivered to every part through a physical medium, and each part's dynamics depend on it. That is the minimal sense in which a relation is available to a part. The parts do not store, categorize, or act on the order parameter in order to change it. Chapter 15 distinguishes this sensitivity from the richer capacities that the rest of the book requires of a reader.

13.7 Audit

Canonical phenomenon

The laser threshold: below it, independent emission and incoherent light; above it, one cavity mode grows and the atomic dipoles oscillate in step with it.

Minimal model

The two-mode system (13.1) and its reduction (13.3); the order-parameter equation of the single-mode laser (13.6); the relative-phase dynamics of coordination (13.7).

What is relational here?

The order parameter, a projection of the whole configuration onto an unstable pattern; the slaving function $\boldsymbol{f} = \boldsymbol{h}(\boldsymbol{s})$, a relation between fast and slow variables; in coordination, the relative phase of two limbs.

What physically reads it?

In the laser, the field: it is produced by the atoms and drives each of them. In the other examples the order parameter enters the dynamics of the parts through their coupling. Nothing reads it in a richer sense.

What existing field explains it?

Bifurcation theory and the center manifold theorem (Guckenheimer and Holmes 1983; Kuznetsov 2004); laser physics and synergetics (Haken 1983b; Haken 1983a; Haken 1975); coordination dynamics (Haken et al. 1985; J. A. Kelso 1984; Schöner and J. A. S. Kelso 1988; J. A. S. Kelso 1995). They explain the whole chapter.

What remains unexplained?

None in the physics. The causal interpretation of order parameters is contested, and Chapter 14 examines it.

A discriminating experiment

For a proposed biological order parameter, test the two signatures of an instability with timescale separation: as the control parameter approaches its critical value, the proposed variable's relaxation after a perturbation should slow and its fluctuations should grow, while perturbations of other measured variables should keep relaxing quickly. A variable that shows neither signature is a summary of the data, however good the fit.

13.8 Notes and References

Haken (1983b) is the introductory text of synergetics, with the laser, fluid convection, chemical reactions, and biological examples, and Haken (1983a) treats instability hierarchies in more depth. The equivalence between the

single-mode laser and the Lorenz equations is in Haken (1975). For the mathematics of reduction near bifurcations, including center manifolds and normal forms, see Guckenheimer and Holmes (1983) and Kuznetsov (2004); Strogatz (2015) gives an accessible introduction to the bifurcations themselves. The coordination experiments are J. A. Kelso (1984); the Haken–Kelso–Bunz model is Haken et al. (1985); the theory of critical fluctuations and slowing in behavioral transitions is Schöner and J. A. S. Kelso (1988); J. A. S. Kelso (1995) synthesizes the program. On the status of macroscopic variables and the vocabulary of emergence, see Anderson (1972) and the discussion in Chapter 14.

13.9 Exercises

Exercise 13.1. Derive (13.5) from (13.4). Write the reduced equation for x to order x^5. Then simulate (13.1) with $\gamma = 10$ and $\lambda = 0.1$ from several initial conditions and compare the full trajectory of $x(t)$ with the solutions of (13.3) and of your higher-order reduction.

Exercise 13.2. Repeat the simulation of Exercise 13.1 with $\gamma = 0.2$ and $\lambda = 0.1$. Describe how the reduced equation fails, and state, in terms of the eigenvalues of the linearization, the condition under which adiabatic elimination is justified.

Exercise 13.3. For the stochastic order-parameter equation $\dot{x} = \lambda x - x^3/\gamma + \sqrt{2Q}\,\eta(t)$ with Gaussian white noise η, show that the stationary density is $p(x) \propto \exp[-V(x)/Q]$ with $V(x) = -\lambda x^2/2 + x^4/(4\gamma)$. Sketch $p(x)$ for $\lambda < 0$, $\lambda = 0$, and $\lambda > 0$, and compute the variance of x at $\lambda = 0$ as a function of Q and γ. Relate your result to the change in the statistics of laser light at threshold.

Exercise 13.4. For (13.7), compute $V''(\pi)$ and show that anti-phase coordination is stable exactly when $b/a > 1/4$. Compute the relaxation time of small perturbations around $\phi = \pi$ as a function of b/a and show that it diverges as $b/a \to 1/4$ from above.

Exercise 13.5. Show that the Kuramoto order parameter r is a relational variable in the sense of Chapter 1. Compute r for three configurations of N phases: all equal, equally spaced around the circle, and independently uniform (give the expected value of r^2 in the last case). Explain in which of these configurations each oscillator can be said to be driven by the order parameter.

Exercise 13.6. Choose a system from your own field for which an order-parameter description has been proposed. Identify the control parameter, the proposed order parameter, and the variables assumed to be slaved. State what measurement would test for a spectral gap, and what result would show that the order parameter is a summary and not the slow variable of an instability.

14

Causal Emergence and Coarse-Graining

A temperature, a relative phase, and an order parameter are coarse-grained variables: many microscopic states map to each of their values. Coarse-graining is usually defended as a convenience that trades detail for tractability. A research program begun by Hoel, Albantakis, and Tononi claims more. On a measure they call effective information, a macroscopic description of a system can carry more causal information than the microscopic description it coarse-grains, and they call this *causal emergence* (E. P. Hoel et al. 2013).

The thesis of this chapter is that the claim is correct as a statement about a measure on transition structures and contested as a claim about where causation resides. We compute the measure in a small example, show exactly where the macroscopic advantage comes from in that example, and set the claim beside its critics and beside two alternatives: an information-theoretic criterion of emergence based on synergy, and the causal states of computational mechanics. For this book the lesson is narrower than either side of the dispute. Whether a coarse-grained relational variable does causal work has to be shown by interventions at the coarse-grained level, and a coarse-graining chosen by an analyst is a different thing from a coarse-graining that a physical system performs on its own inputs.

14.1 Micro, Macro, and Coarse-Graining Maps

Definition 14.1 (Coarse-graining)**.** Let S be a set of microscopic states. A *coarse-graining* is a surjective map $\mu : S \to M$ onto a set of macroscopic states; each macrostate $m \in M$ corresponds to the block $\mu^{-1}(m)$ of microstates that map to it. For a Markov chain on S with transition probabilities $T(s' \mid s)$, the coarse-graining is *lumpable* if the probability of moving from s into a block depends only on the block containing s:

$$\sum_{s' \in \mu^{-1}(m')} T(s' \mid s) = T_M(m' \mid \mu(s)) \quad \text{for all } s, m'.$$

In that case the macrostates form a Markov chain with transition matrix T_M.

A coarse-graining is a relation on microstates: it declares which of them count as the same. The temperature of a gas groups all molecular configurations with the same mean kinetic energy; the relative phase of two oscillators groups all pairs of phases with the same difference; the macrostate "the sign of a frequency difference is positive" groups every waveform whose slower beat component lies above the faster one. Some groupings respect the dynamics, in the sense of lumpability, and support a macroscopic dynamics of their own. Others mix microstates with different futures and do not.

14.2 Effective Information

Effective information was introduced by Tononi and Sporns (2003) to quantify the causal interactions within a system, and E. P. Hoel et al. (2013) adapted it to compare levels of description. The idea is to measure how much the state of a system at one time constrains its state at the next, when the system is set into each of its possible states by intervention with equal probability.

Definition 14.2 (Effective information)**.** Let a system with n states have transition probabilities $T_i = T(\cdot \mid i)$, the i-th row of its transition matrix. Its *effective information* is the mutual information between the current state and the next when the current state is set by an intervention distributed uniformly over all n states:

$$\begin{aligned} \mathrm{EI}(T) = I(X_t; X_{t+1})\big|_{X_t \sim \mathrm{uniform}} = H(\bar{T}) - \frac{1}{n}\sum_{i=1}^{n} H(T_i), \\ \bar{T} = \frac{1}{n}\sum_{i=1}^{n} T_i. \end{aligned} \tag{14.1}$$

The two terms have natural readings. Writing $\log_2 n - \frac{1}{n}\sum_i H(T_i)$ for the *determinism* of the system (how precisely each intervention fixes the next state) and $\log_2 n - H(\bar{T})$ for its *degeneracy* (how often different interventions lead to the same next states), effective information is determinism minus degeneracy (E. P. Hoel et al. 2013). A system scores highly when each state leads to a single next state and different states lead to different next states.

14.3 A Worked Example

Take eight microstates. From each of the states $1, \dots, 7$ the system jumps to one of the states $1, \dots, 7$ at random, each with probability $1/7$; from state 8 it stays in state 8. With J_7 the 7×7 matrix of ones, the transition matrix is

$$T_{\mathrm{micro}} = \begin{pmatrix} \frac{1}{7}J_7 & 0 \\ 0 & 1 \end{pmatrix}. \tag{14.2}$$

The average row $\bar{T}$ puts probability $1/8$ on each of the eight states, so $H(\bar{T}) = 3$ bits; seven rows have entropy $\log_2 7$ and the eighth has entropy zero. Hence

$$\mathrm{EI}(T_{\mathrm{micro}}) = 3 - \tfrac{7}{8}\log_2 7 \approx 0.544 \text{ bits}. \tag{14.3}$$

Now coarse-grain into two macrostates, $A = \{1, \ldots, 7\}$ and $B = \{8\}$. The coarse-graining is lumpable, and the macrostate dynamics is the identity: A always goes to A, B always goes to B. Its effective information is

$$\mathrm{EI}(T_{\mathrm{macro}}) = 1 \text{ bit}. \tag{14.4}$$

The macroscopic description has almost twice the effective information of the microscopic one. This is causal emergence in the sense of E. P. Hoel et al. (2013): the coarse-grained model is more deterministic and less degenerate than the fine-grained model it summarizes.

Where does the extra information come from? The micro-level measurement intervenes uniformly on eight states, so the macrostate A is set with probability $7/8$ and B with probability $1/8$. The macro-level measurement intervenes uniformly on two macrostates, so A and B are each set with probability $1/2$. The two numbers compare different interventions. Hold the intervention fixed at the micro-level uniform distribution and compute the macro-level mutual information instead:

$$I(M_t; M_{t+1})\big|_{\mathbb{P}(A)=7/8,\ \mathbb{P}(B)=1/8} = H\big(\tfrac{7}{8}, \tfrac{1}{8}\big) = 3 - \tfrac{7}{8}\log_2 7 \approx 0.544 \text{ bits}, \tag{14.5}$$

exactly the micro-level value. In this example the whole of the macroscopic advantage comes from the change in the intervention distribution. The following theorem says when that happens.

Theorem 14.3 (Where the advantage comes from)**.** *Let μ be a coarse-graining of a Markov chain, with $M_t = \mu(X_t)$, and suppose that (i) X_{t+1} depends on X_t only through M_t, and (ii) given M_{t+1}, the microstate X_{t+1} is independent of M_t. Then for every distribution of interventions on X_t,*

$$I(X_t; X_{t+1}) = I(M_t; M_{t+1}),$$

where on the right M_t carries the distribution induced by the intervention on X_t.

Proof. By (i), $X_t \to M_t \to X_{t+1}$ is a Markov chain and M_t is a function of X_t, so $I(X_t; X_{t+1}) = I(M_t; X_{t+1})$. Since M_{t+1} is a function of X_{t+1}, the chain rule gives $I(M_t; X_{t+1}) = I(M_t; M_{t+1}) + I(M_t; X_{t+1} \mid M_{t+1})$, and the last term vanishes by (ii). □

The example satisfies both conditions: which of the states $1, \ldots, 7$ the system occupies makes no difference to its future, and where it lands inside a block is uniform whatever block it came from. For coarse-grainings of this

kind, the effective information of the macroscopic model exceeds that of the microscopic one only because the macro-level measure intervenes differently. The theorem does not cover every case in the literature on causal emergence, and it does not show that the choice of a uniform macro-level intervention is wrong. It shows exactly what is being compared.

14.4 Causal Emergence and Its Critics

E. P. Hoel et al. (2013) present causal emergence as evidence that macroscopic descriptions can be causally privileged: a scientist who wants the most informative causal model of a system may do better at a coarser level. E. Hoel (2017) develops the claim with an analogy from communication theory. A noisy channel transmits information best when the sender uses a well-chosen subset of its possible inputs, as an error-correcting code does; a coarse-graining, on this view, is a code that uses the causal structure of a system more efficiently than the full microscopic repertoire does.

The critics accept the arithmetic and dispute the interpretation. Dewhurst and Philosophy Documentation Center (2021) argues that effective information measures properties of a model's transition structure under a stipulated intervention distribution, so that causal emergence in this sense is neither a discovery about causation in the world nor emergence in the sense philosophers have debated. Eberhardt and L. L. Lee (2022) argue that the macroscopic advantage arises from the way the coarse-graining and the uniform macro-level intervention distort the relation between the model and the system it represents, so that comparing effective information across levels compares different interventions rather than the same causal structure described at two grains. Theorem 14.3 makes that point exact for one class of cases.

The debate touches an older argument. Kim (1998) argued that if every physical effect has a sufficient physical cause, then a macroscopic property that supervenes on microscopic properties is either causally idle or an unexplained second cause of the same effect. Interventionist accounts of causation answer that causal relevance is a matter of invariant dependence under interventions on a variable at its own level of description (Woodward 2003): a macroscopic variable is a cause of an outcome if manipulating it, however realized, changes the outcome in a stable way. The interventionist answer puts the weight on a practical condition. A macro-level intervention is well defined only if the different microscopic ways of realizing it have the same effect. When they do not, the macroscopic variable is an ambiguous handle, and no measure computed on a transition matrix repairs the ambiguity.

14.5 Emergence as Synergy

Rosas et al. (2020) proposed a different criterion, based on information that a macroscopic feature carries about the future and that no part carries alone. Let $X_t = (X_t^1, \dots, X_t^n)$ be the parts of a system and $V_t = f(X_t)$ a feature that supervenes on them. Their practical criterion is

$$\Psi_{t,t'}(V) = I(V_t; V_{t'}) - \sum_{j=1}^{n} I(X_t^j; V_{t'}), \tag{14.6}$$

and $\Psi > 0$ is a sufficient condition for V to be causally emergent in their sense: the feature predicts its own future better than all the parts taken separately predict it. The criterion rests on the partial information decomposition (Appendix H), and it detects information that is synergistic across parts (Williams and Beer 2010; Mediano et al. 2021).

Example 14.4 (A parity that persists). Let a system consist of two bits. At each step the parity $V_t = X_t^1 \oplus X_t^2$ is preserved, and the bits themselves are redrawn uniformly among the two configurations with that parity. Start with the parity equally likely to be 0 or 1. Then $I(V_t; V_{t+1}) = 1$ bit, since the parity never changes. Each bit alone is uniform and independent of the parity, so $I(X_t^j; V_{t+1}) = 0$ for $j = 1, 2$. Hence $\Psi = 1$ bit, the maximum possible for a binary feature. The persistent variable is a relation between the bits, invisible to each bit, and it is exactly the kind of irreducibly joint relation met in the exclusive-or example of Chapter 1.

The synergy criterion avoids the choice of an intervention distribution, since it is computed from observed statistics. It inherits the open problems of partial information decomposition, where several incompatible measures of synergy are in use, and it measures predictive information in observational data rather than the effects of interventions. It is a good detector of relational variables that carry information about the future beyond their relata, and it is silent about whether anything in the system makes use of them.

14.6 Causal States

Computational mechanics gives the most principled answer to the question of how to coarse-grain (Crutchfield and K. Young 1989; Shalizi and Crutchfield 2001). Group two histories of a process together whenever they predict the same future.

Definition 14.5 (Causal state)**.** For a stationary process, two pasts $\overleftarrow{x}$ and $\overleftarrow{x}'$ are *predictively equivalent* if they induce the same conditional distribution over futures: $\mathbb{P}(\overrightarrow{X} \mid \overleftarrow{x}) = \mathbb{P}(\overrightarrow{X} \mid \overleftarrow{x}')$. The equivalence classes are the process's *causal states*; the machine that moves between them as symbols are emitted is

its ε-machine, and the entropy of the causal-state distribution is its *statistical complexity* C_μ.

Shalizi and Crutchfield (2001) show that the causal states are the minimal sufficient statistic for prediction: any coarse-graining of the past that predicts as well is a refinement of them, and any coarser one predicts worse. The causal-state partition is fixed by the process itself and by nothing an analyst chooses.

Example 14.6 (The golden mean process). Consider binary sequences with no two consecutive 0s, generated as follows: after a 1, the next symbol is 0 or 1 with equal probability; after a 0, the next symbol is 1. Every past ending in 1 predicts the same future, and every past ending in 0 predicts the same future, so there are two causal states. The stationary probabilities are $2/3$ for the first and $1/3$ for the second, so $C_\mu = H(2/3, 1/3) \approx 0.918$ bits, and the entropy rate is $2/3$ of a bit per symbol. An infinite variety of pasts collapses without loss into a relation between the last symbol and what can follow it.

14.7 Whose Coarse-Graining?

Effective information, the synergy criterion, and causal states share a feature that matters for the rest of this book. Each is computed by an analyst. The analyst chooses the coarse-graining (or, for causal states, derives it from the process), chooses the intervention distribution or the observational data, and evaluates the result. The macrostate is a variable of the analyst's model of the system.

A physical system can also perform a coarse-graining on its own inputs. Midbrain neurons of the electric fish combine amplitude and phase information from the body surface and respond selectively to the sign of the frequency difference between the fish's discharge and a neighbor's (Heiligenberg and G. Rose 1985; G. Rose and Heiligenberg 1986b). Very many different waveforms at the skin share a sign; the sign is what these neurons report, and it is what the fish's frequency shift follows (Heiligenberg 1991a). Their activity is a physically implemented coarse-graining, and the fish's pacemaker responds to the coarse-grained variable (Chapter 21). Stimulating the subdivision of the diencephalic nucleus electrosensorius associated with frequency increases produces the slow rises of discharge frequency characteristic of the jamming avoidance response (C. H. Keller et al. 1990). That is an intervention on a coarse-grained variable, at its own level, in a system that computes the variable itself.

None of the measures in this chapter distinguishes the two situations. An analyst's coarse-graining and a system's coarse-graining can have the same effective information and the same synergy. What distinguishes them is physical: in the second case the coarse-grained variable is the state of a part of

the system, and other parts respond to it. Chapter 15 makes that distinction the subject of Part IV.

14.8 Audit

Canonical phenomenon

An eight-state Markov chain whose two-block coarse-graining has almost twice the effective information of the chain itself; a two-bit system whose parity persists while its bits are redrawn.

Minimal model

Effective information (14.1); Theorem 14.3 on intervention distributions; the synergy criterion Ψ (14.6); causal states (Definition 14.5).

What is relational here?

Macrostates are equivalence relations on microstates; the parity is a relation between bits; causal states are relations between pasts and futures.

What physically reads it?

In every computation of the chapter, the analyst. In the electric fish, sign-selective neurons perform a coarse-graining that the rest of the fish uses.

What existing field explains it?

Information theory (Cover and Thomas 2006); causal emergence and its critics (E. P. Hoel et al. 2013; E. Hoel 2017; Dewhurst and Philosophy Documentation Center 2021; Eberhardt and L. L. Lee 2022); information decomposition (Williams and Beer 2010; Rosas et al. 2020; Mediano et al. 2021); computational mechanics (Crutchfield and K. Young 1989; Shalizi and Crutchfield 2001); interventionist causation and the exclusion argument (Woodward 2003; Kim 1998).

What remains unexplained?

The measures do not distinguish an analyst's coarse-graining from one that a system performs and uses. That distinction is the subject of Part IV, and it is where this chapter's residue lies.

A discriminating experiment

In the electric fish, present stimuli that differ in their microscopic details (amplitudes, contamination geometry, waveform) but share the sign of the frequency difference, and stimuli that match in detail but differ in sign. If the sign-selective population and the pacemaker respond alike within each sign class and differently across classes, and if direct stimulation of the coding population produces the behavior regardless of the stimulus, the coarse-grained variable is implemented and used by the system. An

analyst's macrovariable passes the first part of this test and cannot pass the second.

14.9 Notes and References

Effective information comes from Tononi and Sporns (2003); its use to compare levels of description, and the term causal emergence, from E. P. Hoel et al. (2013), with the communication-theoretic development in E. Hoel (2017). Dewhurst and Philosophy Documentation Center (2021) and Eberhardt and L. L. Lee (2022) are the critiques we follow. The synergy criterion is Rosas et al. (2020); its foundations are the partial information decomposition of Williams and Beer (2010) and the integrated information decomposition of Mediano et al. (2021). Computational mechanics begins with Crutchfield and K. Young (1989); Shalizi and Crutchfield (2001) give the optimality and minimality results for causal states. The exclusion argument is set out in Kim (1998), and the interventionist framework in Woodward (2003); Pearl (2009) develops the formal theory of intervention. Anderson (1972) remains the physicist's classic statement that new levels bring new variables without violating reduction.

14.10 Exercises

Exercise 14.1. Show that (14.1) equals determinism minus degeneracy as defined in Section 14.2. Compute the effective information of a permutation matrix on n states and of the matrix whose rows are all uniform, and interpret the two extremes.

Exercise 14.2. For the chain (14.2), consider the coarse-graining into $\{1,2,3,4\}$ and $\{5,6,7,8\}$. Show that it is not lumpable, and explain in terms of the futures of the grouped microstates why no macroscopic Markov chain can represent it.

Exercise 14.3. Construct a lumpable coarse-graining that violates condition (ii) of Theorem 14.3. Compute $I(X_t;X_{t+1})$ and $I(M_t;M_{t+1})$ under the same intervention distribution and show that they differ. Which of the two is larger, and must it always be?

Exercise 14.4. In Example 14.4, change the dynamics so that the first bit is copied, $X^1_{t+1}=X^1_t$, while the parity is still preserved. Compute Ψ for the parity. Then compute Ψ for the first bit regarded as a macroscopic feature, and explain the difference.

Exercise 14.5. Verify the stationary distribution, statistical complexity, and entropy rate of the golden mean process in Example 14.6. Then find the causal states of the process in which 1s occur in blocks of exactly two, separated by single 0s.

Exercise 14.6. Describe a macroscopic intervention in a system of your choice that is ambiguous in the sense of Section 14.4: two microscopic realizations of the same macro-level setting with different effects. Explain what this implies for effective information computed at the macroscopic level.

Exercise 14.7. Using Section 14.7, write a protocol for the electric fish that distinguishes a coarse-grained variable computed by the fish from one computed only by the experimenter. List the stimulus sets, the recordings, and the stimulation experiment, and state the outcome that would count against the claim that the fish uses the sign of the frequency difference as a variable.

Relative phase, the Kuramoto order parameter, and the order parameters of synergetics are relational variables that close on themselves: their dynamics can be written without the details of the parts. Physics and dynamical systems theory already own this territory, and they own it well. Nothing in Parts I to III needed anyone to read anything. The envelope, the locked phase, and the order parameter are there whether or not any system uses them. Part IV asks what changes when a physical system does use one.

Part IV

The Reader Problem

Who, or what, reads the relation?

15

What Is a Reader?

Parts I to III described relations that exist in configurations: fringes in overlapping gratings, envelopes in summed tones, locked phase differences, order parameters, frustrated compromises. In each case the relation was available to an outside observer, or it entered the equations of motion of the parts, and nothing in the system did anything with it. This part of the book concerns physical processes that respond to a relation, make it available to further processes, and sometimes act on it. We call such a process a *reader*.

The word needs discipline, because it slides between very different things: a photodiode, a thermostat, a neuron, an animal, a person reading a page. The thesis of this chapter is that reading is a family of distinct capacities, ordered roughly as interaction, sensitivity, detection, discrimination, and interpretation, and that each capacity can be present without the next. The *reading loop* $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$ is the formal object that keeps them apart. It names the mixture that delivers a relation, the operation that makes the relation explicit, the use made of it, and the dynamics by which that use changes the next mixture. The loop is standard control theory with the relational structure made visible, and the chapter says so; what it adds is a way to state precisely which capacities a given system has, and which it lacks.

15.1 Five Senses of "Reader"

The preface distinguished four senses of "reader." The distinction needs one more step here, because the gap between detecting a variable and using it to act is where the fish of Part V lives.

1. A *transducer* converts one physical quantity into another in a systematic way. A thermocouple converts a temperature difference between two junctions into a voltage; a photodiode converts light into current.
2. A *classifier* maps its inputs into a small set of categories. A comparator sorts voltages into "above threshold" and "below"; a smoke detector sorts air into "alarm" and "no alarm."
3. A *controller* uses what it reads to act, and its action changes what it will read next. A thermostat switches a heater according to the room

temperature relative to a setpoint.

4. An *interpreter* treats its read states as standing for conditions of the world, under a standard of correctness that belongs to the system itself. A reading can then be mistaken.
5. A *subject* is a reader for which there is something it is like to read.

The senses are not a strict ladder in which each presupposes the one before. Every physical reader is a transducer, but a controller need not classify: a proportional controller acts on a continuous error signal without ever sorting it into categories. The fourth and fifth senses are the subjects of Parts VI and X, and nothing in this chapter decides whether any physical system satisfies them. The table below lists, for each sense, its home discipline, because the book's method requires naming the prior art before naming anything else.

Sense	What it adds	Home discipline
Transducer	Systematic conversion of one physical quantity into another	Physics, instrumentation
Classifier	A small set of output categories with separable statistics	Signal detection theory
Controller	Action that changes the next input	Control theory, cybernetics
Interpreter	Conditions of correctness that belong to the system	Theoretical biology, philosophy of representation
Subject	Something it is like to read	Consciousness science, philosophy of mind

Table 15.1: Five senses of "reader" and the fields that already study each.

15.2 From Interaction to Interpretation

A second way to order the same territory follows a single variable of a system and asks what it does with a single feature of its input. The five rungs below are defined for a system with an input y, a feature $u = u(y)$ of that input (possibly relational), and a variable z of the system.

Definition 15.1 (The rungs of reading)**.**

(i) *Interaction.* The system's dynamics depends on its input: some variable of the system evolves differently for different y.

(ii) *Sensitivity.* The variable z depends systematically on the feature u: interventions that change u, holding the other features of y fixed, change z in a reproducible way.

(iii) *Detection.* The variable z has states that separate the presence of u from its absence (or two ranges of u) reliably enough to support different downstream responses.

(iv) *Discrimination.* The variable z takes distinguishable states for two or more values of u, where u may be a relation among components of the input that no single component carries.

(v) *Interpretation.* Some process of the system uses the states of z as standing for conditions of the world, and a standard of correctness that belongs to the system makes some uses right and others wrong.

The rungs are defined relative to a feature. A system can be sensitive to one feature of its input and blind to another, and the relational features are often the ones it is blind to. The definitions of sensitivity and discrimination are counterfactual, in the interventionist sense of Woodward (2003): they say what would happen to z if u were changed and nothing else. Detection borrows its criterion from signal detection theory, in which a detector is characterized by how well the distributions of its decision variable under two conditions are separated, independently of where it places its criterion (Green and Swets 1966).

The claim that the rungs are distinct is established by examples, each of which reaches one rung and stops.

Example 15.2 (Interaction without sensitivity to the relation). Two tones at frequencies ω_1 and ω_2, both far above their difference $|\omega_1 - \omega_2|$, drive a linear band-pass filter tuned to the difference frequency. The filter interacts with the input, since its state is driven by it. Its output is negligible whatever the detuning, because a linear system adds no frequencies and the input has no energy near $|\omega_1 - \omega_2|$ (Chapters 3 and 4). The filter has no variable sensitive to the beat.

Example 15.3 (Sensitivity without detection). The height of the column in a liquid thermometer is a smooth and reproducible function of temperature. It is sensitive to temperature over its whole range. It has no decision stage and no categories: nothing in the thermometer separates "warm" from "cold," and nothing downstream of it responds differently to the two.

Example 15.4 (Detection without discrimination of the relation). Pass the two-tone signal of Example 15.2 through a rectifier and a low-pass filter, and compare the result with a threshold. The output fluctuates at the beat rate $|\omega_1 - \omega_2|$ whenever the tones are close in frequency, and the comparator flags the presence of a beat. The device detects beating. It cannot discriminate the sign of the detuning, whether the second tone lies above or below the first, because the envelope of the sum is the same in both cases (Chapter 16).

Example 15.5 (Discrimination without control). A lock-in amplifier multiplies its input by a reference oscillation, $\cos \omega_r t$ and $-\sin \omega_r t$, and low-pass filters both products (Dicke 1946; Scofield 1994). For an input $A\cos(\omega t + \varphi_0)$ with ω close to ω_r, the two outputs are, up to a constant factor, the in-phase and quadrature components of the phasor $A e^{\mathrm{i}[(\omega-\omega_r)t+\varphi_0]}$. The instrument discriminates the phase of the input relative to its reference; when the input

frequency differs slightly from the reference, the phasor rotates, and the sense of rotation discriminates the sign of $\omega - \omega_r$. In its usual use the lock-in acts on nothing. Its outputs are displayed, and the experimenter interprets them.

Example 15.6 (Control without interpretation). A thermostat reads the room temperature relative to a setpoint and switches a heater, and the heater changes the temperature it will read next. The loop is closed. Its standard of success, a room held near the setpoint, was fixed by whoever set the dial. The thermostat's reading can be inaccurate, when its sensor drifts, but whether the reading is right or wrong is settled by a standard supplied from outside the device, and nothing about the thermostat's own continuation depends on it.

The last rung is where the book's harder questions start. Interpretation requires that a system's readings be correct or incorrect by a standard that belongs to the system. Whether any physical system meets that requirement, and what it would take, is the subject of Part VI. The electric fish is the book's candidate: its reading of a frequency difference serves an electrolocation system whose success or failure matters to the fish, and it can be wrong (Chapter 23).

Remark 15.7. The feature at the top of each example is a relation. The filter and the envelope detector face a relation between two external tones. The lock-in, the thermostat, and, in later chapters, the phase-locked loop and the fish read a relation between an input and a reference. The lock-in and the thermostat *hold* their reference: an oscillator, a setpoint. The phase-locked loop and the fish *are* their reference: the relation they read is between the world and their own emitted signal. The second kind of reader is where agent-centered measurement begins (Chapters 22 and 34).

15.3 The Reading Loop

The five rungs describe a reader in isolation. Most readers that matter act on what they read, and their actions change what they read next. The reading loop puts the pieces in one structure.

Definition 15.8 (Reading loop)**.** Let S_{A} and S_{E} be the state spaces of an agent and its environment, Y a space of mixtures, Z a space of read variables, and $\mathcal{A}$ a space of actions. A *reading loop* is a tuple $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$ of maps

$$\mathsf{M} : S_{\mathrm{A}} \times S_{\mathrm{E}} \to Y, \qquad y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}}) \qquad \text{(mixing)} \tag{15.1}$$

$$\mathsf{D} : Y \to Z, \qquad z = \mathsf{D}(y) \qquad \text{(reading)} \tag{15.2}$$

$$\pi : Z \to \mathcal{A}, \qquad a = \pi(z) \qquad \text{(policy)} \tag{15.3}$$

$$\mathsf{F} : S_{\mathrm{A}} \times \mathcal{A} \to TS_{\mathrm{A}}, \qquad \dot{s}_{\mathrm{A}} = \mathsf{F}(s_{\mathrm{A}}, a) \qquad \text{(intervention)} \tag{15.4}$$

together with environmental dynamics $\dot{s}_{\mathrm{E}} = G(s_{\mathrm{E}}, s_{\mathrm{A}})$, which may be trivial. The closed-loop dynamics of the agent is

$$\dot{s}_{\mathrm{A}} = \mathsf{F}\big(s_{\mathrm{A}}, \pi(\mathsf{D}(\mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})))\big). \tag{15.5}$$

Two refinements are needed in practice. The reader D usually has internal state (a filter integrates, a neuron adapts), so it maps histories of the mixture to present values of z: formally a causal operator $z(t) = \mathsf{D}[y](t)$ that depends on $y(s)$ for $s \le t$. And each map may carry noise. Neither refinement changes the structure.

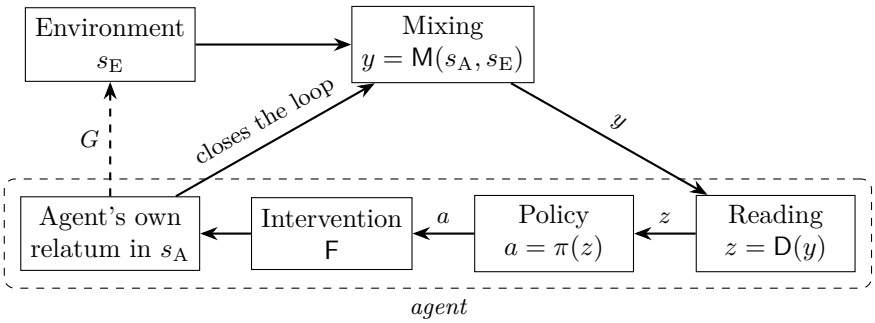


Figure 15.1: The reading loop $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$. The loop is participatory when the agent's own state enters the mixture as one relatum of the relation it reads and the agent's action changes that state. The dashed arrow G marks action on the environment.

Three kinds of loop matter, and the distinctions between them do real work in later parts.

Definition 15.9 (Open, closed, and participatory loops)**.** A reading loop is *open* if the action a affects neither the agent's contribution to the mixture nor the environment, so that what is read does not depend on what was done. It is *closed* if the action affects the next mixture, through F or through G. It is *participatory* if, in addition, the relation that D reads has one relatum in the agent's own state, and the action changes that relatum.

A thermostat is closed but not participatory: it changes the room, one relatum of the relation it reads, while the other relatum, the setpoint, stays where it was put. A phase-locked loop and the electric fish are participatory: each changes its own oscillator, one relatum of the relation it reads, in order to change the relation.

Closing a loop has a consequence that is easy to state and easy to forget. The statistics of what a closed-loop reader reads are partly made by its own policy.

Proposition 15.10 (Policy-dependent readings)**.** *Consider a participatory loop in discrete time in which the relation r_t is shifted by the action and by environmental noise, $r_{t+1} = r_t + a_t + \eta_t$, the reader returns a noisy value $z_t = r_t + \nu_t$, and the policy nulls the relation in proportion to the reading, $a_t = -k z_t$ with $0 < k < 2$. With η_t and ν_t independent, zero-mean, with*

variances σ_η^2 and σ_ν^2, the stationary variance of the read variable is

$$\mathrm{Var}(z) = \frac{\sigma_\eta^2 + k^2\sigma_\nu^2}{k(2-k)} + \sigma_\nu^2, \tag{15.6}$$

which depends on the gain k of the policy.

Proof. Substituting the reading and the policy, $r_{t+1} = (1-k)r_t - k\nu_t + \eta_t$. For $0 < k < 2$ the recursion is stable, and its stationary variance V satisfies $V = (1-k)^2V + k^2\sigma_\nu^2 + \sigma_\eta^2$, so $V = (\sigma_\eta^2 + k^2\sigma_\nu^2)/(k(2-k))$. Then $\mathrm{Var}(z) = V + \sigma_\nu^2$. □

The model is a toy, and the point is general. An observer who recorded only the read variable of a closed-loop system and treated its statistics as a property of the environment would get the environment wrong, by an amount that depends on a policy the observer did not measure. Active sensing systems are the extreme case: an animal that moves its sensors, or emits its own probing signal, shapes the statistics of its own input (M. E. Nelson and MacIver 2006). Chapter 59 turns this into an experimental method.

15.4 Four Loops

The same structure describes devices and animals that differ in almost every other respect. Setting them side by side shows what the structure captures and what it leaves out.

The thermostat. The mixture M is the thermal coupling between room and sensor. The reader D is a sensor and a comparator that return the sign of the difference between the sensed temperature and the setpoint, a relation between the room and a reference the device holds. The policy π switches the heater, and the intervention changes the room through the environmental dynamics G. The loop is closed and regulates the relation toward zero, in the sense of classical cybernetics (Wiener 1948; Ashby 1956).

The lock-in amplifier. The mixture is formed inside the instrument, by multiplying the input with the reference; the reader is the low-pass filter that turns the products into in-phase and quadrature outputs (Dicke 1946; Scofield 1994). There is no policy and no intervention. The loop is open, and all of the instrument's reading is delivered to an outside interpreter.

The phase-locked loop. A phase detector multiplies the input with the output of a voltage-controlled oscillator; a loop filter extracts the phase error; the error sets the control voltage; the oscillator's frequency changes; and the loop settles when the oscillator tracks the input's phase (Horowitz and Hill

2015). Here the reader's reference is one of the reader's own states, and the policy changes it so as to null the relation it reads. The loop is participatory.

The electric fish. The weakly electric fish *Eigenmannia* emits a quasi-sinusoidal field at a few hundred hertz and senses it through electroreceptors in its skin (Heiligenberg 1991a). When a neighbor discharges at a nearby frequency, the two fields mix at the skin, and the mixture is modulated in amplitude and phase at the frequency difference $\Delta f = f_e - f_s$ (Chapter 20). Two classes of receptor afferents encode amplitude and timing (Scheich et al. 1973); their signals pass through the hindbrain to the midbrain torus semicircularis, where neurons compare timing between body regions and combine it with amplitude to become selective for the sign of Δf (Heiligenberg and G. Rose 1985; G. Rose and Heiligenberg 1985a; G. Rose and Heiligenberg 1986b). The diencephalic nucleus electrosensorius and two prepacemaker pathways, one for increases and one for decreases, then drive the medullary pacemaker in a push-pull arrangement (C. H. Keller et al. 1990; Metzner 1993b). The fish raises its frequency if the neighbor's is lower and lowers it if the neighbor's is higher, so that $|\Delta f|$ grows (Metzner 1993b). The computation uses the electrosensory mixture alone, with no internal reference to the fish's own pacemaker command (Heiligenberg 1991a; Kawasaki 1993). The loop is participatory, like the phase-locked loop's, with the opposite aim: the phase-locked loop nulls the relation it reads, and the fish enlarges it. Chapter 21 derives how the sign can be read without a reference, and Chapter 22 places the fish among the other ways animals separate the consequences of their own actions from the world's.

Loop	**Relation read by** D	**Reference**	**Kind**
Thermostat	Room temperature relative to setpoint	Held (setpoint)	Closed
Lock-in amplifier	Input phase relative to reference	Held (oscillator)	Open
Phase-locked loop	Input phase relative to own oscillator	Is the reader's own state	Participatory, nulling
Electric fish	Sign of Δf between neighbor and own discharge	Is the fish's own discharge	Participatory, enlarging

Table 15.2: Four reading loops. All read relations; they differ in where the reference lives and in whether the loop closes through the reader's own relatum.

15.5 Reading Is Implemented

A reader, as defined here, is a physical process. The definitions of sensitivity and discrimination require that the state z *would have been different* had the relation been different, with the other features of the input held fixed.

That requirement rules out a cheap move. Given any physical system and any relation in its environment, an analyst can usually find some function of the system's states that happens to track the relation over a recorded episode, and declare the system a reader. The counterfactual requirement blocks this: the system's causal organization must produce the dependence under intervention, and the dependence must hold for relations the system did not happen to encounter. Chapter 17 develops the point in the setting where it was first sharpened, the argument that every physical system implements every computation, and the counterfactual and mechanistic replies to it (Chalmers 1996a; Piccinini 2015).

The requirement also clarifies what a reader delivers. It delivers a physical state with its own consequences: the voltage on a capacitor, the firing of a population of neurons, the position of a relay. Pattee argued that measurement in nature is a physical process of a special kind, one that maps a dynamical state onto a record that another process uses, and that describing it as measurement requires a cut between the system described by laws and the record described by its function (Pattee 2001). The reading loop places that cut at D: upstream of it is dynamics, downstream of it is a variable that the policy treats as information. Chapter 25 examines whether the cut is a convenience of description or a feature of the systems themselves.

15.6 What the Loop Leaves Open

The reading loop is control theory written with its relational structure made explicit. Wiener's cybernetics described animals and machines with the same feedback diagrams (Wiener 1948); Ashby formalized regulation and the requisite variety of regulators (Ashby 1956); Conant and Ashby argued that every good regulator of a system must be a model of that system (Conant and Ross Ashby 1970); the internal model principle of control theory states a precise version of that claim for linear systems (Francis and Wonham 1976). Nothing in Definition 15.8 goes beyond this tradition, and the book does not claim otherwise.

The loop's value lies in what it lets us say sharply. A phase-locked loop and an electric fish have the same loop structure. An adaptive radio that moves its carrier away from interference, as Bluetooth radios have done with adaptive frequency hopping since the 2003 version of the specification (*Specification of the Bluetooth System, Version 1.2* 2003), shares the fish's input-output function in outline, and cognitive radios extend the idea to radios that sense their spectral environment and reconfigure themselves (Mitola and Maguire 1999; Haykin 2005). Whatever the fish has that the radio lacks cannot be the loop's structure, and it cannot be the input-output function.

Three candidates remain, and each defines a later part of the book. The first is who built the reader: the radio's reader exists because engineers designed it,

while the fish grows, maintains, and reproduces its own (Part VI). The second is whether the reading has conditions of correctness that belong to the system: the radio's frequency hops are good or bad by its designers' standards, the fish's frequency shifts by the standard of its own electrolocation (Chapter 28). The third is whether there is anything it is like to be the reader (Part X). The first two are questions that theoretical biology already asks and partly answers. The third is the question the book began with, and the reading loop does not touch it.

15.7 Audit

Canonical phenomenon

A thermostat, a lock-in amplifier, a phase-locked loop, and the jamming avoidance response of the electric fish, side by side.

Minimal model

The rungs of reading (Definition 15.1); the reading loop $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$ (Definition 15.8); open, closed, and participatory loops; policy-dependent read statistics (Proposition 15.10).

What is relational here?

In every case the read variable is a relation between an input and a reference: temperature and setpoint, input and reference oscillator, input and the reader's own oscillator, neighbor's discharge and the fish's own.

What physically reads it?

A comparator; a mixer with low-pass filters; a phase detector with a loop filter; electroreceptors, hindbrain, and midbrain circuits of the fish.

What existing field explains it?

Control theory and cybernetics (Wiener 1948; Ashby 1956; Conant and Ross Ashby 1970; Francis and Wonham 1976); signal detection theory (Green and Swets 1966); instrumentation (Dicke 1946; Scofield 1994; Horowitz and Hill 2015); the neuroethology of the jamming avoidance response (Heiligenberg 1991a; Metzner 1993b); Pattee's account of measurement (Pattee 2001). The loop formalism belongs to these fields.

What remains unexplained?

The difference between loops whose reader and standard of success are supplied from outside and loops that build and maintain their own (Part VI), and the question whether any loop has a subject (Part X).

A discriminating experiment

To establish that a system reads a relation, vary only the relation and record the candidate read variable and the action. In the fish, replace the fish's own discharge and the neighbor's with electronic mimics (Heiligenberg's S1 and

S2), change only the sign of Δf with its magnitude and the amplitudes held fixed, and record sign-selective neurons and the pacemaker; then interrupt the candidate pathway and check that the action no longer follows the relation (Heiligenberg 1991a; Metzner 1993b). A phase-locked loop passes the same test. That is the point: the test establishes reading, and nothing more.

15.8 Notes and References

The vocabulary of transducers, detectors, and controllers comes from instrumentation and control engineering; Horowitz and Hill (2015) covers the electronic readers of this chapter, including phase-locked loops, and Scofield (1994) gives a frequency-domain account of the lock-in amplifier, whose modern form descends from Dicke (1946). Signal detection theory is set out in Green and Swets (1966). The cybernetic tradition begins with Wiener (1948) and Ashby (1956); the good-regulator theorem is Conant and Ross Ashby (1970), and the internal model principle Francis and Wonham (1976). Shannon's theory of communication (Shannon 1948) supplies the measures of how much a reader's state can say about its input, and F. I. Dretske (1981) built a theory of perceptual knowledge on it; both return in Part VI. Pattee's account of measurement as a physical process is Pattee (2001). On active sensing and the way an animal's own actions shape its inputs, see M. E. Nelson and MacIver (2006). The electric-fish material follows Heiligenberg (1991a) and the primary studies cited in the text; adaptive frequency hopping is specified in *Specification of the Bluetooth System, Version 1.2* (2003), and cognitive radio is described in Mitola and Maguire (1999) and Haykin (2005).

15.9 Exercises

Exercise 15.1. Place each of the following on the rungs of Definition 15.1 with respect to the feature named, and say which of the five senses of Section 15.1 it satisfies: a thermocouple (temperature difference between its junctions); a smoke alarm (smoke density); a camera's autofocus (image sharpness); a lock-in amplifier used in a feedback loop that stabilizes a laser's frequency (the laser's frequency relative to a reference cavity). For each, state one intervention that would establish its rung.

Exercise 15.2. For the input $A\cos(\omega t+\varphi_0)$ and a lock-in referenced at ω_r, compute the outputs $X=\mathrm{LPF}[2x\cos\omega_r t]$ and $Y=\mathrm{LPF}[-2x\sin\omega_r t]$, assuming an ideal low-pass filter that removes components near $2\omega_r$. Show that $X+\mathrm{i}Y=A\mathrm{e}^{\mathrm{i}[(\omega-\omega_r)t+\varphi_0]}$, and that the sense of rotation of (X,Y) gives the sign of $\omega-\omega_r$.

Exercise 15.3. Derive (15.6), plot $\mathrm{Var}(z)$ against k for $\sigma_\eta=\sigma_\nu=1$, and find the gain that minimizes it. An observer who ignores the policy estimates the

environmental noise as $\mathrm{Var}(z)$; compute the size of the error as a function of k. Repeat the analysis for the enlarging policy $a_t = +kz_t$ and describe what happens to the linear model.

Exercise 15.4. Write the phase-locked loop and the jamming avoidance response of the fish in the notation of Definition 15.8, naming M, D, π, and F for each. For each, identify the relatum that the agent changes. Model the fish's policy as $\dot{f}_{\mathrm{s}} = -k\tanh(\Delta f / w)$ with $k, w > 0$, a smooth stand-in for a response to the sign of Δf, and linearize both loops about $\Delta f = 0$; show that the fixed point is stable for the phase-locked loop and unstable for the fish.

Exercise 15.5. Design a bench experiment that separates detection of a beat from discrimination of its sign in an electronic reader of your choice. Specify the stimuli, the variable you record, and the result that would place the device on the detection rung but not on the discrimination rung.

Exercise 15.6. Find an example in biology, other than the electric fish, of a participatory reading loop: a system that emits or moves in order to sense, and whose sensed relation has one relatum in its own state. Describe it in the notation of Figure 15.1, and state what would count as evidence that the organism reads the relation rather than one of its relata.

Exercise 15.7. An adaptive radio and the electric fish share the input-output function "move away from an interfering frequency." List three measurable properties that differ between them, and for each, state which of the candidates in Section 15.6 it bears on.

16

Reader-Relative Variables

Two tones sound together. A linear filter tuned to their difference frequency receives nothing. A square-law detector receives a steady fluctuation at the beat rate. A lock-in amplifier referenced to one of the tones receives a slowly rotating phasor whose sense of rotation says which tone is higher. Play the tones to the two ears separately and there is no acoustic mixture at all, yet listeners hear a beat, because the mixture is formed inside the auditory brainstem. One physical situation presents different variables to different receivers.

The thesis of this chapter is that given a physical mixture $y = \mathsf{M}(A, B)$, a physical reader D_i makes some relational variables of y dynamically explicit and leaves others implicit, and different readers applied to the same mixture make different variables explicit. The claim is about physics and information. It requires no metaphysics, and it has a precise limit: no reader can make explicit any information about a relation that the mixture does not already carry. What a reader contributes is format, the conversion of information that is present but unusable by some downstream process into a physical state that process can respond to.

16.1 One Mixture, Five Readers

Let two sources emit tones of amplitudes a and b at nearby frequencies ω_1 and ω_2, both far above their detuning $\Delta\omega = \omega_1 - \omega_2$. Their superposition at a receiver is

$$\begin{aligned} y(t) &= a\cos(\omega_1 t + \psi_0) + b\cos\omega_2 t \\ &= \mathrm{Re}\Big[\mathrm{e}^{\mathrm{i}\omega_2 t}\big(b + a\,\mathrm{e}^{\mathrm{i}(\Delta\omega\, t+\psi_0)}\big)\Big]. \end{aligned} \tag{16.1}$$

The second form shows the relational content of the mixture: relative to the second tone, the first is a phasor of length a rotating at the detuning. The rate of rotation is $|\Delta\omega|$; its sense is the sign of $\Delta\omega$; its instantaneous angle is the beat phase. We follow what five readers do with y.

A linear band-pass filter. A linear system responds at the frequencies present in its input and at no others. The input has energy near ω_1 and ω_2

and none near $|\Delta\omega|$, so a filter tuned to the difference frequency returns a negligible output, whatever the detuning (Chapter 3). The relation is in y; it is not available to this reader.

A square-law detector. Squaring the mixture and removing everything near $2\omega_1$, $2\omega_2$, and $\omega_1 + \omega_2$ with a low-pass filter leaves

$$\mathrm{LPF}[y^2] = \tfrac{1}{2}(a^2 + b^2) + ab\cos(\Delta\omega\, t + \psi_0). \tag{16.2}$$

The detector's output oscillates at the beat rate with depth proportional to ab (Chapter 4). It makes $|\Delta\omega|$ and the beat phase explicit. Because the cosine is even, it does not make the sign of $\Delta\omega$ explicit.

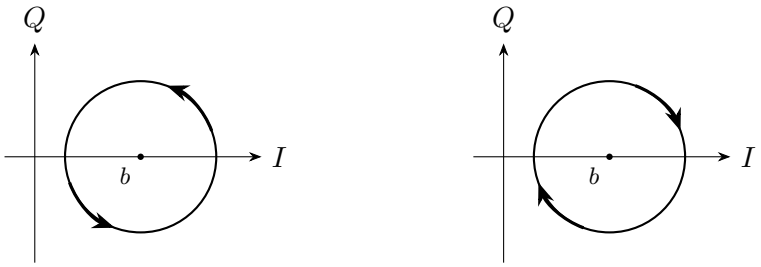
An envelope detector. A rectifier followed by a smoothing filter tracks the envelope $A(t) = |b + a\mathrm{e}^{\mathrm{i}(\Delta\omega t + \psi_0)}| = \left(a^2 + b^2 + 2ab\cos(\Delta\omega t + \psi_0)\right)^{1/2}$. Its output carries the same relational variables as the square-law detector's, and loses the same one.

A quadrature lock-in amplifier. The instrument multiplies y by $2\cos\omega_2 t$ and, separately, by $-2\sin\omega_2 t$, and low-pass filters both products. The two outputs, the in-phase component I and the quadrature component Q, combine into

$$u(t) = I(t) + \mathrm{i}Q(t) = b + a\,\mathrm{e}^{\mathrm{i}(\Delta\omega\, t + \psi_0)}. \tag{16.3}$$

The point (I, Q) circles the point $(b, 0)$ at rate $|\Delta\omega|$, counterclockwise when $\Delta\omega > 0$ and clockwise when $\Delta\omega < 0$ (Figure 16.1). The lock-in makes explicit the magnitude of the detuning, its sign, both amplitudes, and the phase of the first tone relative to the second. It can do so because it holds a reference, an internal oscillation at ω_2 against which rotation has a direction (Dicke 1946; Scofield 1994).

The binaural pathway. Deliver the first tone to the left ear and the second to the right. No acoustic mixture exists anywhere outside the head. Each auditory nerve phase-locks to its own tone, and binaural neurons in the superior olivary complex respond to the difference in timing between the two ears (Goldberg and P. B. Brown 1969). When the tones differ slightly in frequency, the interaural phase difference advances steadily at the detuning, neurons in the inferior colliculus fire in step with the resulting binaural beat (Kuwada et al. 1979), and listeners hear a beat that is present in neither ear, for carrier frequencies low enough for the effect to occur (Licklider et al. 1950; Oster 1973). Here the mixing operation M happens inside the reader, after each relatum has been transduced separately.



$\Delta\omega > 0$: counterclockwise $\Delta\omega < 0$: clockwise

Figure 16.1: Output of a quadrature lock-in referenced to the second tone, (16.3). The phasor $u(t) = b + ae^{i(\Delta\omega t+\psi_0)}$ circles the point b with radius a. Its rate gives $|\Delta\omega|$; its sense gives the sign of $\Delta\omega$. A square-law or envelope detector sees only the distance of u from the origin, which is the same for both senses of rotation.

Reader	**Explicit relational variables**	**Implicit (not available)**	**Reference held**
Linear filter at $\|\Delta\omega\|$	None	All	None
Square-law detector	$\|\Delta\omega\|$, beat phase, product ab	Sign of $\Delta\omega$	None
Envelope detector	$\|\Delta\omega\|$, beat phase, envelope depth	Sign of $\Delta\omega$	None
Quadrature lock-in	$\|\Delta\omega\|$, sign of $\Delta\omega$, a, b, relative phase	Absolute phase of the pair	Oscillator at ω_2
Binaural pathway	Interaural phase difference and its rate	(Sign: open question)	None; two channels

Table 16.1: One relational situation, five readers. Which variables become explicit depends jointly on the mixture and on the reader.

16.2 Explicit Variables

The examples call for a definition of what it means for a reader to make a relation explicit. The definition has to refer to what comes after the reader, because "explicit" means usable, and usable means usable by something.

Definition 16.1 (Dynamically explicit variable)**.** Let a reader produce $z = \mathsf{D}(y)$ from a mixture y, and let a downstream process with state w evolve as $\dot{w} = \Phi(w, z)$. A relational feature $r = r(y)$ is made *dynamically explicit* by D for that process if, over the conditions of interest, z is approximately a function of r, $z \approx g(r)$, and the process's state changes when r changes: interventions on r produce, through z and Φ, systematically different trajectories of w.

Explicitness is therefore a three-place relation among a feature, a reader, and a downstream process. The output of a square-law detector makes the beat rate explicit for a linear filter tuned to it and for an ear, and makes

nothing explicit for a process that responds only to steady voltages. The same waveform can be explicit for one process and implicit for another sitting beside it.

16.3 What a Reader Cannot Make Explicit

The failure of the square-law detector to register the sign of the detuning is an instance of a general limit.

Proposition 16.2 (Symmetry bound)**.** *Let T be a transformation of inputs that leaves a reader's output unchanged, $\mathsf{D}(Ty) = \mathsf{D}(y)$ for all y in some set, and let f be a feature with $f(Ty) \neq f(y)$ on that set. Then no process downstream of D can respond differently to y and to Ty, and D makes f explicit for no downstream process on that set.*

Proof. Every downstream process receives $z = \mathsf{D}(y)$, which is identical for y and Ty; its trajectories are therefore identical for the two inputs, whatever the value of f. □

For the square-law detector, take T to mirror the first tone's frequency about the second and reverse its phase: $\omega_1 \mapsto 2\omega_2 - \omega_1$ and $\psi_0 \mapsto -\psi_0$. Then $\Delta\omega \mapsto -\Delta\omega$, while (16.2) is unchanged because $\cos(-\Delta\omega t - \psi_0) = \cos(\Delta\omega t + \psi_0)$. The detector and everything downstream of it are blind to the sign.

The proposition says what a reader must have in order to read the sign: something that breaks the symmetry T. The lock-in breaks it with a reference oscillation that gives its two outputs an order, so that rotation has a direction. A reader that could measure instantaneous frequency would break it with a time base, since the instantaneous frequency of the mixture swings to one side of the mean when the envelope is large and to the other when it is small, in a direction set by the sign of the detuning (Exercise 16.4). And a reader with two channels that receive the two sources in different proportions can break it without any reference at all, by comparing the timing of the two channels with each other. That third solution is the one the electric fish uses: two patches of skin receive a neighbor's field with different strengths, and the joint behavior of the amplitude at one patch and the phase difference between patches rotates in a direction set by the sign of Δf (Chapter 21).

16.4 No Metaphysics Required

The phrase "reader-relative variable" invites a stronger reading than the one intended, so we state the intended one exactly. The relational features of a mixture are features of the mixture. An analyst who records $y(t)$ with enough bandwidth and precision can compute the detuning, its sign, and the beat

phase, whatever readers happen to be present, and no reader can extract more information about a relation than the mixture carries (Section 16.7). What depends on the reader is which of those features becomes a physical state that some further process can respond to. That dependence is ordinary physics: it follows from the reader's nonlinearity, its bandwidth, and the references it holds.

The claim should be kept apart from a much stronger one with a similar sound. Relational quantum mechanics holds that the values of physical quantities are defined only relative to other physical systems, as a feature of quantum theory itself (Rovelli 1996). Whatever the merits of that interpretation, it is a thesis about the foundations of physics. The reader-relativity of this chapter is classical and operational, and it holds whichever interpretation of quantum mechanics is correct.

16.5 Description and Dynamics

Searle distinguished features of the world that exist independently of observers from features that exist only relative to the attitudes or uses of observers, such as something's being a bathtub or a computer (Searle 1992). It is tempting to file reader-relative variables under the second heading. They belong under the first.

An analyst who displays the envelope of y on a screen has produced a description of the mixture. The voltage on the smoothing capacitor of an envelope detector is a physical state of the detector, with physical consequences: it can drive a relay, bias a transistor, or trigger an alarm. Which variable that voltage tracks depends on which detector is present, and that is the sense in which the variable is reader-relative. The voltage itself is as intrinsic as any other physical quantity. The distinction between an analyst's description and a detector's dynamics is exactly the distinction that Chapter 14 drew between an analyst's coarse-graining and a system's.

The distinction carries a burden that Chapter 17 takes up. For almost any physical system and almost any relation, an analyst can find some function of the system's states that tracks the relation over a recorded episode. If that were enough for reading, every rock would read every beat. The counterfactual condition in the next section is what rules the cheap move out.

16.6 Counterfactual Response

Definition 16.3 (Counterfactual encoding)**.** A reader's output z *encodes* a relational feature r over a range of conditions if, for interventions that change r within that range while holding fixed the other features to which the reader is sensitive, z changes as a fixed function of r.

The definition applies the interventionist account of causal relevance (Woodward 2003) to the output of a reader. It is demanding in the right way.

Example 16.4 (Three encoding tests). To show that an envelope detector encodes $|\Delta\omega|$, vary the detuning with the amplitudes fixed and observe the rate of the output follow it. To show that it does not encode the sign, reverse the sign with everything else fixed and observe no change in the output. To show that the lock-in encodes the sign, reverse it and observe the phasor reverse its rotation. A function of a rock's states that happened to track a beat on one afternoon would fail the first of these tests on the next.

The definition also makes the fish's case precise. To show that the fish encodes the sign of Δf, rather than some correlated feature of particular stimuli, the experimenter must reverse the sign while holding the magnitude and the amplitudes fixed, and must do so over the range of geometries in which the neighbor's field falls on the fish's body. The mimic experiments of Heiligenberg and his colleagues, in which the fish's own signal and the neighbor's are replaced by electronic substitutes (their S1 and S2) whose frequencies and amplitudes can be set independently, make tests of this kind possible (Heiligenberg 1991a).

16.7 Channels and Bottlenecks

The relation r, the mixture y, and the reader's output z form a Markov chain, $r \to y \to z$: once y is fixed, z depends on nothing further about r. The data-processing inequality then bounds what any reader can achieve.

Proposition 16.5 (No reader creates information)**.** *For any reader,* $I(r; z) \leq I(r; y)$.

The proof is the standard one for Markov chains (Cover and Thomas 2006). The square-law detector's output carries no more information about the detuning than the mixture did; it carries less, since it has discarded the sign. What the detector does is change the format of the information it keeps. Information about $|\Delta\omega|$ that sat in the mixture at frequencies near ω_1 and ω_2, where a slow linear process could not use it, now sits at frequency $|\Delta\omega|$, where such a process can. Explicitness is a property of format relative to a class of downstream processes, bounded above by information.

A reader that keeps what matters and discards the rest is a compression. The information bottleneck method makes the trade-off precise: among all stochastic maps from y to a compressed variable z, choose one that minimizes $I(y; z) - \beta I(z; r)$ for a relevant variable r and a multiplier $\beta > 0$, trading compression against relevance (Tishby et al. 2000). Much of what readers do can be described in these terms. A reader that outputs only the sign of a detuning keeps one bit per decision out of a waveform that carries a great

deal more, and the bit it keeps is the one a frequency-shifting policy needs. The description is normative: it says what an optimal reader would keep for a given use. Whether biological readers approach the bound is an empirical question, and the bottleneck formulation is useful because it turns it into one. The explicit variable is a physical state, and keeping it, copying it, and erasing it have physical costs (Landauer 1991).

16.8 When the Reader Is a Relatum

In every example so far the relata were external to the reader, or, in the lock-in, one relatum was a reference the reader held for the purpose. In active sensing the reader emits one of the relata itself. The electric fish's own discharge is one term of the mixture on its skin, and a neighbor's discharge, or the distortion caused by an object, is the other (Chapter 19). The variables that the fish's readers make explicit are then of the form "world relative to my own signal." They are centered on the fish by construction, before any model of a self is involved, and Chapters 22 and 34 develop what follows from that.

The lock-in's reference plays for the lock-in the role the fish's discharge plays for the fish, with one difference that turns out to matter. The lock-in's reference is kept inside the instrument. The fish's reference is also what it broadcasts into the world, so the fish's own activity shapes the mixture that it reads. The reader has become one of the relata, and the relation it reads is partly of its own making.

16.9 Audit

Canonical phenomenon

One two-tone situation presented to a linear filter, a square-law detector, an envelope detector, a quadrature lock-in amplifier, and, split between the ears, the binaural auditory pathway.

Minimal model

The mixture $y = \mathsf{M}(A, B)$ and readers $z_i = \mathsf{D}_i(y)$; explicit variables (Definition 16.1); the symmetry bound (Proposition 16.2); the data-processing inequality (Proposition 16.5).

What is relational here?

The magnitude and the sign of the detuning, the beat phase, and, in the binaural case, the interaural phase difference.

What physically reads it?

Each device in turn; in the binaural case, brainstem and midbrain neurons that compare timing between the ears.

What existing field explains it?

Signal processing and communication theory (Oppenheim and Schafer 2010; Horowitz and Hill 2015; Scofield 1994; Shannon 1948; Cover and Thomas 2006); auditory neurophysiology and psychophysics (Goldberg and P. B. Brown 1969; Kuwada et al. 1979; Licklider et al. 1950; Oster 1973); the information bottleneck (Tishby et al. 2000); interventionist causation (Woodward 2003). They explain which variables each reader makes explicit.

What remains unexplained?

Nothing about which variables which devices extract. The residue concerns why some readers exist at all, who built them, and what their readings are for: the questions of Parts V and VI.

A discriminating experiment

Hold the physical mixture fixed and change the reader: swap a square-law detector for a quadrature demodulator, or, in a neural system, change the reader pharmacologically or by adaptation (Chapter 58). A downstream process that followed $|\Delta\omega|$ should come to follow the sign of $\Delta\omega$ only when the new reader breaks the symmetry of Proposition 16.2.

16.10 Notes and References

Demodulation, envelope detection, and quadrature detection are standard in communication engineering; Oppenheim and Schafer (2010) treat the signal processing and Horowitz and Hill (2015) the circuits. Scofield (1994) gives a clear frequency-domain account of the lock-in amplifier, which descends from Dicke (1946). Binaural beats were studied psychophysically by Licklider et al. (1950) and popularized by Oster (1973); the physiology of interaural phase sensitivity begins with Goldberg and P. B. Brown (1969), and Kuwada et al. (1979) recorded inferior colliculus neurons driven by binaural beats. The data-processing inequality is in every information theory text; Cover and Thomas (2006) is standard. The information bottleneck is Tishby et al. (2000). Searle's distinction between intrinsic and observer-relative features is developed in Searle (1992); relational quantum mechanics is Rovelli (1996). Landauer's argument that information is physical is Landauer (1991).

16.11 Exercises

Exercise 16.1. Derive (16.2) for unequal amplitudes $a \neq b$. Show that the output's modulation depth is maximal when $a = b$ for fixed $a^2 + b^2$, and explain why a square-law detector is least informative about the beat when one tone is much weaker than the other.

Exercise 16.2. Derive (16.3), stating the assumption about the low-pass filter that you need. Show that $\arg[u(t) - b]$ increases in time when $\Delta\omega > 0$ and

decreases when $\Delta\omega < 0$. What does the lock-in output when its reference is placed at the mean frequency $(\omega_1 + \omega_2)/2$ instead?

Exercise 16.3. Verify that the transformation T of Section 16.3 leaves the square-law output unchanged and reverses the sign of $\Delta\omega$. Show that T changes the lock-in's output, and identify the feature of the lock-in that breaks the symmetry.

Exercise 16.4. Use the fish's convention for this exercise: $y(t) = \cos\omega_{\mathrm{s}} t + m\cos(\omega_{\mathrm{e}} t + \psi_0)$ with $0 < m < 1$ and $\Delta\omega = \omega_{\mathrm{e}} - \omega_{\mathrm{s}}$. Write y as $A(t)\cos(\omega_{\mathrm{s}} t + \varphi(t))$ with $A\mathrm{e}^{\mathrm{i}\varphi} = 1 + m\mathrm{e}^{\mathrm{i}\psi}$ and $\psi = \Delta\omega t + \psi_0$. Show that $\dot\varphi = \Delta\omega\,(m^2 + m\cos\psi)/(1 + m^2 + 2m\cos\psi)$, evaluate it at the maxima and minima of A, and show that the time average of $A\dot\varphi$ has the sign of $\Delta\omega$. State what a reader must have in order to measure $\dot\varphi$.

Exercise 16.5. Using Proposition 16.5 and Proposition 16.2, show that no processing of the square-law output, however elaborate, can recover the sign of $\Delta\omega$ over the set of mixtures closed under T. Then explain why this does not contradict the fact that the sign can be recovered from y itself.

Exercise 16.6. Design an electrophysiological or psychophysical experiment that tests whether the binaural pathway encodes the *direction* in which the interaural phase difference advances, in the sense of Definition 16.3. Specify the stimuli, the variables held fixed, and the result that would count as a negative answer.

Exercise 16.7. Four waveforms are equally likely: small or large $|\Delta\omega|$, each with either sign. The relevant variable is the sign. Find a deterministic reader $z = \mathsf{D}(y)$ that minimizes $I(y; z)$ subject to $I(z; r) = 1$ bit, and compute $I(y; z)$. Then describe a reader that keeps the magnitude and discards the sign, and compute the same two quantities for it.

17

Does Everything Compute Everything?

> [T]he wall behind my back is right now implementing the Wordstar program, because there is some pattern of molecule movements that is isomorphic with the formal structure of Wordstar.
>
> John R. Searle, *The Rediscovery of the Mind* (1992)

The preceding chapters treated a reader as a physical operation whose output depends systematically on a relation in its input. The word "systematically" hides a problem that philosophers of computation have argued about for four decades. If being a reader, or being a computer, means only that the states of a physical system can be put in correspondence with the states of some abstract process, then almost every system reads almost every relation and computes almost every function, and both concepts are empty. Hilary Putnam proved a theorem to that effect, and John Searle drew the conclusion that computation is in the eye of the beholder.

The thesis of this chapter is that the triviality arguments show what implementation cannot be, and that every serious response to them adds a condition on the physical system itself: counterfactual support, causal organization among parts, mechanistic function, content, or use. The weakly electric fish meets all of these conditions at once. In the fish, the correspondence between physical states and the computational distinction that matters is fixed by the animal's own causal organization and by what the animal does with the result, so no outside interpreter decides which distinction the circuit draws. That is the first precise sense in which a reader can be *endogenous*, and it is where the book's question stops being a question about descriptions and becomes a question about physical systems.

17.1 Physical Computation

Computation is defined abstractly. Turing's machines (Turing 1937), finite-state automata, and their many equivalents specify computations as rule-governed transitions among states, with no mention of what the states are made of. A physical system computes, on the standard view, when it *implements* such an abstract structure, and the implementation relation is the bridge between the mathematics and the world. Everything in this chapter turns on what that bridge requires.

We work with the simplest abstract structures.

Definition 17.1 (Finite-state automaton)**.** An *inputless finite-state automaton* is a pair (Σ, δ) with a finite set of states Σ and a transition function $\delta\colon \Sigma \to \Sigma$. An automaton *with inputs and outputs* adds an input set $\mathcal{I}$, an output set $\mathcal{O}$, a transition function $\delta\colon \Sigma \times \mathcal{I} \to \Sigma$, and an output function $\lambda\colon \Sigma \to \mathcal{O}$.

The most direct account of implementation asks only for a correspondence that respects the transitions.

Definition 17.2 (Simple mapping account)**.** Let P be a physical system whose state at step t is $p(t)$. P *implements* the inputless automaton (Σ, δ) over steps $0, 1, \ldots, T$ if there is a map h from physical states to Σ such that

$$h\big(p(t+1)\big) = \delta\big(h(p(t))\big) \qquad \text{for } t = 0, \ldots, T-1. \tag{17.1}$$

Figure 17.1 draws the condition as a square that must commute: following the physics and then applying h gives the same formal state as applying h and then following the rule. Every account of implementation in this chapter keeps the square. They differ in what they demand of h, of the physical states it is defined on, and of the arrows it is supposed to cover.

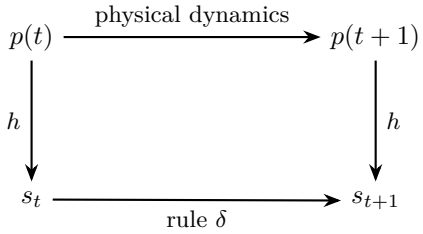


Figure 17.1: The implementation square. A mapping h from physical states to formal states implements a rule when the square commutes. Accounts of implementation differ over which squares must commute (only those on the actual run, or also those on counterfactual runs) and over what may serve as h.

17.2 The Triviality Problem

The simple mapping account asks for too little, and the following result shows how little.

Theorem 17.3 (Triviality for an actual run)**.** *Let P be a physical system whose states $p(0), p(1), \dots, p(T)$ at $T+1$ successive steps are pairwise distinct. Then for every inputless automaton (Σ, δ) and every initial formal state $s_0 \in \Sigma$, the system P implements (Σ, δ) over steps $0, \dots, T$ under the simple mapping account, with $h(p(0)) = s_0$.*

Proof. Define $h(p(t)) = \delta^t(s_0)$ for $t = 0, \dots, T$, where δ^t is the t-fold composition of δ, and let h take arbitrary values on all other physical states. Because the states $p(t)$ are pairwise distinct, h is well defined. Then $h(p(t+1)) = \delta^{t+1}(s_0) = \delta(\delta^t(s_0)) = \delta(h(p(t)))$, which is (17.1). □

The hypothesis costs almost nothing. A physical system subject to the smallest external influence almost never returns to exactly the same microstate, so its states along any run are distinct. The theorem then says that such a system implements every inputless automaton, over any finite stretch of time, in whatever initial state we please. The physics contributes nothing to the result except distinctness; the map does all the work.

The consequence reaches past computation. If "reading a relation" meant only that some map takes the reader's input states to labels of the relation, then every system with distinct states would read every relation, and the reading loop of Chapter 1 would be a way of talking, not a feature of the world. The triviality problem is therefore the reader problem in its most general form.

17.3 Putnam

Putnam (1988, appendix) proved that every ordinary open system is a realization of every abstract finite automaton. An *ordinary open system*, in his sense, is a physical system subject to external influences (gravitational and electromagnetic influences from its surroundings are enough), and his argument relies on two principles about such systems: that they do not pass through the same maximal physical state twice in a given interval (noncyclical behavior), and that their states vary continuously with time. Given an automaton's run, Putnam divides the time interval into subintervals, one per step, and maps to each formal state the disjunction of all physical states the system occupies during the subintervals assigned to that formal state. Distinctness guarantees that the assignment is consistent. Theorem 17.3 is a stripped-down version of the same construction.

Putnam's target was functionalism, the view that to have a mind is to have the right functional organization, where functional organization is spelled out as the realization of an automaton. If every ordinary system realizes

every automaton, then every system has every functional organization, and functionalism in that form collapses. Later commentators have stressed that the theorem concerns inputless automata and finite runs (Chalmers 1996a; Sprevak 2018). Those restrictions turn out to be where the responses gain their grip.

17.4 Searle

Searle reached a similar verdict by a different route and drew a stronger conclusion (Searle 1990; Searle 1992). On the standard definition of implementation, he argued, a sufficiently complex object implements a program whenever some pattern in its physical behavior is isomorphic to the program's formal structure; a wall, with its enormous number of molecular movements, contains such a pattern for any program of modest size. From this he concluded that computational properties are not intrinsic to physical systems. Syntax, on his view, is not a physical property at all: to call a physical state a zero or a one is to assign it a role that exists only relative to an observer who uses it that way. Computation is therefore observer-relative, and a science that explains minds computationally explains nothing about what is intrinsic to brains.

Searle's argument has two parts that are worth separating. The first is the triviality of mapping-based implementation, which he shares with Putnam and which Theorem 17.3 makes precise. The second is a thesis about the nature of syntax: that no physical organization, however rich, makes a state a symbol without an interpreter. The responses surveyed below answer the first part directly. They answer the second only if the conditions they add to implementation can be satisfied without an outside interpreter, and that is exactly the question Section 17.8 puts to the fish.

17.5 Counterfactual Causal Structure

Chalmers (1996a) diagnosed the weakness of the simple mapping account in its restriction to the actual run. A computer running a program does not merely pass through a sequence of states that can be labeled after the fact. It would have gone through different states had its inputs or initial state been different, and it would have done so in accordance with the same rules. The implementation square must commute for counterfactual runs as well as for the actual one.

Definition 17.4 (Counterfactual implementation)**.** A physical system P implements the inputless automaton (Σ, δ) if there is a partition of P's possible states into classes and a map h from classes onto Σ such that, for every formal state s and every physical state p in a class mapped to s, the lawful evolution of P from p reliably reaches a state in a class mapped to $\delta(s)$. The requirement

holds for all formal states and all their physical realizers, whether or not the actual run visits them.

Putnam's construction fails this test. Its map is defined only on the states the system actually occupies, and nothing in the physics guarantees that a slightly different state, assigned to the same formal state, would evolve into a state assigned to the right successor.

Chalmers then showed that the counterfactual requirement is still not enough for inputless automata. Consider a system with two parts: a *clock*, which passes through a long sequence of distinct states, and a *dial*, which can be set to any of several positions and stays where it is set. Map each pair of a clock state and a dial position to a formal state, choosing the dial position to encode the initial state of the automaton and the clock state to encode the number of steps taken. Every transition then holds reliably and counterfactually, because the clock advances and the dial stays put whatever position it is in (Chalmers 1996a). The moral Chalmers drew is that inputless automata are too weak a notion of computational structure: they have no internal organization for an implementation to mirror.

His proposal was the *combinatorial-state automaton*.

Definition 17.5 (Combinatorial-state automaton)**.** A *combinatorial-state automaton* (CSA) has internal states that are vectors $\boldsymbol{s} = (s^1, \ldots, s^k)$ of substates, inputs and outputs that may also be vectors, and transition rules that fix each new substate s'^j as a function of the old state vector and the input. A physical system *implements* a CSA if its state can be decomposed into k distinct components, with each component's physical states mapped to the corresponding formal substates, such that the causal transitions among the components reliably mirror every formal rule, counterfactually as well as actually.

A CSA with k binary substates has 2^k states, and its rules describe how each part changes as a function of the others. To implement it, a physical system must have distinct parts that causally interact in the pattern the rules describe. A wall's molecules do not come organized into components whose interactions reliably track an arbitrary table of rules, and the clock-and-dial trick has no analogue, because a dial cannot stand in for k interacting parts. The CSA account keeps the functionalist's idea of abstract organization while requiring that the organization be *physical*, carried by the causal structure among the system's parts (Chalmers 2011).

Copeland (1996) reached a related verdict from the side of labeling. A labeling of physical states that is devised after the fact, with knowledge of the run it is meant to describe, is not an honest model of the computation; the labeling must be fixed in advance and must support the transitions the specification demands, including those that are never exercised. Both proposals

put the weight on counterfactuals. The next two families of response add conditions of a different kind.

17.6 Mechanistic Accounts

Engineers do not worry that walls run word processors, and the mechanistic account of computation explains why. On Piccinini (2015)'s version, a physical computing system is a functional mechanism: a system of components whose teleological function is to manipulate vehicles according to rules that are sensitive to certain degrees of freedom of those vehicles and insensitive to the medium that realizes them. The rules of a digital circuit are defined over voltage ranges, not over the particular metal that carries the current; that medium independence is what lets the same computation be realized in other materials. A wall has components, but none of them has the function of manipulating vehicles, and so the wall computes nothing, whatever mappings can be drawn onto it. Miłkowski (2013) develops a closely related account in which a computational explanation is a description of an information-processing mechanism, to be tested like any other mechanistic hypothesis: by intervening on its parts and checking whether the system behaves as the description predicts.

The mechanistic view blocks triviality by adding a condition on the system's organization, and its cost is the notion of function it relies on. Something must make it the case that a component has the function of manipulating vehicles, as opposed to merely being describable that way. In artifacts the answer is design. In organisms the answer has to come from the organism: from selection history, from the component's contribution to the organism's self-maintenance, or from both. Part VI takes up that question, and the organizational account of function developed there (Mossio et al. 2009) is one way of paying the mechanist's bill.

17.7 Computation Versus Description

Two further families of response make the role of the interpreter explicit rather than eliminating it.

The *semantic view* holds that computational identity depends on content. A familiar case shows why.

Example 17.6 (One gate, two computations). Take a gate whose output voltage is high exactly when both input voltages are high. If high voltage represents 1, the gate computes AND; if high voltage represents 0, the same gate, with the same physics and the same mechanism, computes OR. Nothing in the voltages decides between the two.

Shagrir (2020) argues from cases of this kind that which computation a system performs is fixed partly by what its states represent, so that an

account of implementation that ignores content leaves computational identity underdetermined. Mechanists reply that the gate's function in the larger system fixes the reading; the reply works when there is a larger system whose use of the gate settles the matter.

The *pragmatic* or representational view makes use central. Horsman et al. (2014) define physical computing as the use of a physical system to predict the outcome of an abstract evolution: a representation relation carries abstract states to physical states, the physical system evolves, and the result is carried back and trusted as the outcome of the abstract process. Computing, on this account, requires a *computational entity* that sets up and uses the representation, and the authors take care to allow that the entity need not be a human user.

Table 17.1 summarizes the families of response in the taxonomy of Sprevak (2018); Godfrey-Smith (2009) surveys the same arguments as they bear on functionalism.

Table 17.1: Responses to the triviality arguments and the condition each adds to the simple mapping account.

Family	**Added condition**	**Representative sources**
Causal and counterfactual	The square commutes for all realizers of all formal states; for CSAs, distinct physical parts mirror the substate rules	Chalmers (1996a) and Chalmers (2011), Copeland (1996)
Mechanistic	Components have the function of manipulating medium-independent vehicles	Piccinini (2015), Miłkowski (2013)
Semantic	The states have content, and content fixes which computation is performed	Shagrir (2020)
Pragmatic and representational	An entity uses the system, through a representation relation, to predict an abstract evolution	Horsman et al. (2014)

The families disagree about much, and they converge on one point. Each locates what the simple mapping account lacks in a feature of the implementing system that an arbitrary wall does not have: reliable causal structure among parts, functional organization, states with content, or a user. Where they diverge is over whether the last two can be supplied from inside the system or must come from outside it. That divergence is the reader problem again, and the fish is a case in which it can be examined in a mapped circuit instead of in thought experiments.

17.8 What an Endogenous Reader Adds

The jamming avoidance response of *Eigenmannia* requires the fish to determine the sign of the frequency difference $\Delta f = f_e - f_s$ between a neighbor's discharge and its own: it lowers its frequency when $\Delta f > 0$ and raises it when $\Delta f < 0$ (Watanabe and Takeda 1963; Metzner 1993b). Chapter 21 derives the computation in full. Here we need only its logical form, which follows from the two-patch model used throughout Part V.

Two patches of skin receive the neighbor's field at different strengths, with contamination ratios $m_2 > m_1$. At each patch the mixture of the two fields is modulated in amplitude and in phase at the beat rate $|\Delta f|$. Define two binary cues and the target variable:

$$U = [\dot{A}_2(t) > 0], \qquad V = [\varphi_2(t) - \varphi_1(t) > 0], \qquad S = [\Delta f > 0], \tag{17.2}$$

where U says whether the amplitude at the more contaminated patch is rising, V says whether that patch leads the other in phase, and brackets denote the truth value of the enclosed statement.

Proposition 17.7 (The sign of Δf is an exclusive or of two cues)**.** *In the two-patch model with small contamination, and at almost every moment of a beat cycle,*

$$S = U \oplus V. \tag{17.3}$$

If the moment of observation is uniformly distributed over the beat cycle, each cue is independent of S on its own, $I(U;S) = I(V;S) = 0$, while the pair determines it: $I(U,V;S) = H(S)$.

Proof. With beat phase $\psi(t) = \Delta\omega\, t + \psi_0$, the model gives $\dot{A}_2 \approx -A_s m_2 \Delta\omega \sin\psi$ and $\varphi_2 - \varphi_1 \approx (m_2 - m_1)\sin\psi$. Their product is approximately $-A_s m_2 (m_2 - m_1)\Delta\omega \sin^2\psi$, which has the sign of $-\Delta\omega$ whenever $\sin\psi \neq 0$. So the cues agree ($U = V$) exactly when $\Delta f < 0$, and disagree exactly when $\Delta f > 0$, which is (17.3). For either sign of Δf, $\sin\psi$ is positive for half of each cycle and negative for the other half, so each cue takes each value half the time regardless of S. □

The proposition is exact within an idealized model: two patches, small contamination, and pure sinusoids.[1] Its content is that the distinction the fish needs is carried only by a relation between two relations. Neither cue, on its own, carries any information about the sign; their combination carries all of it. This is the exclusive or of Chapter 1 in a living animal, and it means that no

[1] The real fish pools the comparison over many patches of skin with different contamination ratios, and the behavior emerges from the whole population (Heiligenberg, Baker, et al. 1978). The proposition states the logical form of the problem each comparison solves; it is not a model of the circuit.

reader that monitors amplitude alone, or phase alone, can do what the fish does.

Now apply the conditions of the preceding sections.

Causal organization among parts. The two cues are carried by physically distinct channels. One class of electroreceptor afferents, the P-units, encodes amplitude; another, the T-units, fires one phase-locked spike per discharge cycle and encodes timing (Scheich et al. 1973). The two classes project along separate pathways to the hindbrain (Metzner and Heiligenberg 1991). Differential phase is computed in the midbrain, where neurons in deeper layers combine amplitude and phase information and become selective for the sign of Δf (C. Carr et al. 1986; Heiligenberg and G. Rose 1985; G. Rose and Heiligenberg 1986b; Heiligenberg and G. Rose 1986). On the motor side, frequency rises and frequency falls are controlled by two separate pathways converging on the pacemaker in a push-pull arrangement (Metzner 1993b). The state of the circuit decomposes into components that carry the two cues and the output, and their interactions carry the rule. This is the kind of organization the CSA account demands.

Counterfactual support. The relevant counterfactuals have been tested by intervention rather than inferred. In the standard preparation, the fish's own discharge is replaced by an electronic mimic whose frequency the experimenter controls, so that Δf can be set to values the animal would never produce on its own; the pacemaker still shifts in the direction the rule prescribes (Heiligenberg 1991a). The African *Gymnarchus* determines the direction of its response from the signal mixture alone, without reference to its own pacemaker command (Kawasaki 1993). And the cues can be manipulated without any neighbor at all: combinations of amplitude and phase modulation that mimic the relation between the cues evoke frequency shifts whose tuning, sensitivity, and dynamics are continuous with the jamming avoidance response (Carlson and Kawasaki 2007). The square of Figure 17.1 commutes on counterfactual runs, and the experiments that show it intervene on the relational cues themselves:

$$\Delta f > 0 \;\Rightarrow\; \text{lower } f_{\mathrm{s}}, \qquad \Delta f < 0 \;\Rightarrow\; \text{raise } f_{\mathrm{s}}. \tag{17.4}$$

Function. The circuit's components have a function in the mechanist's sense, and the function has a source that does not depend on a designer. The response keeps the fish's own frequency away from its neighbor's; the two fish shift their discharge frequencies apart and so avoid jamming each other's electrolocation (Kawasaki 1996). The correct response appears in young fish without training (Viete and Heiligenberg 1991).

Who fixes the mapping? In a laboratory lock-in amplifier, the map from voltages to "the component of the signal at the reference frequency" is fixed

by the engineer who built the instrument and reads its display. In Searle's wall, nothing fixes any map at all, which is why every map is equally available. In the fish, the physical states that count as "$\Delta f > 0$" are those that drive the pathway that lowers the pacemaker frequency, and they count as that because of what they do downstream in the animal. The distinction has been incorporated into the causal architecture. No outside interpreter is needed to decide which distinction the circuit draws, because the circuit's own effects decide it.

This is what an endogenous reader adds to the analysis of implementation: a case in which the conditions that the responses to triviality require from outside (a user, a designer, an assignment of content) are met from inside, by the organization of the system and the use it makes of its own states. It does not yet show that this use is anything more than mechanistic function. An adaptive radio that moves away from an interfering carrier also satisfies counterfactual implementation, has components with functions, and uses its readings to act (*Specification of the Bluetooth System, Version 1.2* 2003); its readings are fixed by engineers who built it for their purposes. The difference between the fish and the radio is not in the input-output function, which they share, but in where the reader comes from and what its use maintains. Chapter 18 asks how many computations a single substrate supports and which of them are exploited; Chapters 25 and 26 ask what it takes for a system to build and maintain its own readers.

17.9 Audit

Canonical phenomenon

The jamming avoidance response: a fish reads the sign of the frequency difference between its own discharge and a neighbor's and shifts its frequency in the direction that increases the difference.

Minimal model

Implementation as a commuting square (Figure 17.1); the triviality theorem for actual runs (Theorem 17.3); counterfactual implementation and combinatorial-state automata (Definitions 17.4 and 17.5); the sign of Δf as the exclusive or of an amplitude cue and a differential-phase cue (Proposition 17.7).

What is relational here?

The sign of Δf, a property of the pair of discharges; its cues, each of which is a relation between the two fields; and the exclusive or, a relation between the cues.

What physically reads it?

Distinct afferent channels for amplitude and timing, their separate hindbrain pathways, sign-selective neurons in the midbrain that combine them, and

push-pull motor pathways to the pacemaker. The mapping from physical states to the distinction is fixed by what those states cause.

What existing field explains it?
The philosophy of computation owns the triviality arguments and all four families of response (Putnam 1988; Searle 1992; Chalmers 1996a; Copeland 1996; Piccinini 2015; Shagrir 2020; Horsman et al. 2014; Sprevak 2018); neuroethology owns the circuit (Heiligenberg 1991a). The exclusive-or form of the sign problem follows from the standard amplitude-phase analysis of the response (G. Rose and Heiligenberg 1986b).

What remains unexplained?
Whether endogenous use adds anything to mechanistic function. The fish and an adaptive radio share the counterfactual structure and the input-output function; what distinguishes them is the origin of the reader and what its use maintains, which the counterfactual and mechanistic accounts do not by themselves address.

A discriminating experiment
Present the cues of (17.2) with their relation reversed at one patch (amplitude rising while that patch lags) and consistent at another, in preparations where the relevant midbrain populations are recorded. If the circuit implements the exclusive or over the pooled population, the behavioral response should follow the population's majority vote and sign-selective neurons should respond according to the combination at their own receptive field; if it merely correlates with Δf, conflicting cue combinations should produce no consistent response.

17.10 Notes and References

Putnam's theorem appears in the appendix of Putnam (1988). Searle's argument is in Searle (1990) and in chapter 9 of Searle (1992); his earlier Chinese room argument (Searle 1980) concerns a different question, whether running a program suffices for understanding, and should not be confused with the triviality argument. Chalmers (1996a) is the classic counterfactual response and introduces the clock-and-dial construction and combinatorial-state automata; Chalmers (2011) places the CSA account at the foundation of computational cognitive science. Copeland (1996) develops the requirement that labelings be honest. Godfrey-Smith (2009) examines what the triviality results mean for functionalism, and Sprevak (2018) is the most useful survey of the arguments and the responses, organized into causal, semantic, teleological, and anti-realist families.

The mechanistic account is developed at book length by Piccinini (2015) and Miłkowski (2013). Shagrir (2020) defends the semantic view against mechanistic critics. Horsman et al. (2014) give the abstraction/representation account

of when a physical system computes, and their notion of a computational entity is the closest point of contact between the philosophy of implementation and the reader problem of this book. Marr (1982)'s distinction between the computational, algorithmic, and implementational levels returns in Chapter 24, where two fish lineages implement the same algorithm in different brain regions. The circuitry of the jamming avoidance response is reviewed by Heiligenberg (1991a); the amplitude-phase analysis of the sign problem is in G. Rose and Heiligenberg (1986b) and Heiligenberg and G. Rose (1986).

17.11 Exercises

Exercise 17.1. Prove Theorem 17.3. Then take the automaton $\Sigma = \{a, b, c\}$ with $\delta(a) = b$, $\delta(b) = a$, $\delta(c) = c$, and a physical run that starts in a state mapped to a. Which formal transition is never exercised on the run? Show that the map constructed in the proof says nothing about how the physical system would behave from a state mapped to c, and explain why this violates Definition 17.4.

Exercise 17.2. Construct Chalmers's clock-and-dial implementation of the automaton of the previous exercise. Specify the clock states, the dial positions, and the mapping, and verify that every transition holds counterfactually. What feature of the automaton makes the construction possible, and why does it fail for a combinatorial-state automaton with two interacting binary substates?

Exercise 17.3. A combinatorial-state automaton has k binary substates and a transition table that specifies each new substate as an arbitrary Boolean function of the old state vector. How many distinct transition tables are there? For $k = 10$, compare this number with the number of distinct inputless automata with 2^{10} states. What does the comparison suggest about how much physical organization an implementation of a CSA requires?

Exercise 17.4. Write the truth table of a gate whose output voltage is high exactly when both inputs are high. Show that it computes AND under one assignment of 0 and 1 to voltage levels and OR under the other. For each of the following, say what (if anything) fixes the assignment: a gate inside a calculator; the same gate lying loose on a bench; the sign-selective neurons of the fish.

Exercise 17.5. In the two-patch model, let ψ be uniform on $[0, 2\pi)$ and let the sign of Δf be $+$ or $-$ with probability $1/2$, independently of ψ. Compute $I(U; S)$, $I(V; S)$, $I(U; V)$, and $I(U, V; S)$ for the cues of (17.2). Now suppose the fish could observe only the sign of the amplitude derivative at two patches, $U_1 = [\dot{A}_1 > 0]$ and $U_2 = [\dot{A}_2 > 0]$. Show that $I(U_1, U_2; S) = 0$.

Exercise 17.6. A laboratory claims that a gene regulatory network in a bacterium "computes" an exclusive or of two nutrient signals. Design the experiments that would establish counterfactual implementation in the sense of

Definition 17.4: which input combinations, which interventions on internal components, which controls. Which additional evidence would be needed to show that the cell uses the result?

Exercise 17.7. Apply the abstraction/representation account of Horsman et al. (2014) to the jamming avoidance response. What is the abstract evolution, what is the representation relation, and what, if anything, plays the role of the computational entity? Compare your answer with the same analysis for an adaptive radio.

18

Polycomputation

The implementation debate of Chapter 17 asked whether a physical system computes one thing, or everything, or nothing in particular. Biology and engineering both suggest a fourth answer: a single substrate can compute several things at once, each available to a different reader. A telephone cable carries many conversations; a recurrent network can feed many readouts; an evolved granular material can report two logical operations from the same grain at two frequencies. Bongard and Levin call this *polycomputing* and argue that living matter does it everywhere (Bongard and Levin 2023).

The thesis of this chapter is that polycomputing is real and common, and that it sharpens the reader problem instead of dissolving it. Once a substrate is allowed to support many computations, the question "what does this substrate compute?" has no single answer. The question that has an answer is which of the describable computations are *exploited* by physical readers inside the system, and that question can be settled by experiment.

18.1 One Substrate, Multiple Computations

The oldest engineered polycomputer is the frequency-division multiplex. Several messages $m_1(t), \dots, m_K(t)$, each varying slowly, modulate carriers at well-separated frequencies $\omega_1, \dots, \omega_K$, and the sum travels on one wire:

$$y(t) = \sum_{k=1}^{K} m_k(t) \cos \omega_k t. \tag{18.1}$$

A receiver tuned to channel j multiplies y by $\cos \omega_j t$ and passes the product through a low-pass filter. By the product-to-sum identity (Appendix A), the product contains $\frac{1}{2} m_j(t)$ near zero frequency and every other term near a frequency $\omega_k \pm \omega_j$ that the filter removes, so the receiver recovers $\frac{1}{2} m_j(t)$ (Oppenheim and Schafer 2010). Each receiver reads a different message from the same physical signal, and none of them disturbs the others.

Definition 18.1 (Polycomputing substrate)**.** A *polycomputing substrate* is a physical system S, with state trajectory $y(t)$, together with a family of

physically realized readers $\mathsf{D}_1, \ldots, \mathsf{D}_K$ such that each reader's output $z_k = \mathsf{D}_k(y)$ implements a computation C_k in the counterfactual sense of Chapter 17, and the computations are carried by the same trajectory over the same interval of time.

Two features of the definition matter. The readers are physical: a demodulator, a readout neuron, a measurement apparatus, not an arbitrary map drawn by a theorist. And the computations are simultaneous: the substrate is not switched between tasks, as a general-purpose computer is, but carries all of them in one trajectory. Without the first feature the definition inherits the triviality problem of Chapter 17, since any substrate would support any computation under some map. Without the second it describes ordinary reuse.

The multiplex is designed so that its channels do not interfere, and its readers are built by the same engineers who built the transmitter. The more interesting cases are those in which neither is true.

18.2 Observer-Dependent Decompositions

Bongard and Levin define polycomputing as "the ability of the same substrate to simultaneously compute different things, and make those computational results available to different observers" (Bongard and Levin 2023). They argue for an observer-dependent, pragmatic view of biological computation, in which the question of what a system computes is always relative to an observer that reads it, and in which the overloading of several functions on the same hardware is a design principle of evolved systems.

The word "observer" carries the weight, and it can be read in two ways. Read as an external interpreter (a scientist choosing a map), observer dependence is the triviality problem restated: every substrate computes whatever any theorist can map onto it. Read as a physical system coupled to the substrate (a downstream neuron, a cell responding to its neighbors' membrane voltage, a frequency-selective sensor), observer dependence is the claim of Chapter 16 applied to computation: different physical readers make different variables of the same substrate dynamically explicit. The second reading is the one that makes polycomputing a scientific claim, and it is the one we adopt.

Remark 18.2. On the second reading, "observer dependent" means "reader dependent," and a reader is itself a physical system with a counterfactual structure that can be tested. The decomposition of a substrate into computations is then relative to the set of readers that exist, not to the set of descriptions that could be written down.

Levin's wider program supplies the biological motivation for taking readers seriously. On his account, a self is demarcated by the spatiotemporal boundary of the events it can measure, model, and try to affect, which he calls its cognitive light cone (Levin 2019), and developmental bioelectricity, the use of

ion channels and gap junctions to form electrical networks among non-neural cells, is one medium through which cells come to read the states of other cells (Levin 2022). The empirical content of these proposals is still being tested. Their structure is the one this chapter needs: a tissue is a substrate whose states are read by many internal readers at once, each with its own reach.

18.3 Granular Polycomputation

Parsa et al. (2023) made the idea concrete in a material with no electronics at all. In a granular assembly, vibration can be treated as the information-bearing mode: inputs are vibrations applied at designated grains, and outputs are read from the vibration of another grain. Using evolutionary search over the composition of the assembly, they produced a material in which a single grain acts simultaneously as two different NAND gates at two different frequencies. NAND gates are universal, in the sense that any Boolean function can be built from them, and the authors found the polycomputation distributed across each evolved material, a property they take to suggest robustness.

Example 18.3 (Two gates in one grain). Label the two frequencies f_a and f_b. At the output grain, the component of the vibration at f_a reports one NAND of the inputs, and the component at f_b reports another (Parsa et al. 2023). A sensor that filters the grain's motion around f_a reads the first gate; a sensor that filters around f_b reads the second. Neither gate is "the" computation of the material. Which one is performed, for any purpose, depends on which filter is attached.

The example has the structure of the multiplex with one difference that matters. The channels of the multiplex were designed to be independent; the gates of the granular material were found by search, and their independence is an empirical property of the evolved configuration. In both cases, the substrate by itself reports nothing. The readers, here frequency-selective measurements set up by the experimenters, turn implicit components of the grain's motion into outputs. This is Chapter 3 and Chapter 4 again: a frequency component of a vibration becomes a variable only when some operation selects it.

18.4 Latent Computations

A substrate can carry far more computations than anything reads. The limit on what can be read is set by information theory.

Theorem 18.4 (What a reader can extract)**.** *Let the substrate's trajectory Y and a variable C of interest be random variables, and let a reader produce $Z = \mathsf{D}(Y)$, where D may be noisy but depends on C only through Y. Then*

$$I(Z;C) \le I(Y;C). \tag{18.2}$$

Proof. The variables form a Markov chain $C \to Y \to Z$, and (18.2) is the data processing inequality (Cover and Thomas 2006). □

The theorem puts a ceiling over every reader: no operation on the substrate extracts more about C than the substrate carries. It says nothing about whether any reader reaches the ceiling. That gap defines the next notion.

Definition 18.5 (Latent computation)**.** A computation C of a system's inputs is *latent* in a substrate S if $I(Y;C)$ is large (so that some reader could extract C) but no physically realized reader in the system extracts it.

Reservoir computing turns latent computation into a technique. A fixed recurrent network, driven by an input stream, is read by one or more trained readouts, and several readouts, each trained for a different task, can extract different functions of the input history from the same network at the same time (Maass et al. 2002). Before the readouts are trained, those functions are latent in the network's state. Training a readout does not change the network; it creates a reader.

The fish supplies a biological case. At any pair of skin patches, the amplitude and differential-phase modulations carry the sign of the frequency difference as an exclusive or (Chapter 17). The information is present at every patch pair on the body. It becomes a variable for the fish because midbrain neurons combine the two cues and become selective for the sign of Δf (Heiligenberg and G. Rose 1985; G. Rose and Heiligenberg 1986b). A single amplitude-coding afferent carries the magnitude of the beat, and by Theorem 18.4 applied to one cue alone, nothing about its sign.

18.5 Exploited Versus Merely Describable Mappings

The distinction the chapter needs can now be stated.

Definition 18.6 (Exploited mapping)**.** A mapping C defined on a substrate S is *exploited* by a system containing S if the system contains a physical reader D such that (i) the output of D depends on C counterfactually, in the sense of Chapter 17, and (ii) the output of D has downstream effects on the system's behavior or maintenance. A mapping that satisfies neither condition is *merely describable*.

Exploitation is a property of a system, not of a substrate. The same trajectory can carry a mapping that one system exploits and another leaves latent, depending on which readers are attached. This is why polycomputing does not reopen the triviality problem: a substrate supports indefinitely many describable mappings, and the claim that one of them is exploited is a claim about a specific physical reader, which can be tested.

The tests are interventions. To show that C is exploited, change C while holding other features of the substrate as fixed as possible, and look for the corresponding change in the reader's output and downstream; remove or silence the reader, and look for the loss of the downstream effect; present values of C that never occur naturally, and look for the reader's response. The fish experiments of Chapter 17 have exactly this form: the relation between the cues is set by the experimenter, and the pacemaker responds (Carlson and Kawasaki 2007).

Proposition 18.7 (Exploitation requires a counterfactual path)**.** *If a mapping C is exploited by a system in the sense of Definition 18.6, then some intervention on C (holding the rest of the substrate fixed) changes the system's downstream behavior or maintenance.*

Proof. By condition (i), an intervention on C changes the output of D; by condition (ii), that output has downstream effects. Composing the two dependencies gives the claim. □

The proposition is simple, and its value is in the contrapositive: if no intervention on C makes a downstream difference, then C is not exploited, however elegantly it can be described. Grid-cell interference models and retinal moiré models of orientation maps, reviewed in Chapter 47, were beautiful describable mappings whose exploitation the evidence did not support.

18.6 Internal Readers

In an organism the readers are inside. Write $S \xrightarrow{\mathsf{D}_1} C_1$ and $S \xrightarrow{\mathsf{D}_2} C_2$ for two readers extracting two computations from the same substrate (Figure 18.1). The substrate might be the spike trains of a population of afferents, the membrane voltages of a sheet of cells, or the vibration of a grain.

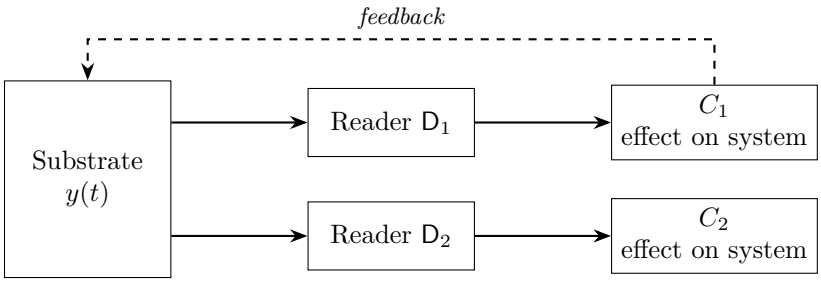


Figure 18.1: Internal polycomputing. Two physical readers extract different computations from one substrate. When a reader's output feeds back on the substrate (dashed), the substrate is read by a part of itself.

The electrosensory system of the wave-type fish is such an arrangement. The same electroreceptor afferents carry the amplitude and timing information

used for electrolocation and for the jamming avoidance response (Heiligenberg 1991a). G. Rose, C. Keller, et al. (1987) recorded from the midbrain of *Sternopygus*, a gymnotiform that performs no jamming avoidance response, and found neurons that discriminate the sign of the difference frequency between interfering discharges. They suggested that such sign-selective neurons serve a more general function in detecting moving objects and conspecifics, and that they could have been assembled into a jamming avoidance response in the course of evolution. Kawasaki reached a parallel conclusion from the independently evolved response of *Gymnarchus*: its computational rules are identical to those of *Eigenmannia*, which suggests that in both lineages the circuit evolved from preexisting mechanisms for electrolocation (Kawasaki 1993). On this reading, the jamming avoidance response is a second use attached, in the course of evolution, to a reading that already existed. *Sternopygus* shows the intermediate case: a relation read by neurons, and exploited for something, without being used to control the discharge frequency.

18.7 Can One Subsystem Read Another?

Yes, and in the sense of Chapter 17 this is routine. The midbrain of the fish reads the hindbrain; the diencephalic nucleus that selects the direction of the frequency shift reads the midbrain; the prepacemaker nuclei read that nucleus; the pacemaker reads the prepacemaker nuclei (C. H. Keller et al. 1990; Metzner 1993b). Each stage's physical states drive the next stage counterfactually, and each stage makes explicit a variable that was implicit at the stage before. The chain is a sequence of readers, each exploiting a mapping on the output of the last.

What such a chain does not settle is whether any stage reads *for* the system as a whole or only for the next stage. In an engineered pipeline the question has a clear answer: the designer's purpose unifies the stages. In an organism the unity has to come from the organism, and that is the question of Part VI.

18.8 Can a Substrate Read Itself?

A substrate reads itself when a reader that is part of the substrate responds to the substrate's own states and its response changes those states, which is the dashed arrow of Figure 18.1. Two worries attach to the idea.

The first is regress. If a reader D is part of S and reads S, it reads its own state among others, and an explanation of what D reads seems to require an explanation of what reads D. The regress stops when the reader responds to a relation that includes its own activity without needing to represent that activity separately. The fish is the example. Its receptors respond to the mixture of its own field and its neighbor's, and the relation it extracts (the sign of Δf) has the fish's own frequency as one of its terms. Nothing in the circuit

reads the circuit; the circuit reads a relation of which the fish is a relatum (Kawasaki 1993). Chapter 22 develops this point, and Part IX takes up the systems in which readers do read other readers.

The second worry is that self-reading is too cheap: every feedback loop is a substrate acting on itself. The answer again lies in Definition 18.6. A thermostat's switch reads the temperature of a room that the switch helps to set, and it exploits exactly one mapping, the one its designer chose. A polycomputing substrate that reads itself exploits several mappings at once, through several readers, and the readers' outputs interact in the substrate they read. Whether that difference matters for anything beyond engineering is an open question, and Chapter 26 states the version of it that has a literature.

18.9 Audit

Canonical phenomenon

One grain of an evolved granular material acts as two NAND gates at two frequencies; one population of electroreceptor afferents serves electrolocation and jamming avoidance.

Minimal model

Frequency-division multiplexing (18.1); the polycomputing substrate (Definition 18.1); the data processing bound on what any reader can extract (Theorem 18.4); latent and exploited mappings (Definitions 18.5 and 18.6).

What is relational here?

The computations C_k are functions of the substrate's trajectory relative to a reader: a frequency component of a vibration, a combination of amplitude and phase cues, a linear readout of a network state.

What physically reads it?

Demodulators, frequency-selective sensors, trained readouts, and, in the fish, midbrain neurons that combine amplitude and phase. Without them the computations are latent.

What existing field explains it?

Communication engineering owns multiplexing (Oppenheim and Schafer 2010); information theory owns the extraction bound (Cover and Thomas 2006); reservoir computing owns multiple readouts of one substrate (Maass et al. 2002); unconventional computing owns polycomputing materials (Parsa et al. 2023); Bongard and Levin (2023) name and generalize the idea for biology.

What remains unexplained?

The chapter replaces "what does this substrate compute?" with "which mappings are exploited, by which readers?" That question is answerable case by case. What it leaves open is whether a system whose readers are

all internal exploits mappings for itself or only for its parts, which is the question of Part VI.

A discriminating experiment

In a system proposed to polycompute (a tissue, a granular material, a neural population), identify a candidate mapping C, intervene on C while holding other features of the substrate fixed, and measure downstream effects. Then silence the candidate reader and repeat. A mapping is exploited if and only if the first manipulation has effects that the second abolishes; otherwise it is latent or merely describable.

18.10 Notes and References

The term *polycomputing* and its application to biology are due to Bongard and Levin (2023), who review biological and technological examples and argue for an observer-centered framework. Parsa et al. (2023) report the granular material with two NAND gates at two frequencies. Levin's framework for diverse intelligences and his proposal that developmental bioelectricity scales cell-level feedback into tissue-level agency are developed in Levin (2019) and Levin (2022). The liquid state machine of Maass et al. (2002) is one of the two founding formulations of reservoir computing, in which a fixed dynamical system supports many trained readouts. The data processing inequality and related bounds are in Cover and Thomas (2006). On multiplexing and demodulation see Oppenheim and Schafer (2010). For the view that physical computing requires an entity that uses a system to predict an abstract evolution, which is the pragmatic counterpart of the reader-based view adopted here, see Horsman et al. (2014). The evolutionary argument that jamming avoidance was built on electrolocation circuitry is in G. Rose, C. Keller, et al. (1987) and Kawasaki (1993).

18.11 Exercises

Exercise 18.1. For the multiplex (18.1) with $K = 2$, multiply $y(t)$ by $\cos \omega_1 t$ and use the product-to-sum identity to list every frequency component of the product. If each message has bandwidth B, what separation between ω_1 and ω_2 lets a low-pass filter recover m_1 without contamination from m_2?

Exercise 18.2. Prove (18.2) from the chain rule for mutual information. Then use it to show that a reader of the amplitude modulation at one skin patch of the fish can extract no information about the sign of Δf when the beat phase is uniformly distributed.

Exercise 18.3. In the two-patch model of the fish, classify each of the following as exploited or latent for the jamming avoidance response, and justify each answer with an intervention: the beat rate $|\Delta f|$; the sign of Δf; the absolute

phase of the field at one patch; the ratio m_2/m_1 of the two contamination ratios.

Exercise 18.4. Simulate a random recurrent network of $N = 200$ tanh units driven by a scalar input $u(t)$. Train two linear readouts on the same network state, one to output $u(t-5)$ and one to output $u(t)\,u(t-2)$. Verify that both succeed at the same time. Which computation does the network perform? Answer the question twice, once for the network alone and once for the network with its readouts.

Exercise 18.5. In the granular material of Example 18.3, the frequency-selective readers are laboratory instruments. Sketch a physical addition to the material that would exploit both gates without an experimenter: what would read each frequency component, and what downstream effect would each reading have on the material?

Exercise 18.6. A colleague claims that a sheet of cultured cells polycomputes, carrying one signal in slow membrane-voltage patterns and another in fast calcium transients. Design the interventions of Section 18.5 that would establish whether each signal is exploited by the cells, and state the result that would count against the claim.

Exercise 18.7. Write the fish's reading loop in the notation of Figure 18.1, with the fish's own discharge as part of the substrate. Show that the reader responds to a relation that includes its own activity without any component that represents the reader's state, and explain why no regress arises.

A reader is a physical operation whose output depends systematically on a relation in its input. Different readers make different relational variables dynamically explicit, and none of this requires metaphysics. The triviality arguments of Putnam and Searle show what goes wrong when the reader is allowed to be an outside interpreter; counterfactual and mechanistic accounts of implementation show how to block them. The question that survives is which distinctions are exploited by causal subsystems inside the system itself. Part V follows one animal in which the whole chain, from waveform to behavior, has been mapped.

Part V

The Electric Fish: A Complete Reader

Which neural circuits extract the relation?

19

Active Electrosensation

Most senses receive energy that the world supplies. Light arrives from the sun and bounces off surfaces; sound arrives from whatever vibrates; odorants arrive on the wind. A weakly electric fish supplies its own carrier. Hundreds of times a second an electric organ in its tail drives current into the water, the current flows back to the fish's skin, and electroreceptors in the skin measure how the returning field has been shaped by whatever lies nearby. The fish does not sense the world directly. It senses the world's effect on a signal that the fish itself produced.

That arrangement is the subject of this part of the book, and this chapter establishes its first consequence. In active electrosensation the variable a receptor measures is, from the first synapse, a relation between the animal's own emission and the environment's distortion of it. The fish's sensory variables are indexed to its own activity before any self-model, memory, or inference is involved. Later chapters build on that fact: Chapter 20 shows what happens when a second emitter enters the same water, Chapter 21 shows how the fish reads the resulting relation, and Chapter 22 places the whole system in the history of ideas about self-generated sensation.

19.1 The Electric Organ Discharge

Weakly electric fishes emit fields too weak to stun prey or deter predators; the fields serve for sensing and for communication (Lissmann 1958; Heiligenberg and Bastian 1984). Two broad signal types exist. *Pulse-type* species emit brief discharges separated by silent intervals, at rates that vary with behavior. *Wave-type* species emit a continuous, quasi-sinusoidal discharge whose frequency is held nearly constant. Two wave-type fishes carry this part of the book. *Eigenmannia*, the glass knifefish, is a South American gymnotiform whose electric organ discharge (EOD) is a continuous wave at a few hundred hertz (Heiligenberg 1991a). *Gymnarchus niloticus* is African; its discharge is a continuous wave of about 400 Hz (Kawasaki and Y. Guo 1996), and it belongs to a lineage whose other members, the mormyrids, emit pulses.

The discharge is driven by a medullary pacemaker nucleus whose neurons fire once per cycle and set the frequency (Metzner 1993b). Each pacemaker cycle recruits electromotor neurons, which fire the electric organ, which drives current through the water around the fish and back through its skin. The fish is therefore both the source of a field and an object within it: the current that leaves the tail returns through the head and trunk, and the skin carries a transepidermal voltage whose spatial pattern depends on the fish's own body geometry and on everything nearby that conducts better or worse than water.

The story of how this became known begins with Hans Lissmann. In 1951 he reported that *Gymnarchus* emits continuous weak electric signals from its tail (Lissmann 1951). With Machin he then showed that the fish locates objects by the distortions they cause in its self-generated field, and worked out the physics of that distortion (Lissmann and Machin 1958). The discovery changed the status of the electric organ in weakly electric fish from a curiosity to a sense organ's transmitter.

19.2 Electroreceptors

Two families of electroreceptors are relevant here (Heiligenberg and Bastian 1984; Bullock, Hopkins, et al. 2005). *Ampullary* receptors respond to low-frequency fields, including the weak bioelectric fields of other animals; they serve passive electroreception, and many fishes that emit no field at all have them. *Tuberous* receptors respond to the high-frequency field of the fish's own discharge; they are the receivers of the active system.

In *Eigenmannia*, tuberous afferents fall into two physiological classes that divide the work of reading the field (Scheich et al. 1973; Mathieson et al. 1987). *P-units* (probability coders) fire with a probability per EOD cycle that increases with the local amplitude of the field; their firing rate reports amplitude. *T-units* (time coders) fire one spike per EOD cycle, locked to a fixed phase of the local signal; their spike times report timing. The separation matters later. Amplitude and timing are the two modulations that a neighbor's field imposes on the fish's own, and they travel along separate pathways into the brain before they are combined (Chapter 21).

19.3 The Self-Generated Carrier

Consider a patch of skin at location $\boldsymbol{x}$ on the fish's body. With no object nearby, the local transepidermal signal is a carrier at the fish's own frequency,

$$y_0(\boldsymbol{x}, t) = A_0(\boldsymbol{x}) \cos(\omega_s t + \vartheta(\boldsymbol{x})), \tag{19.1}$$

where $A_0(\boldsymbol{x})$ is the unperturbed local amplitude and $\vartheta(\boldsymbol{x})$ a local phase offset set by the geometry of the fish and the water. An object whose electrical properties differ from the water's distorts the field. To first approximation for

purely resistive objects, the distortion changes the local amplitude and leaves the carrier frequency alone:

$$y(\boldsymbol{x}, t) = A_0(\boldsymbol{x})\big(1 + \epsilon(\boldsymbol{x})\big)\cos(\omega_{\mathrm{s}} t + \vartheta(\boldsymbol{x})). \tag{19.2}$$

Definition 19.1 (Electric image)**.** The *electric image* of an object is the spatial pattern of fractional amplitude change it produces on the fish's skin,

$$\epsilon(\boldsymbol{x}) = \frac{A(\boldsymbol{x}) - A_0(\boldsymbol{x})}{A_0(\boldsymbol{x})},$$

where A and A_0 are the local amplitudes with and without the object.

A good conductor concentrates current and raises the local amplitude over the region of skin nearest to it; an insulator diverts current and lowers it (Lissmann and Machin 1958; Emde 1999; Caputi and Budelli 2006). Objects with capacitive properties, as living tissue has, also alter the timing and waveform of the local signal, and some species exploit that difference to distinguish living from non-living objects (Emde 1999).

The electric image is a relational variable in the sense of Chapter 1, and of a particular kind. Its relata are the fish's own emission and the object. It is a ratio of what the skin receives to what the skin would have received from the fish's emission alone.

Model 19.2 (Idealized sphere in a uniform field)**.** Treat the fish's field near a small object as uniform, with strength E_0, and the water and the object as ohmic conductors with conductivities σ_{w} and σ_{o}. Solving Laplace's equation for a sphere of radius a gives a perturbation of the potential outside the sphere of dipolar form,

$$\delta V(r, \theta) = \chi\, E_0\, \frac{a^3}{r^2}\cos\theta, \qquad \chi = \frac{\sigma_{\mathrm{o}} - \sigma_{\mathrm{w}}}{\sigma_{\mathrm{o}} + 2\sigma_{\mathrm{w}}}, \tag{19.3}$$

where r is the distance from the sphere's center and θ the angle from the field direction. The contrast factor χ runs from $-\frac{1}{2}$ for a perfect insulator to $+1$ for a perfect conductor, and vanishes when the object matches the water.

The model is the standard electrostatic idealization; the fish's field is neither uniform nor far from the body, and the skin's own conductance shapes the image, so real images are computed numerically (Caputi and Budelli 2006). Two features survive every refinement. The sign of the image reports whether the object conducts better or worse than the water. And the image is *linear in the fish's own field*: doubling the emission doubles both A_0 and $A - A_0$.

Proposition 19.3 (Self-normalization of the electric image)**.** *In a linear conducting medium, the electric image $\epsilon(\boldsymbol{x})$ is independent of the strength of the fish's own discharge.*

Proof. In a linear medium, the field produced by the electric organ, and the field perturbed by any fixed configuration of objects, both scale linearly with the organ's output. Scaling the output by a factor $\lambda > 0$ therefore multiplies $A(\boldsymbol{x})$ and $A_0(\boldsymbol{x})$ by the same λ, and their ratio is unchanged. □

The proposition is the electrosensory case of the invariance argument of Chapter 1. Discharge amplitude varies with the fish's condition and with water conductivity, and any variable that tracks the object without tracking these nuisances must be a ratio: the world relative to the fish's own emission. Because the emission is the fish's own, the fish can hold that reference nearly constant, and its brain can compare the amplitude on one patch of skin with the amplitude on others, or with the same patch a moment earlier. The reference is carried by the signal itself.

19.4 The Reading Loop of Electrolocation

In the notation of Chapter 15, electrolocation instantiates the first three terms of the reading loop with unusual clarity.

- *Agent state* s_{A}: the parameters of the fish's discharge (frequency, amplitude, waveform) and the position and posture of its body.
- *Environment state* s_{E}: the conductivity, capacitance, size, and position of nearby objects.
- *Mixing* $y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})$: the physics of current flow in the water, which delivers to each patch of skin a signal shaped jointly by the emission and the objects, (19.2).
- *Reading* $z = \mathsf{D}(y)$: transduction by tuberous receptors (P-units for amplitude, T-units for timing) and the central comparisons that extract image contrast.

The loop closes through movement. Fish swim, bend, and hover in ways that change the geometry of their own field relative to objects, and those movements are shaped by the sensory feedback they produce. *Eigenmannia* tracking a moving refuge makes small back-and-forth movements that alter the spatiotemporal pattern of electrosensory input (Stamper, Roth, et al. 2012), and those active sensing movements are regulated in closed loop so as to control the sensory slip between fish and refuge (Biswas et al. 2018). Chapter 59 returns to these experiments as a model for how loops can be opened and closed deliberately.

Example 19.4 (Two readers of one electric image). A metal rod and a plastic rod of the same size, placed at the same distance from the fish, produce electric images of the same shape and opposite sign. A reader of local amplitude alone, applied to one patch of skin, reports a number; the number is uninterpretable until it is compared with the amplitude the patch would have had without the object. A reader of spatial contrast across patches, or of temporal contrast

as the fish moves, reports the sign of χ directly and does not need to know the fish's own emission strength. The second reader is the one the fish's brain implements: the variable that matters is relational at the level of the skin and relational again at the level of the comparison.

19.5 Active and Passive Sensing

The term *active sensing* is used in two ways in the literature, and the distinction matters for what follows (M. E. Nelson and MacIver 2006). In the narrow sense, an active sensing system emits the energy it senses: electrolocation, echolocation in bats and toothed whales. In the broader sense, any sensing in which the animal's own movements shape the input is active: whisking, haptic exploration, saccadic vision. Both senses share a structure that matters to this book.

Definition 19.5 (Self-indexed sensory variable)**.** A sensory variable $z = \mathsf{D}(\mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}}))$ is *self-indexed* if it depends on the environment only through a relation between an environmental quantity and a quantity the agent itself produces or controls.

The electric image is self-indexed: it is a ratio of perturbed to self-generated field. A bat's echo delay is self-indexed: it is the interval between the bat's call and the returning echo. A passively sensed light intensity is not self-indexed, although it can be made so by a movement that changes the relation between eye and source.

Passive electroreception provides the contrast within a single animal. When ampullary receptors detect the low-frequency field of a prey animal, the measured variable is the world's own field; the fish contributes only its position. When tuberous receptors detect an electric image, the measured variable exists only because the fish emitted the carrier. The same animal thus has a passive sense and an active one, and the active one produces self-indexed variables at the receptor. We call the resulting structure an *agent-centered measurement frame*: information available to the system is indexed by the system's own intervention. Chapter 22 develops the idea with the established vocabulary of reafference, and Chapter 34 follows it toward the question of perspective.

The step should not be overread. A radar also measures the world relative to its own emission, and an automatic door senses the reflection of the infrared light it emits. Self-indexing is necessary for the kind of perspective this book pursues and far from sufficient for it; Chapter 48 asks what, if anything, the fish has that the radar lacks.

19.6 The Vulnerability of a Self-Generated Carrier

A carrier that the fish emits is a carrier that another fish can disturb. Wave-type fishes often live in groups, and the density of neighbors varies across species in ways that correlate with the electrosensory interference they face (Stamper, Carrera-G, et al. 2010). When a neighbor discharges at a frequency close to the fish's own, the neighbor's field adds to the fish's field on its skin. The sum is modulated in amplitude and in phase at the difference of the two frequencies, and those modulations fall in the same range as the slow modulations produced by objects as the fish swims past them (Heiligenberg 1991a). Electrolocation, which depends on reading small changes relative to a stable carrier, is degraded.

The fish's solution is to move its own frequency away from its neighbor's, and to do so in the right direction. That behavior, the jamming avoidance response, is the subject of Chapter 20. It requires a reading this chapter has not yet needed: not the ratio of perturbed to unperturbed field, but the sign of a frequency difference between two emitters, one of which is the reader itself.

19.7 Audit

Canonical phenomenon

A weakly electric fish detects nearby objects by the distortion they cause in the field of its own electric organ discharge.

Minimal model

The local skin signal $y = A_0(1+\epsilon)\cos(\omega_s t + \vartheta)$; the electric image $\epsilon(\boldsymbol{x})$ (Definition 19.1); the dipolar sphere model with contrast factor χ (Model 19.2); self-normalization (Proposition 19.3).

What is relational here?

The electric image: a ratio between what the skin receives and what the fish's own emission alone would deliver. Its relata are the fish's discharge and the object.

What physically reads it?

Tuberous electroreceptors (P-units for amplitude, T-units for timing) and central circuits that compare amplitude across patches of skin and across time.

What existing field explains it?

Electrostatics (Laplace's equation for a conducting sphere), sensory physiology of electroreception (Heiligenberg and Bastian 1984; Bullock, Hopkins, et al. 2005), active sensing (M. E. Nelson and MacIver 2006), and the electrolocation literature from Lissmann and Machin (1958) onward. Nothing in the chapter goes beyond them.

What remains unexplained?

None in the mechanism. The chapter isolates one structural fact for later use: in active sensing, the sensory variable is self-indexed at the receptor, before any self-model exists. Whether that fact bears on perspective is a question for Parts VIII and X.

A discriminating experiment

Scale the fish's emission artificially (in a curarized preparation with an electronic substitute for the discharge) while holding objects fixed. If central electrosensory responses to an object track the image ratio ϵ rather than absolute amplitude, the system reads the self-indexed variable, as Proposition 19.3 predicts it should.

19.8 Notes and References

The discovery of active electrolocation is due to Lissmann (1951) and Lissmann and Machin (1958); Lissmann (1958) discusses the function and evolution of electric organs more broadly. Heiligenberg and Bastian (1984) reviews the electric sense of weakly electric fish, and the handbook edited by Bullock, Hopkins, et al. (2005) covers electroreception across taxa. Emde (1999) reviews active electrolocation of objects, including capacitance detection; Caputi and Budelli (2006) treat the peripheral formation of electric images in detail. The receptor classes of *Eigenmannia* were characterized by Scheich et al. (1973). M. E. Nelson and MacIver (2006) analyze active sensing systems in general, including the sensory consequences of emission and movement; Stamper, Roth, et al. (2012) and Biswas et al. (2018) show how movement enters the electrosensory loop. Heiligenberg's monograph (Heiligenberg 1991a) remains the best single account of the *Eigenmannia* system from receptor to behavior.

19.9 Exercises

Exercise 19.1. Derive (19.3) by solving Laplace's equation outside and inside a sphere with the boundary conditions of continuous potential and continuous normal current density. Check the limits $\sigma_o \to \infty$ and $\sigma_o \to 0$.

Exercise 19.2. In the idealized model, the fish's own field near the body falls off with distance, and the object's perturbation must travel back to the skin. Assuming a dipolar field for the fish and the dipolar response (19.3) for the object, estimate how the amplitude of the electric image scales with the object's distance. Explain why this idealization overstates the steepness for objects close to the body, and what it implies about the range of electrolocation relative to the fish's size.

Exercise 19.3. Prove Proposition 19.3 directly from (19.3). Then suppose the water's conductivity changes. Which features of the electric image are still invariant, and which change?

Exercise 19.4. Classify the following sensory variables as self-indexed or not, in the sense of Definition 19.5: a bat's echo delay; the loudness of a distant thunderclap; the texture signal from a rat's whisker sweeping a surface; the temperature sensed by a stationary thermoreceptor; the optic flow produced by walking. For those that are not, describe a movement or emission that would make them so.

Exercise 19.5. Write explicit expressions for the output of the two readers in Example 19.4: one that reports $A(\boldsymbol{x}_1)$ at a single patch, and one that reports $A(\boldsymbol{x}_1)/A(\boldsymbol{x}_2)$ for two patches at different distances from the object. Show that the second is invariant under changes of emission strength and the first is not.

Exercise 19.6. Describe a pulse radar in the notation M, D, s_{A}, s_{E} used in Section 19.4. List every structural feature of the fish's electrolocation loop that the radar shares, and every feature it lacks. Keep the list; Chapter 48 returns to it.

20

The Jamming Problem

Two wave-type fishes that meet with nearly equal discharge frequencies each become part of the other's sensory world. Each fish's skin receives its own carrier plus its neighbor's field, and the sum beats. The beat modulates exactly the quantities that electrolocation depends on, amplitude and timing, at a rate that can be as slow as the modulations produced by objects. The fish's response is to change its own frequency, in the direction that increases the difference, until the beat is fast enough to be ignored.

This chapter treats that behavior, the jamming avoidance response (JAR), as the book's first complete example of a reader inside the relation it reads. The thesis is structural. In the JAR an animal changes one of the relata (its own discharge frequency) in order to change the relation (the beat) that degrades its measurements, and the change is driven by a reading of that relation. Everything about how the reading works is deferred to Chapter 21; here we establish what the relation is, why it is a problem, and what the loop that removes it looks like.

20.1 Two Nearby Discharge Frequencies

Let the fish's own discharge produce, at a patch of skin, a carrier of amplitude A_s and angular frequency $\omega_\mathrm{s} = 2\pi f_\mathrm{s}$, and let a neighbor's discharge contribute a field of amplitude A_e and angular frequency $\omega_\mathrm{e} = 2\pi f_\mathrm{e}$ at the same patch. Write $m = A_\mathrm{e}/A_\mathrm{s}$ for the *contamination ratio* at the patch. The neighbor is usually farther from the patch than the fish's own electric organ, so $m < 1$, and because the neighbor's field falls off across the fish's body, m differs from patch to patch. The local signal is

$$y(t) = A_\mathrm{s}\big[\cos(\omega_\mathrm{s} t) + m\cos(\omega_\mathrm{e} t + \psi_0)\big] = A_\mathrm{s}\,\mathrm{Re}\big\{\mathrm{e}^{\mathrm{i}\omega_\mathrm{s} t}\big(1 + m\,\mathrm{e}^{\mathrm{i}\psi(t)}\big)\big\}, \quad (20.1)$$

with the *beat phase*

$$\psi(t) = \Delta\omega\, t + \psi_0, \qquad \Delta\omega = \omega_\mathrm{e} - \omega_\mathrm{s} = 2\pi\Delta f, \qquad \Delta f = f_\mathrm{e} - f_\mathrm{s}. \quad (20.2)$$

The sign convention is fixed for the whole book: $\Delta f > 0$ means the neighbor's frequency is higher.

The complex factor $1 + m e^{i\psi}$ is the *envelope* of the mixture relative to the fish's own carrier. Its modulus is the local amplitude and its argument is the local phase shift:

$$A(t) = A_s \, |1 + m \, e^{i\psi(t)}| = A_s \sqrt{1 + m^2 + 2m \cos \psi(t)}, \tag{20.3}$$

$$\varphi(t) = \arg\bigl(1 + m \, e^{i\psi(t)}\bigr) = \operatorname{atan2}\bigl(m \sin \psi(t),\, 1 + m \cos \psi(t)\bigr). \tag{20.4}$$

The mixture is therefore a carrier at the fish's own frequency whose amplitude and phase are both modulated at the beat rate $|\Delta f|$. For small contamination, $A \approx A_s(1 + m \cos \psi)$ and $\varphi \approx m \sin \psi$: the two modulations are sinusoids in quadrature.

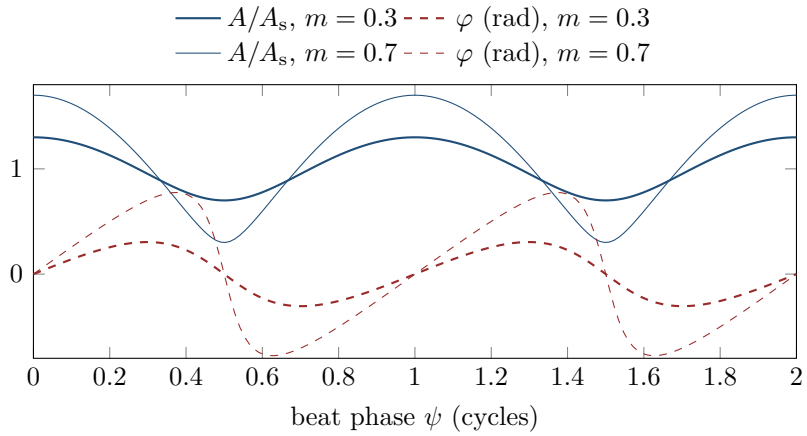


Figure 20.1: Amplitude and phase modulation of the local mixture (20.3)–(20.4) over two beat cycles, for a weakly contaminated patch ($m = 0.3$) and a strongly contaminated one ($m = 0.7$). Both modulations repeat at the beat rate $|\Delta f|$. At small m they are sinusoids a quarter cycle apart; at larger m the phase modulation becomes asymmetric.

The mixture is the physical relation between the two fishes, delivered to each patch of each fish's skin. In the terms of Chapter 3, it is a sum of two tones; in the terms of Chapter 4, the tuberous receptors that encode its amplitude and timing are the nonlinear readers that turn the implicit beat into explicit variables.

20.2 Why a Slow Beat Is a Problem

Electrolocation reads small, slow changes in the amplitude and timing of the carrier as the fish moves relative to objects (Chapter 19). A neighbor's field produces changes of the same two kinds. When the frequency difference is large, the beat is fast and can be separated from object-induced modulations by the temporal filtering of the electrosensory pathways. When the difference

is small, the beat falls into the same slow range as the object signals, and the two cannot be separated by filtering (Heiligenberg 1991a). The neighbor's field is then a jamming signal in the engineering sense: interference inside the band the receiver needs.

In the notation of Chapter 19, an object produces a slow modulation $\epsilon(\boldsymbol{x}, t)$ of the local amplitude as the fish swims. With a neighbor present, the patch receives

$$A(\boldsymbol{x}, t) \approx A_{\mathrm{s}}(\boldsymbol{x})\big[1 + \epsilon(\boldsymbol{x}, t) + m(\boldsymbol{x}) \cos \psi(t)\big], \tag{20.5}$$

to first order in both perturbations. If the spectrum of ϵ and the beat rate $|\Delta f|$ overlap, no linear filter applied to A can remove the second term without also damaging the first. The only remedy available to a receiver that cannot change its neighbor is to change the relation: to move $|\Delta f|$ out of the band.

20.3 The Jamming Avoidance Response

The behavior was discovered by Watanabe and Takeda (1963), who found that an alternating-current stimulus near the discharge frequency of *Eigenmannia* shifts the fish's own frequency. Bullock, Hamstra, et al. (1972a) and Bullock, Hamstra, et al. (1972b) named it the jamming avoidance response and characterized it in detail. The rule is simple to state: the fish raises its discharge frequency when its neighbor's is lower, and lowers it when its neighbor's is higher (Metzner 1993b). In our notation, the JAR drives f_{s} so as to increase $|\Delta f|$:

$$\Delta f > 0 \implies f_{\mathrm{s}} \downarrow, \qquad \Delta f < 0 \implies f_{\mathrm{s}} \uparrow . \tag{20.6}$$

Four facts about the response matter for the rest of the book.

It is computed from sensory input alone. The fish does not compare the neighbor's frequency with an internal copy of its own pacemaker command; the information is in the mixture on its skin (Heiligenberg 1991a; Heiligenberg 1991b). The experimental demonstration uses fish whose own discharge has been silenced pharmacologically and replaced by an electronic mimic whose frequency the experimenter controls, the "S1" of Heiligenberg's papers, with the neighbor's field supplied as a second signal, "S2". The pacemaker, still running but no longer driving the field on the skin, shifts in the direction dictated by the sign of the frequency difference between the two electronic signals. Chapter 22 examines why this matters.

It is distributed. Many patches of skin contribute to the decision, and the behavior emerges from the pooled activity of many imprecisely tuned neurons rather than from a small set of dedicated detectors; Heiligenberg called the arrangement a neuronal democracy (Heiligenberg, Baker, et al. 1978; Heiligenberg 1980; Heiligenberg and G. Rose 1985).

It is innate. A correct JAR appears in young *Eigenmannia* of 1.2 to 1.6 cm body length, at a developmental age of 24 to 32 days (Viete and Heiligenberg 1991).

It is not universal. *Sternopygus* is the only known gymnotiform genus that does not perform a JAR, and the neural mechanisms it uses for electrolocation suggest the substrate from which the response evolved in its relatives (G. Rose, C. Keller, et al. 1987). The African *Gymnarchus* performs a JAR with the same computational rules (Chapter 24).

The response is tuned to small differences. Neurons in the midbrain that encode the sign of Δf respond best to beat rates of roughly 4 to 8 Hz (G. Rose and Heiligenberg 1986b); the behavior itself is elicited when the neighbor's frequency is close to the fish's own, and fades as the difference grows (Bullock, Hamstra, et al. 1972a; Heiligenberg 1991a).

20.4 Changing a Relatum to Change a Relation

The JAR has a structure that none of the systems in Parts I to III had. A beat between two oscillators is a relation. In Parts I to III, relations were either read by an outside observer (the moiré fringes in Chapter 2) or entered the dynamics of the oscillators symmetrically (the relative phase of coupled oscillators in Chapter 9). Here one of the relata reads the relation and acts on itself to change it.

Definition 20.1 (Relatum-changing response)**.** Let a relation $r = \mathcal{R}(s_{\mathrm{A}}, s_{\mathrm{E}})$ hold between an agent and its environment. A *relatum-changing response* is a reading loop $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$ in which the policy acts only on the agent's own state, and does so as a function of $\mathsf{D}(\mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}}))$, such that the relation changes in a direction specified by the reading.

For the JAR, the agent state is the discharge frequency f_{s}, the environment state is the neighbor's frequency f_{e}, the relation is Δf, the reading (Chapter 21) extracts $\operatorname{sgn}(\Delta f)$, and the policy moves f_{s} in the opposite direction:

$$\begin{aligned}(f_{\mathrm{s}}, f_{\mathrm{e}}) &\xrightarrow{\mathsf{M}} y \xrightarrow{\mathsf{D}} \operatorname{sgn}(\Delta f) \\ &\xrightarrow{\pi} \dot{f}_{\mathrm{s}} = -\kappa \operatorname{sgn}(\Delta f)\, G(|\Delta f|) \xrightarrow{\mathsf{F}} f_{\mathrm{s}}'.\end{aligned} \tag{20.7}$$

Here $\kappa > 0$ sets the rate and $G \geq 0$ describes how the response depends on the size of the difference, positive for small $|\Delta f|$ and falling to zero for large ones. Both are illustrative placeholders for the measured tuning.

Proposition 20.2 (The coincidence of frequencies is repelling)**.** *Suppose the neighbor's frequency f_{e} is constant and the fish obeys (20.7) with $G(u) > 0$ for $0 < u < u_{\max}$. Then along any solution with $0 < |\Delta f| < u_{\max}$, $|\Delta f|$ strictly increases.*

Proof. With f_{e} constant, $\dot{\Delta f} = -\dot{f}_{\mathrm{s}} = \kappa \operatorname{sgn}(\Delta f)\, G(|\Delta f|)$. Then $\frac{\mathrm{d}}{\mathrm{d}t}|\Delta f| = \operatorname{sgn}(\Delta f)\, \dot{\Delta f} = \kappa\, G(|\Delta f|) > 0$ whenever $0 < |\Delta f| < u_{\max}$. □

The proposition is a statement about the model; its fit to the animal is the subject of the closed-loop experiments below. Its content is that the state of equal frequencies, $\Delta f = 0$, is an unstable point of the relation: the fish's behavior makes the relation run away from coincidence until it leaves the response range. When two fish both perform the response, each moves away from the other, and the frequency difference grows from both ends.

Example 20.3 (The animal as participant in its own measurement). An instrument that measures a frequency difference normally stands outside the two sources. The fish is one of the sources. Its measurement of Δf is taken from a mixture it helped produce, and its response changes one of the quantities being measured. The loop (20.7) is closed: the next mixture on the skin depends on the last reading. In the language of Chapter 1, the fish is a reader inside the relation it reads, and the reader subsequently modifies the relation.

20.5 The Loop Closed and Opened in the Laboratory

Under natural conditions the JAR runs in closed loop: the fish's frequency shift changes the beat, which changes the input, which changes the next shift. Experimenters can open the loop, by holding the mimic's and the neighbor's frequencies fixed, or close it artificially, by making the stimulus frequency follow the fish's own.

Madhav et al. (2013) did the second. They clamped the jamming stimulus to track the fish's own discharge frequency with a set offset, so that the frequency difference the fish experienced could be held at a chosen value regardless of what the fish did. In this arrangement the JAR could be stabilized, and its dynamics measured near what is normally an unstable equilibrium. Near the equilibrium, a delayed first-order linear model fitted the behavior. Farther away, the dynamics were nonlinear, and a model with competing motor and sensory components predicted a "snap-through" bifurcation as a function of the clamped difference, which the experiments then confirmed. The study gives a dynamical description of the JAR spanning its local linear and global nonlinear regimes, and it is a model for the kind of experiment Part XIII proposes: one that changes the loop's structure while holding its parts fixed (Chapter 59).

20.6 Participation Without a Participant

The JAR invites a description in agential language: the fish "avoids" jamming, "decides" which way to shift, "knows" whether its neighbor is higher. Each of these words has a precise surrogate in the loop (20.7). "Avoids" means that the closed-loop dynamics make coincidence repelling (Proposition 20.2).

"Decides" means that a population of neurons, pooled across the body surface, drives two separate motor pathways whose balance sets the direction of change (Chapter 21). "Knows" means that a variable correlated with $\mathrm{sgn}(\Delta f)$ is explicit in the fish's brain and causally controls the frequency shift.

The surrogates are exact and they leave something unaddressed. An adaptive radio transmitter that shifts its channel away from interference satisfies every one of them, and nobody calls the radio a participant in anything (Chapter 48). The fish differs from the radio in ways that later parts of the book examine: it built its own reader (Chapter 26), its reading can be wrong in a way that matters to it (Chapter 23), and its reading loop is part of an organism that maintains itself (Chapter 28). None of those differences is visible in the loop (20.7) itself.

20.7 Audit

Canonical phenomenon

Two wave-type electric fish with nearby discharge frequencies shift their frequencies apart, each in the direction that increases the difference.

Minimal model

The mixture (20.1) with amplitude and phase modulations (20.3)–(20.4); the JAR rule (20.6); the loop (20.7) and Proposition 20.2.

What is relational here?

The frequency difference $\Delta f = f_\mathrm{e} - f_\mathrm{s}$ and the beat it produces in the amplitude and phase of the local field; neither discharge contains it.

What physically reads it?

Tuberous electroreceptors encode the modulations; midbrain circuits extract the sign of Δf (Chapter 21); prepacemaker and pacemaker nuclei convert it into a frequency change.

What existing field explains it?

Neuroethology of electric fish (Bullock, Hamstra, et al. 1972a; Heiligenberg 1991a), the physics of beats (Chapter 3), and feedback control. The closed-loop dynamics have been measured and modeled (Madhav et al. 2013).

What remains unexplained?

None in the mechanism. The chapter identifies the structure the rest of Part V examines: a reader that is one of the relata and acts on itself to change the relation it reads.

A discriminating experiment

Clamp the stimulus to follow the fish at a fixed offset and vary the offset through zero (Madhav et al. 2013). The loop model predicts an unstable equilibrium at $\Delta f = 0$ and a bifurcation structure set by the sensory tuning

and the motor dynamics; a model in which the fish tracks absolute frequency rather than the relation predicts neither.

20.8 Notes and References

The JAR was discovered by Watanabe and Takeda (1963) and named and analyzed by Bullock, Hamstra, et al. (1972a) and Bullock, Hamstra, et al. (1972b). Heiligenberg, Baker, et al. (1978) and Heiligenberg (1980) established that the response is computed by a distributed evaluation of electroreceptive input across the body surface; Heiligenberg (1991a) and Heiligenberg (1991b) synthesize the program that followed. Viete and Heiligenberg (1991) show the response is innate, G. Rose, C. Keller, et al. (1987) discuss *Sternopygus*, the gymnotiform that lacks it, and Stamper, Carrera-G, et al. (2010) relate interference to group size across species. The closed-loop analysis is Madhav et al. (2013); Cowan et al. (2014) discuss feedback control as a general framework for such biological loops.

20.9 Exercises

Exercise 20.1. Derive (20.3) and (20.4) from (20.1). Show that A is maximal when $\psi = 0$ and minimal when $\psi = \pi$, and that the modulation depth $(A_{\max} - A_{\min})/(A_{\max} + A_{\min})$ equals m.

Exercise 20.2. Expand A and φ to second order in m. Show that at first order the two modulations are sinusoids a quarter cycle apart, and identify the second-order terms that make φ asymmetric in Figure 20.1.

Exercise 20.3. Suppose object-induced modulations $\epsilon(t)$ have power concentrated below 5 Hz (an illustrative figure). For beat rates of 2, 10, and 50 Hz, sketch the spectrum of (20.5) and explain which beats a low-pass filter could separate from ϵ. What does this imply about which frequency differences a JAR should respond to?

Exercise 20.4. Two fish obey (20.7) with the same κ and G. Write the coupled equations for f_1 and f_2, show that their mean frequency is conserved, and find how $|f_1 - f_2|$ evolves. What changes if the two fish have different κ?

Exercise 20.5. In a clamp experiment the stimulus is set to $f_{\mathrm{e}}(t) = f_{\mathrm{s}}(t) + d$ for a fixed offset d. Using (20.7), show that the fish's frequency then drifts at a constant rate determined by d, and explain how measuring that rate for many values of d recovers the function $\operatorname{sgn}(d)\, G(|d|)$. What additional dynamics would reveal a delay in the loop?

Exercise 20.6. Write a loop of the form (20.7) for a transmitter that hops away from an interfering channel. Identify each term. Which of the four facts listed in Section 20.3 does the radio share, and which has no meaning for it?

21

Reading the Sign

To avoid jamming, a fish must shift its frequency in the right direction: down if its neighbor is higher, up if its neighbor is lower (Chapter 20). The beat rate tells it how far apart the two frequencies are. It does not tell it which one is higher. The magnitude $|\Delta f|$ is written plainly in the mixture on the skin; the sign of Δf is not written in any single feature of it.

This chapter shows where the sign is, and how the fish reads it. The thesis has a mathematical half and a biological half. Mathematically, the amplitude modulation of the mixture carries no information about the sign at all, the phase modulation carries it only through small distortions of its waveform, and the *joint* behavior of amplitude and differential phase carries it exactly and at every instant. Biologically, the fish's brain reads that joint behavior: amplitude and timing travel along separate pathways from the skin to the midbrain, where neurons that respond to one combination of amplitude change and phase difference and not to the other give a sign-selective output, and two separate motor pathways convert it into a rise or a fall of the discharge frequency. The sign of a relation is read from a relation between two other relations, by circuits that have been traced from receptor to pacemaker.

21.1 Why the Magnitude Is Not Enough

Recall the local mixture at a patch of skin with contamination ratio m (Chapter 20):

$$A(t) = A_{\mathrm{s}}\,|1 + m\,\mathrm{e}^{\mathrm{i}\psi(t)}|, \qquad \varphi(t) = \arg\!\left(1 + m\,\mathrm{e}^{\mathrm{i}\psi(t)}\right), \tag{21.1}$$

with beat phase $\psi(t) = \Delta\omega\, t + \psi_0$, $\Delta\omega = 2\pi\Delta f$, and $0 \le m < 1$. Both modulations are periodic with period $1/|\Delta f|$, so the beat rate, and with it $|\Delta f|$, can be read from either. The sign is another matter.

Theorem 21.1 (The amplitude is blind to the sign)**.** *For every contamination ratio $0 \le m < 1$, every initial beat phase ψ_0, and every $\Delta\omega$, the amplitude waveform produced by $(\Delta\omega, \psi_0)$ is identical to the one produced by $(-\Delta\omega, -\psi_0)$.*

Proof. $A = A_{\mathrm{s}}\sqrt{1 + m^2 + 2m\cos\psi}$ depends on ψ only through $\cos\psi$, which is even. Replacing $(\Delta\omega, \psi_0)$ by $(-\Delta\omega, -\psi_0)$ replaces $\psi(t)$ by $-\psi(t)$ and leaves $A(t)$ unchanged for all t. □

No reader of the amplitude modulation, however sophisticated, can recover the sign of Δf from it: two worlds that differ in the sign produce the same amplitude record. The theorem is the electrosensory version of a general fact about real signals. A real modulation at rate $|\Delta f|$ has a spectrum symmetric about zero frequency, and the sign of a frequency is invisible in a symmetric spectrum.

The phase modulation is almost blind. Expanding the logarithm of the envelope gives, for $m < 1$,

$$\varphi = \operatorname{Im}\log\bigl(1 + m\mathrm{e}^{\mathrm{i}\psi}\bigr) = m\sin\psi - \frac{m^2}{2}\sin 2\psi + \frac{m^3}{3}\sin 3\psi - \cdots . \tag{21.2}$$

Proposition 21.2 (The phase is blind to the sign at first order)**.** *To first order in m, the phase waveform produced by $(\Delta\omega, \psi_0)$ is identical to the one produced by $(-\Delta\omega, \pi - \psi_0)$. At second order the two differ.*

Proof. The substitution maps $\psi(t)$ to $\pi - \psi(t)$. Then $\sin(\pi - \psi) = \sin\psi$, so the first-order term is unchanged, while $\sin(2\pi - 2\psi) = -\sin 2\psi$, so the second-order term changes sign. □

At weak contamination the phase modulation is a sinusoid, and a sinusoid does not reveal the direction in which its phase advances. At strong contamination the waveform becomes asymmetric (Figure 20.1), and the sign leaks into its shape. That leak is small where jamming usually operates, and in any case the fish has no clock against which to measure the absolute phase φ. It has no internal clock against which to time the signal on its skin (Section 21.6); what it can measure is timing *between* patches of skin.

21.2 Amplitude and Differential Phase

The fish's skin is not a single patch. A neighbor's field reaches different parts of the body with different strengths, so different patches have different contamination ratios, and their modulations differ in depth. Two families of receptors report on each patch (Chapter 19): P-units encode the local amplitude, and T-units fire one spike per cycle at a fixed phase of the local signal, so the relative timing of T-unit spikes from two patches reports their *differential phase* (Scheich et al. 1973; Heiligenberg 1991a).

For two patches with contamination ratios $m_1 < m_2$, define the differential phase

$$\delta(t) = \varphi_2(t) - \varphi_1(t), \tag{21.3}$$

the phase of the more contaminated patch relative to the less contaminated one. Because both patches carry the same fish's carrier, any drift or jitter common to both cancels in δ. The differential phase is frame-independent in the sense of Chapter 1: it does not depend on where the phase origin is placed, and by Theorem 1.5 a reader with no access to that origin can use only variables of this kind. Its reliance on differential phase is also why the jamming avoidance response of *Gymnarchus* survives a large artificial jitter imposed on the phase of the whole stimulus (Y.-X. Guo and Kawasaki 1997).

At small contamination the two variables are

$$A_2 \approx A_{\mathrm{s}}(1 + m_2 \cos\psi), \qquad \delta \approx (m_2 - m_1)\sin\psi. \tag{21.4}$$

They are sinusoids at the beat rate, a quarter cycle apart, and the sign of Δf decides which one leads. The following theorem shows that the relation holds exactly, at every contamination, and at every instant.

21.3 The Sign Is in the Joint Behavior

Theorem 21.3 (Joint modulation determines the sign)**.** *Let two patches of skin have contamination ratios $0 < m_1 < m_2 < 1$ and a common beat phase $\psi(t) = \Delta\omega\, t + \psi_0$ with $\Delta\omega \neq 0$. Let A_j be the amplitude at patch j and $\delta = \varphi_2 - \varphi_1$ the differential phase. Then at every instant at which $\sin\psi(t) \neq 0$,*

$$\operatorname{sgn}\big(\dot{A}_j(t)\,\delta(t)\big) = -\operatorname{sgn}(\Delta f), \qquad j = 1, 2. \tag{21.5}$$

Proof. First, the sign of δ. For fixed ψ, differentiate the phase with respect to the contamination ratio:

$$\frac{\partial}{\partial m}\arg\!\left(1 + m\mathrm{e}^{\mathrm{i}\psi}\right) = \operatorname{Im}\frac{\mathrm{e}^{\mathrm{i}\psi}}{1 + m\mathrm{e}^{\mathrm{i}\psi}} = \frac{\sin\psi}{|1 + m\mathrm{e}^{\mathrm{i}\psi}|^2}.$$

For $\sin\psi > 0$ the phase strictly increases with m, so $\delta = \varphi(\psi; m_2) - \varphi(\psi; m_1) > 0$; for $\sin\psi < 0$ it strictly decreases, so $\delta < 0$. Hence $\operatorname{sgn}\delta = \operatorname{sgn}\sin\psi$.

Second, the sign of $\dot{A}_j$. From $A_j^2 = A_{\mathrm{s}}^2(1 + m_j^2 + 2m_j\cos\psi)$ we get $\mathrm{d}A_j/\mathrm{d}\psi = -A_{\mathrm{s}}^2 m_j \sin\psi / A_j$, so $\operatorname{sgn}(\mathrm{d}A_j/\mathrm{d}\psi) = -\operatorname{sgn}\sin\psi$, and $\dot{A}_j = \Delta\omega\,\mathrm{d}A_j/\mathrm{d}\psi$.

Combining, $\operatorname{sgn}(\dot{A}_j\,\delta) = \operatorname{sgn}(\Delta\omega)\cdot(-\operatorname{sgn}\sin\psi)\cdot\operatorname{sgn}\sin\psi = -\operatorname{sgn}(\Delta\omega)$, and $\operatorname{sgn}(\Delta\omega) = \operatorname{sgn}(\Delta f)$. □

In words: *when the amplitude is rising while the more contaminated patch leads the less contaminated one, the neighbor is lower* ($\Delta f < 0$) and the fish should raise its frequency; *when the amplitude is rising while the more contaminated patch lags, the neighbor is higher* ($\Delta f > 0$) and the fish should lower it. Heiligenberg and Rose stated the behavioral rule in these terms, for a negative Δf, with the amplitude read in a region A and phase compared with a region B (Heiligenberg and G. Rose 1986); in our model their statement holds with A the less contaminated and B the more contaminated region.

Three features of the theorem matter for the rest of the book.

The rule is pointwise. It does not require averaging over a beat cycle: except at the two instants per cycle where $\sin\psi = 0$, every moment of the joint signal carries the sign. A population of receptors sampling many moments and many pairs of patches can only strengthen the evidence.

The rule is invariant. It uses the direction of amplitude change and the sign of a phase difference, so it is unchanged by the overall strength of either discharge, by the polarity of the neighbor's field at a patch (reversing the polarity shifts ψ by π and leaves the theorem intact), and by any timing drift common to both patches.

The rule needs an ordering of patches. For a pair of patches a, b in either order, (21.5) generalizes to $\operatorname{sgn}(\dot{A}_a\,(\varphi_a - \varphi_b)) = -\operatorname{sgn}(\Delta f)\,\operatorname{sgn}(m_a - m_b)$. The information needed to orient each comparison is itself in the mixture, since the more contaminated patch is the one with the deeper amplitude modulation. When many pairs vote and are pooled, as they are in the fish (Section 21.5), a comparison with the wrong orientation is outvoted.

Corollary 21.4 (Sense of rotation)**.** *Plot the differential phase δ on the horizontal axis and the amplitude A_2 on the vertical axis. As time runs, the point $(\delta(t), A_2(t))$ traces a closed curve once per beat cycle, clockwise if $\Delta f > 0$ and counterclockwise if $\Delta f < 0$.*

Proof. On the right half of the curve ($\delta > 0$), Theorem 21.3 gives $\dot{A}_2 < 0$ when $\Delta f > 0$: the point moves down on the right and, by the same argument, up on the left. That is clockwise motion; reversing the sign of Δf reverses every velocity. □

Figure 21.1 makes the point visually. The two panels contain the same closed curve. Every value of amplitude and every value of differential phase occurs in both; the only difference is the order in which they occur. The variable the fish needs is the orientation of a trajectory in a two-dimensional space of relational variables, a relation among relations. G. Rose and Heiligenberg (1986b) called it "the sense of rotation in an amplitude-phase plane" and found neurons that read it.

21.4 The Neural Sign Computation

The circuit that reads the sign in *Eigenmannia* has been mapped with unusual completeness (Heiligenberg 1991a). Figure 21.2 summarizes it.

Receptors and hindbrain. P-units and T-units project along separate pathways to the electrosensory lateral line lobe (ELL) of the hindbrain, which relays both to the dorsal torus semicircularis of the midbrain (Scheich et al. 1973; Metzner and Heiligenberg 1991).

Differential phase. Neurons in the torus encode phase differences between sinusoidal electric stimuli (Bastian and Heiligenberg 1980). Differential phase

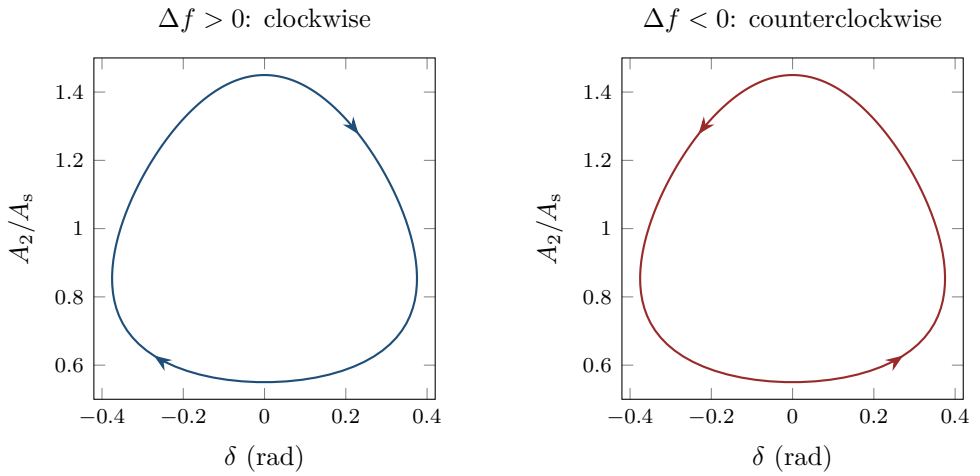


Figure 21.1: The joint trajectory of differential phase and amplitude over one beat cycle, for patches with contamination ratios $m_1 = 0.1$ and $m_2 = 0.45$ (illustrative values). The two curves are the same set of points traversed in opposite directions: neither coordinate alone distinguishes them, and the sense of rotation does (Corollary 21.4).

sensitivity originates in the small cells of lamina 6, the exclusive termination site of the phase-coding pathway (G. Rose and Heiligenberg 1985a). The circuit compares the timing of signals from different parts of the body surface and detects differences as small as 400 ns at the level of behavior (G. Rose and Heiligenberg 1985b; C. Carr et al. 1986). Individual differential-phase neurons in the torus are far less acute, with thresholds from 6.5 µs to several hundred microseconds (G. Rose and Heiligenberg 1986a). Acuity sharpens along the pathway by convergence: single neurons of the prepacemaker nucleus respond to disparities as small as 1 µs, a result of a high degree of spatial convergence of afferent inputs (Kawasaki, G. Rose, et al. 1988).

The combination. Deeper laminae of the torus contain neurons that respond to both amplitude and phase modulation, and some are selective for the sign of Δf (Heiligenberg and G. Rose 1985). These sign-selective units read the sense of rotation in the amplitude–phase plane and are also sensitive to the beat rate, most of them responding best to rates of about 4 to 8 Hz (G. Rose and Heiligenberg 1986b). Amplitude information gates phase information in these neurons (Heiligenberg and G. Rose 1986); the gating is a neural counterpart of the product in (21.5). No single neuron needs to be precise. The behavior emerges from a large population of imprecisely tuned neurons, and the system has no critical bottleneck of dedicated detectors (Heiligenberg and G. Rose 1985).

From sign to motor command. The nucleus electrosensorius of the diencephalon contains finely tuned filters for the JAR; stimulating its “increase” subdivision produces slow rises of the discharge frequency, and stimulating

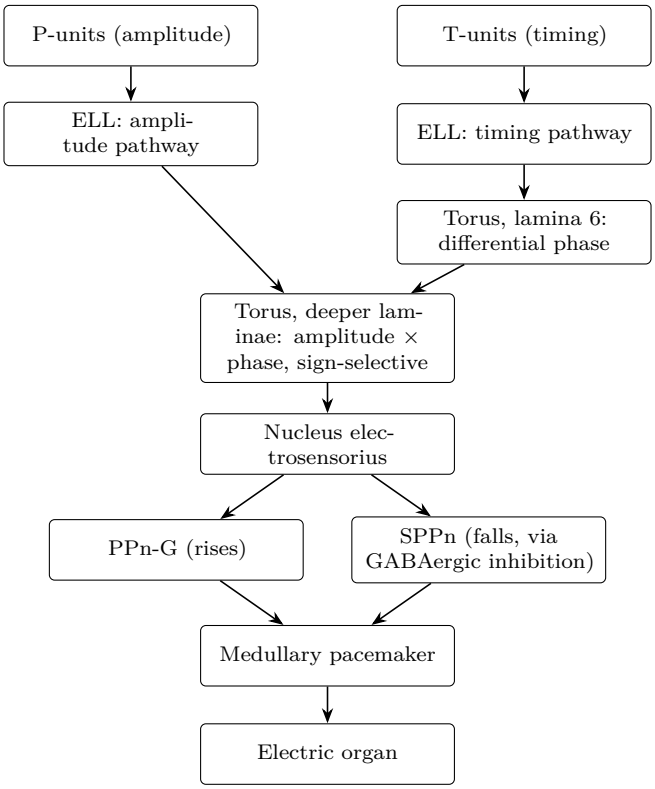


Figure 21.2: The jamming avoidance pathway of *Eigenmannia*, simplified from Heiligenberg (1991a), C. H. Keller et al. (1990), and Metzner (1993b). Amplitude and timing travel separately to the midbrain, where they are combined; two motor pathways, one for rises and one for falls, converge on the pacemaker.

its "decrease" subdivision produces slow falls (C. H. Keller et al. 1990). Metzner (1993b), using selective lesions and pharmacology, showed that the two directions are controlled by separate motor pathways. Rises run from the dorsal nucleus electrosensorius through the diencephalic prepacemaker nucleus (PPn-G) to the pacemaker, with AMPA-type receptors mediating the final synapse. Falls run from the ventral nucleus electrosensorius as GABAergic inhibition of the sublemniscal prepacemaker nucleus (SPPn), which is tonically active and drives the pacemaker through NMDA-type receptors; inhibiting it lowers the frequency. Metzner concluded that there is no single population of final "recognition units" for the sign; the motor control of the JAR is a push-pull interaction of two independent pathways. Glutamate applied to PPn-G produces slow, gradual accelerations of the discharge, while a neighboring subdivision (PPn-C) produces the fast, abrupt modulations used in communication (Kawasaki, Maler, et al. 1988).

21.5 A Neuronal Democracy

Theorem 21.3 is a statement about one pair of patches. The fish has thousands of receptors over its whole body, and the neighbor's field contaminates them unevenly. Early behavioral work showed that the fish evaluates the whole body surface: each region contributes according to the modulations it experiences, and the response follows the balance of the contributions (Heiligenberg, Baker, et al. 1978; Heiligenberg 1980). The physiology of the torus matches: sign selectivity is spread across many neurons of modest precision (Heiligenberg and G. Rose 1985).

The model explains why pooling works. Every correctly oriented pair of patches, at every instant, votes the same way (Theorem 21.3). Noise and occasional mis-oriented comparisons add scatter to the vote without biasing its sign, so the reliability of the population's decision grows with the number of votes. The same arithmetic explains how a behavioral acuity of 400 ns can rest on neurons whose individual thresholds are microseconds: convergence of many independent, imprecise estimates of a shared variable reduces the error of their combination.

Example 21.5 (A vote that cannot be cast by one region). Suppose the neighbor's field were uniform over the whole fish, so that every patch had the same contamination ratio m. Then $\delta \equiv 0$ for every pair, and by Theorem 21.1 the amplitude alone is uninformative. The model predicts that the fish could detect the beat but could not determine its sign. The prediction is the reason geometry matters to the JAR: the sign becomes available only because the neighbor's field falls off across the body, so that different patches are differently contaminated. The relation between the fish and its neighbor is readable only through a relation between parts of the fish.

21.6 No Internal Reference Is Needed

A tempting design for a jamming avoidance circuit would compare the neighbor's frequency with an internal copy of the fish's own pacemaker command, an efference copy in the sense of Chapter 22. The fish does not use one. In preparations where the fish's own discharge is silenced and replaced by an electronic mimic whose frequency is independent of the pacemaker, the pacemaker still shifts in the direction dictated by the sign of the frequency difference between the mimic and the jamming signal (Heiligenberg 1991a). *Gymnarchus* also determines the direction of its shift from the signal mixture alone, without reference to its pacemaker command (Kawasaki 1993), and its pacemaker sends no collaterals to sensory areas that could carry such a reference (Kawasaki 1994).

Theorem 21.3 explains why none is needed. The fish's own carrier is present in the mixture at every patch, so the mixture already contains the reference

that an efference copy would supply. The sign is carried by relations internal to the mixture (the direction of amplitude change and the phase difference between patches), and those relations are measurable without knowing the carrier's absolute timing. The information an internal copy would provide is already present in the relation.

21.7 The Counterfactual Test

Chapter 17 asks what distinguishes a physical system that implements a computation from one that can merely be described as doing so. The standard answer requires that the system's causal structure support the right counterfactuals: had the input been different in the relevant way, the system's state transitions would have differed correspondingly. The jamming avoidance circuit meets that standard for the relation it reads:

$$\Delta f > 0 \implies \text{frequency falls}, \qquad \Delta f < 0 \implies \text{frequency rises}. \tag{21.6}$$

Change the sign of the frequency difference and the neural and behavioral trajectories change systematically: different torus neurons fire, a different motor pathway is recruited, and the pacemaker moves the other way. The experimenter does not decide which distinction matters. The distinction has been built into the causal architecture of the animal, and the animal's own behavior is the readout.

Sign of Δf	Amplitude rising while	Sense of rotation	Response
$\Delta f > 0$ (neighbor higher)	more contaminated patch lags	clockwise	lower f_{s}
$\Delta f < 0$ (neighbor lower)	more contaminated patch leads	counterclockwise	raise f_{s}

Table 21.1: The relational computation of the jamming avoidance response. Axes as in Figure 21.1.

This is a genuinely implemented relational computation. The input is a relation between two discharges; the computed variable is the sign of that relation; the variable is carried by a higher-order relation between two modulations; and the output acts on one of the original relata. Whether it is anything more than a well-engineered control loop is the question Parts VI to X take up. Chapter 23 shows the first place where more vocabulary becomes necessary: the circuit can be fooled.

21.8 Audit

Canonical phenomenon

Eigenmannia shifts its discharge frequency down when a neighbor is higher

and up when a neighbor is lower, without an internal reference to its own motor command.

Minimal model
Amplitude and phase of the envelope $1+me^{i\psi}$ at patches of differing contamination; Theorem 21.1, Proposition 21.2, Theorem 21.3, and Corollary 21.4.

What is relational here?
The sign of Δf, carried by the relative order of two modulations (amplitude change and differential phase), each of which is itself a relation between the fish's field and its neighbor's.

What physically reads it?
T-unit timing compared across body regions in lamina 6 of the torus; amplitude from P-units; sign-selective neurons in deeper torus laminae; filtering in the nucleus electrosensorius; push-pull motor pathways through PPn-G and SPPn to the pacemaker.

What existing field explains it?
Neuroethology (Heiligenberg 1991a; Heiligenberg and G. Rose 1986; G. Rose and Heiligenberg 1986b; Metzner 1993b) explains the mechanism completely at the level used here; signal theory explains why the amplitude alone cannot carry the sign.

What remains unexplained?
None in the mechanism. What the chapter adds is a precise statement of where the sign lives (exactly in the joint modulation, pointwise) and of why convergence on this algorithm is forced by the physics of the mixture, a point Chapter 24 develops.

A discriminating experiment
Deliver amplitude and differential-phase modulations independently, with their relative order set by the experimenter and no second discharge present. If the fish's frequency follows the sense of rotation of the joint modulation rather than the presence of an actual neighbor, the circuit reads the relation of Theorem 21.3. Chapter 23 reports what happens when such combinations occur without a neighbor (Carlson and Kawasaki 2007).

21.9 Notes and References

Heiligenberg (1991a) is the definitive account of the computational rules of the JAR and their neuronal implementation; Heiligenberg (1991b) is a shorter review. The receptor classes are described by Scheich et al. (1973). Phase-sensitive midbrain neurons were first reported by Bastian and Heiligenberg (1980); the structure of the torus circuitry by G. Rose and Heiligenberg (1985a); the joint phase and amplitude computations by Heiligenberg and

G. Rose (1985) and Heiligenberg and G. Rose (1986); and the neurons that read the sense of rotation by G. Rose and Heiligenberg (1986b). Temporal hyperacuity at the behavioral level is reported by G. Rose and Heiligenberg (1985b) and C. Carr et al. (1986), the limits of single torus neurons by G. Rose and Heiligenberg (1986a), and single-neuron hyperacuity in the prepacemaker nucleus by Kawasaki, G. Rose, et al. (1988). The diencephalic sensory-motor interface is described by C. H. Keller et al. (1990), the push-pull motor control by Metzner (1993b), and the prepacemaker subdivisions by Kawasaki, Maler, et al. (1988). The equivalent analysis for *Gymnarchus* is due to Kawasaki (1993) and is taken up in Chapter 24.

21.10 Exercises

Exercise 21.1. Show that $\log(A/A_{\mathrm{s}}) = m\cos\psi - \frac{m^2}{2}\cos 2\psi + \frac{m^3}{3}\cos 3\psi - \cdots$ and use the series to give a second proof of Theorem 21.1.

Exercise 21.2. Using (21.2), compute the waveform of φ to second order for $(\Delta\omega, \psi_0)$ and for $(-\Delta\omega, \pi - \psi_0)$. Sketch both for $m = 0.5$. How could a reader with an absolute phase reference exploit the difference, and why is this of little use to the fish?

Exercise 21.3. Prove the generalization stated after Theorem 21.3: for any two patches a, b with $m_a \neq m_b$, $\operatorname{sgn}(\dot{A}_a(\varphi_a - \varphi_b)) = -\operatorname{sgn}(\Delta f)\operatorname{sgn}(m_a - m_b)$. Show that the modulation depth of A_a determines $\operatorname{sgn}(m_a - m_b)$, so that each comparison can be oriented from the mixture alone.

Exercise 21.4. Simulate N patches with contamination ratios drawn uniformly from $(0, 0.5)$, add independent Gaussian noise to each amplitude and phase sample, and implement a population vote: each pair contributes $\operatorname{sgn}(\dot{A}_a(\varphi_a - \varphi_b))\operatorname{sgn}(m_a - m_b)$. Plot the probability of a correct decision against N and against the noise level. Relate the result to the gap between single-neuron and behavioral acuity.

Exercise 21.5. Verify the claim of Example 21.5 that a uniform neighbor field leaves the sign undetermined. Then let the contamination vary linearly along the body, $m(x) = m_0 + gx$, and show that the information available to the fish grows with the gradient g.

Exercise 21.6. The signed area enclosed by the curve $(\delta(t), A_2(t))$ over one beat cycle, computed as $\frac{1}{2}\oint(\delta\,\mathrm{d}A_2 - A_2\,\mathrm{d}\delta)$, is negative for $\Delta f > 0$ and positive for $\Delta f < 0$. Prove this for small m and interpret the signed area as a readout that a neural circuit could compute.

Exercise 21.7. Design an electronic circuit that implements (21.5) for two input signals: specify how it extracts amplitude, differential phase, and the product, and how it drives a voltage-controlled oscillator. List the counterfactuals your circuit satisfies. Does it implement the same computation as the fish, in the sense of Chapter 17?

22

Reafference

Every animal that moves changes what it senses. Turning the head sweeps the visual scene across the retina; touching an object presses the skin; emitting a call returns an echo. The sensory input of a behaving animal is therefore a mixture of two components: the part caused by the animal's own action and the part caused by the world. In 1950 two papers gave this fact a name and a mechanism. Erich von Holst and Horst Mittelstaedt called the self-caused part *reafference* and the world-caused part *exafference*, and proposed that animals separate them by comparing input with an internal copy of their own motor commands (Holst and Mittelstaedt 1950). Roger Sperry, working independently, called the internal signal a *corollary discharge* (Sperry 1950).

This chapter places the electric fish in that tradition, and the fish shows both halves of it. Pulse-type electric fish subtract a learned prediction of their own reafference, and the prediction is carried by a corollary discharge of the command that fires each pulse. The wave-type jamming avoidance response needs no subtraction and no internal copy, because the fish's own field is physically present in the mixture it reads and serves as the reference directly. The thesis is that both routes produce the same kind of variable, one that measures the world relative to the animal's own activity, and that the second route produces it without any internal model of the self. That variable is what Chapter 19 called an agent-centered measurement frame, and the established name for its study is the study of reafference.

22.1 The Reafference Principle

Write the input to a sensory surface as the sum of a self-caused and a world-caused part,

$$s = \rho(a) + x(w), \tag{22.1}$$

where a is the animal's motor command, $\rho(a)$ the input that command produces through the body and the environment (the reafference), and $x(w)$ the input produced by events w in the world (the exafference).

Definition 22.1 (Efference copy and corollary discharge)**.** An *efference copy* is an internal signal, derived from the motor command a, that is used to predict the reafference $\rho(a)$. A *corollary discharge* is any internal signal derived from a motor command and sent to sensory or other non-motor structures; an efference copy is a corollary discharge used for this particular comparison (Crapse and Sommer 2008).

Model 22.2 (Reafference cancellation)**.** An animal with an internal predictor $\hat{\rho}$ computes

$$\hat{x} = s - \hat{\rho}(a) = x(w) + \big(\rho(a) - \hat{\rho}(a)\big). \tag{22.2}$$

If the predictor is accurate, $\hat{x}$ recovers the exafference; any error in the prediction appears as an apparent event in the world.

Von Holst and Mittelstaedt's demonstrations exploited the second clause. If the relation between command and reafference is surgically reversed, the prediction has the wrong sign, and self-caused input is misread as a world-caused event of twice the size. In one of their demonstrations, flies whose heads had been rotated through 180° walked in circles (Holst and Mittelstaedt 1950). Sperry's fish with surgically inverted eyes circled in the same way: with the retinal image inverted, the visual motion produced by each turn had the opposite sign from the one the corollary discharge predicted, the discrepancy drove a further turn, and the loop ran away (Sperry 1950).

The runaway is a statement about signs in a feedback loop, and it has a direct counterpart in Chapter 21. A reader that combines two signals to extract the sign of a relation fails catastrophically when one of the signals is inverted: the jamming avoidance circuit, fed a relation with the wrong orientation, would move the fish's frequency toward its neighbor's instead of away.

22.2 Reafference Across the Animal Kingdom

The distinction between self-caused and world-caused input is now a general principle of sensorimotor neuroscience (Crapse and Sommer 2008). Corollary discharges suppress or reshape sensory responses in singing crickets and in primates making saccadic eye movements, among many other cases. In humans, the attenuation of self-produced sensation shows up in the familiar inability to tickle oneself: a tactile stimulus produced by one's own movement is perceived as less ticklish, and evokes less activity in somatosensory cortex, than an identical stimulus produced externally (Blakemore et al. 1998). Model 22.2 predicts the specificity of the effect: cancellation depends on the prediction matching the actual consequence of the movement, so a stimulus that departs from the predicted one should be cancelled less.

22.3 Negative Images in Pulse-Type Fish

Pulse-type electric fish face reafference in its starkest form. Each discharge of the electric organ produces a strong stimulus on the fish's own electroreceptors, far stronger than the small distortions caused by prey or objects, and the discharges are emitted irregularly, at intervals the fish controls. In mormyrids, the command that fires each discharge also sends a corollary discharge to the electrosensory lateral line lobe (ELL), a cerebellum-like structure (C. C. Bell 1981; C. Bell et al. 1997).

The ELL uses the corollary discharge to cancel predictable reafference. Principal cells receive the electrosensory input and, through a parallel-fiber system, signals time-locked to the discharge command. The command-locked input learns to form a *negative image* of the sensory response that normally follows each discharge; added to the actual response, the negative image cancels the predictable part and leaves the unexpected part. The image is plastic. If an artificial stimulus is repeatedly paired with the command, a negative image of that stimulus develops, and when the stimulus is then withdrawn, the negative image remains for a while as a response to the command alone: Bell's "efference copy which is modified by reafferent input" (C. C. Bell 1981; Requarth and Sawtell 2011).

Model 22.3 (An adaptive negative image)**.** Let $s_n(\tau)$ be the sensory response at time τ after the n-th discharge, and let the command-locked prediction be $\hat{\rho}_n(\tau)$. The cell's output is $u_n(\tau) = s_n(\tau) - \hat{\rho}_n(\tau)$, and the prediction is updated after each discharge by

$$\hat{\rho}_{n+1}(\tau) = \hat{\rho}_n(\tau) + \eta\, u_n(\tau), \qquad 0 < \eta < 1. \tag{22.3}$$

If the reafferent response is the same after every discharge, $s_n = \rho + x_n$ with x_n zero-mean, then $\hat{\rho}_n \to \rho$ in expectation and $u_n \to x_n$.

The model is a least-mean-squares filter; it is illustrative, and the cellular mechanisms (plasticity at the synapses carrying the command-locked signals) are reviewed by C. Bell et al. (1997) and Requarth and Sawtell (2011). Its structure is what matters here. The fish carries an internal model of the sensory consequences of its own action, learns it from experience, and uses it to compute the world relative to the self by subtraction.

22.4 Reafference as Carrier

Active electrolocation in wave-type fishes uses self-generated input differently. The continuous discharge is reafference, and it is also the carrier on which object information rides (Chapter 19). Subtracting it would remove the signal. Instead, the relevant variable is the fractional change of the carrier, the electric image ϵ, which normalizes the world-caused change by the self-caused carrier.

The two strategies have a common form. Each computes a relation between the input and the animal's own contribution to it, a difference in the pulse-type case and a ratio in the wave-type case, and each yields a variable that measures the world relative to the self. They differ in where the reference comes from. The pulse-type fish must supply it internally, from a learned model driven by the command. The wave-type fish finds it in the signal: the carrier is on the skin, at every patch, continuously. Caputi's review of electric fish as model reafferent systems develops this comparison across species (Caputi 2004).

22.5 When No Efference Copy Is Required

The jamming avoidance response is the limiting case. The fish must determine the sign of the frequency difference between its own discharge and a neighbor's, and an engineer might expect it to compare the neighbor's frequency with an internal copy of its own pacemaker rhythm. It does not. The response is computed from the electrosensory mixture alone: in fish whose discharge has been silenced and replaced by an electronic mimic, the pacemaker shifts according to the sign of the frequency difference between the mimic and the jamming signal, although the mimic's frequency is independent of the pacemaker (Heiligenberg 1991a). *Gymnarchus* determines the direction of its shift "based solely upon the signal mixture" without referencing its pacemaker command (Kawasaki 1993), and it lacks the anatomical substrate for a corollary discharge to sensory areas: its pacemaker projects only through relay nuclei to the electromotor neurons, with no collaterals reaching sensory regions (Kawasaki 1994).

Theorem 21.3 explains why no copy is needed. The sign of Δf is carried by relations internal to the mixture: the direction of amplitude change at a patch and the phase difference between patches. Both are measured against the fish's own carrier, and the carrier is physically present at every patch. Any timing drift of the carrier is common to all patches and cancels in the differential phase; this is why the *Gymnarchus* response survives a large artificial jitter imposed on the whole stimulus (Y.-X. Guo and Kawasaki 1997).

Proposition 22.4 (The jamming computation does not distinguish self from mimic)**.** *In the model of Chapter 21, the sign computed by* (21.5) *depends only on the frequencies and amplitudes of the two signals present on the skin. Replacing the fish's own discharge by any signal of the same frequency and spatial distribution leaves the computed sign unchanged.*

Proof. The quantities A_j and δ in Theorem 21.3 are functions of m_j and $\psi(t)$ alone, which are determined by the amplitudes and frequencies of the two signals at each patch. The origin of the dominant signal does not enter. □

The proposition has an interesting consequence. The "self" against which the jamming computation measures the neighbor is the dominant carrier on the

skin, whatever produced it. In nature that carrier is the fish's own discharge, because the fish's electric organ is closer to its skin than any neighbor's. The computation is centered on the fish by physics and anatomy, and no representation of the fish as the source of the carrier enters it. This is what the experimental mimic exploits: the circuit accepts the mimic as "self" because nothing in the circuit checks.

22.6 Measurement Indexed to One's Own Action

Chapter 19 defined a self-indexed sensory variable as one that depends on the environment only through a relation to a quantity the agent produces or controls. The electric fish shows that such variables can be built in two ways.

Definition 22.5 (Agent-centered measurement frame)**.** A reading loop $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$ has an *agent-centered measurement frame* if the variables $z = \mathsf{D}(\mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}}))$ on which the policy π acts are self-indexed, so that the loop's actions are functions of the world relative to the agent's own activity. The frame is *model-based* if the reference is supplied by an internal prediction of self-caused input, and *carrier-based* if the reference is physically present in the mixture.

The mormyrid negative image is a model-based frame; the wave-type JAR is a carrier-based one. The distinction has a simple operational criterion.

Proposition 22.6 (When a carrier-based frame is available)**.** *A carrier-based frame is available at a sensory surface when the agent's own signal is present in the mixture at that surface throughout the reading, and when the relation between world and self can be computed from quantities that are invariant to the self signal's absolute parameters. When the self signal is transient, intermittent, or overwhelms the world-caused input it modulates, a frame must be model-based.*

The proposition is a structural observation rather than a theorem, and it predicts the division of labor observed in electric fish: continuous wave discharges support carrier-based computations such as the JAR, and intermittent pulse discharges, each of which produces a strong reafferent volley, are handled with learned negative images. It also explains why the most intensively studied examples of reafference in vision and touch are model-based: a saccade or a reach does not leave a continuous self-signal on the retina or skin against which world-caused input can be normalized.

22.7 The Minimal Agent-Centered Frame

The carrier-based frame of the jamming avoidance response is the simplest structure in this book that deserves the name "agent-centered." It exists before

any self-model: the fish does not represent itself as the source of its carrier, and Proposition 22.4 shows that the computation would accept any carrier as its own. It exists before memory, prediction, or learning: the computation is pointwise in time (Theorem 21.3) and innate (Viete and Heiligenberg 1991). And it produces variables of the form "the world relative to my activity": the sign of the neighbor's frequency relative to the fish's own, used to change the fish's own.

What the frame supplies is an origin for the fish's sensory coordinates. Information available to the system is indexed by the system's own intervention. This is the earliest structure we will meet that corresponds, however distantly, to what we intuitively call being "somewhere" rather than nowhere. Chapter 34 takes up the geometry of such coordinates, and Chapter 37 the logic of the indexical "I" that more complex systems use to refer to their own origin.

What the frame does not supply is anything that the adaptive radio of Chapter 48 lacks. A transmitter that monitors interference relative to its own carrier, and hops away from it, also measures the world relative to its own emission and acts on the measurement. The fish's frame is agent-centered in the same physical sense. Whatever additional property would make a system's variables centered on a *subject*, rather than on a source, is not present in Definition 22.5, and the book keeps the two questions apart.

22.8 Audit

Canonical phenomenon

Pulse-type electric fish cancel the predictable electrosensory consequences of their own discharges with a plastic negative image; wave-type fish perform the jamming avoidance response without any internal copy of their own discharge command.

Minimal model

The decomposition $s = \rho(a) + x(w)$; reafference cancellation (Model 22.2); the adaptive negative image (Model 22.3); Propositions 22.4 and 22.6.

What is relational here?

The world relative to the animal's own activity: a difference between input and predicted self-input (pulse-type), or a ratio and phase difference relative to a self-generated carrier (wave-type).

What physically reads it?

Cerebellum-like ELL circuits with command-locked plasticity in mormyrids; in wave-type fishes, the amplitude and differential-phase circuits of Chapter 21, which read relations to the carrier without a corollary discharge.

What existing field explains it?

The reafference principle (Holst and Mittelstaedt 1950), corollary discharge

(Sperry 1950; Crapse and Sommer 2008), cerebellum-like circuits and negative images (C. C. Bell 1981; C. Bell et al. 1997; Requarth and Sawtell 2011), the comparative analysis of electric fish as reafferent systems (Caputi 2004), and the jamming avoidance literature (Heiligenberg 1991a; Kawasaki 1993; Kawasaki 1994). The distinction between model-based and carrier-based frames is a convenient label for a contrast this literature already draws.

What remains unexplained?

A structural point rather than a mechanism: an agent-centered measurement frame can exist with no internal model of the self, centered by physics and anatomy. Whether any further property centers a system's variables on a subject remains open.

A discriminating experiment

In a mormyrid, pair an artificial stimulus with the discharge command and verify that a negative image forms (C. C. Bell 1981); in a wave-type fish, substitute a mimic for the fish's own discharge and verify that the JAR follows the mimic (Heiligenberg 1991a). A system in which the second manipulation abolished the response would be using a model-based frame for jamming avoidance, contrary to what is observed.

22.9 Notes and References

The reafference principle is due to Holst and Mittelstaedt (1950); Sperry (1950) introduced the corollary discharge in the same year. Crapse and Sommer (2008) review corollary discharge across the animal kingdom and clarify the terminology; Blakemore et al. (1998) is the standard study of self-produced tickle. The mormyrid negative image begins with C. C. Bell (1981); C. Bell et al. (1997) review the generation and subtraction of sensory expectations in cerebellum-like structures, and Requarth and Sawtell (2011) review the mechanisms. Caputi (2004) discusses electric fish as model systems for reafferent processing. The evidence that the JAR needs no internal reference comes from Heiligenberg (1991a) for *Eigenmannia* and from Kawasaki (1993) and Kawasaki (1994) for *Gymnarchus*. O'Regan and Noë (2001) develop a related idea for vision, in which perception consists in mastery of the sensorimotor contingencies linking action and input; Chapter 34 returns to it.

22.10 Exercises

Exercise 22.1. Model a fish swimming in a uniform visual surround with a heading h and turning rate $\dot{h} = -k(v - \hat{v})$, where v is the visual motion signal and $\hat{v} = -c\,\dot{h}$ the corollary-discharge prediction of the motion caused by turning. With normal eyes, $v = -\dot{h}$. With inverted eyes, $v = +\dot{h}$. Show that

the first case is stable for suitable $k, c > 0$ and the second produces runaway turning, as in Sperry's fish.

Exercise 22.2. Prove the convergence claim of Model 22.3: show that $\mathbb{E}[\hat{\rho}_n - \rho] = (1-\eta)^n \mathbb{E}[\hat{\rho}_0 - \rho]$, and compute the steady-state variance of $\hat{\rho}_n$ when x_n has variance σ^2. What does the trade-off in η correspond to biologically?

Exercise 22.3. Using Model 22.3, simulate Bell's pairing experiment: pair an artificial stimulus with every discharge for 200 discharges, then remove it. Plot the output u_n during pairing and after removal, and explain the transient response to the command alone.

Exercise 22.4. Add a random phase $\xi(t)$, common to all patches, to the mixture of Chapter 20. Show that $\delta(t)$ in Theorem 21.3 is unaffected and that the amplitude is unaffected. What kind of jitter would defeat the computation?

Exercise 22.5. Classify each of the following as using a model-based or a carrier-based frame, in the sense of Definition 22.5, and justify: cancellation of self-produced tickle; saccadic suppression; bat echolocation; the jamming avoidance response; active electrolocation of a metal object by a wave-type fish; a noise-cancelling headphone.

Exercise 22.6. Design an experiment that would distinguish, in a new species, a model-based from a carrier-based jamming avoidance mechanism. Specify the manipulation, the two predictions, and the control conditions.

23

Phantom Jamming and Error

A reader that responds to a relation can be driven by anything that reproduces the relation. Chapter 21 showed that the jamming avoidance circuit reads the sense of rotation of a joint trajectory of amplitude and differential phase. Nothing in that trajectory certifies that a second fish produced it. A stimulus that rotates the trajectory in the same way, with no neighbor in the water, should move the fish's frequency just as a neighbor would. It does.

This chapter is about what follows from that fact. The thesis is that phantom jamming is the first point in the book at which the vocabulary of error becomes operational: a physical discrimination drives an action appropriate to a situation that does not obtain. Whether the event is an *error* or merely a different response depends on a standard against which the action is measured, and naming that standard is the first step from reading toward meaning. The fish supplies the case; teleosemantics and its rivals supply the candidate standards, and Part VI develops them.

23.1 Mimicking Relational Cues

Carlson and Kawasaki (2007) asked whether stimuli that are not jamming signals can elicit jamming avoidance behavior because they are misperceived as jamming. They presented *Eigenmannia* with combinations of amplitude and phase modulation that did not arise from a second discharge near the fish's frequency. The fish shifted their discharge frequencies in response. The responses to the non-jamming stimuli had tuning, sensitivity, and dynamics similar to genuine jamming avoidance responses, and the magnitudes varied smoothly between the two, so that the phantom responses formed a continuum with the JAR rather than a separate category. The authors also identified natural contexts that could produce such stimuli: localized amplitude modulations, transient changes, and moving objects near the fish. They concluded that the fish perform frequency shifts in response to phantom jamming under ecologically relevant conditions.

The model of Chapter 21 explains why. The circuit reads the product of the direction of amplitude change and the sign of the differential phase

between body regions. Any stimulus that makes those two quantities covary with a consistent sign drives the circuit, and several natural situations can do it. An object moving along the fish's body, for example, changes the local amplitude at the patches it passes, and if it also shifts the timing of the local signal (as capacitive objects do; Chapter 19), the amplitude change and the timing difference between patches can covary. The circuit has no access to the cause of the covariation.

Example 23.1 (A trajectory without a neighbor). Deliver to two regions of the skin amplitude modulations $A_1(t) = A_2(t) = A_\mathrm{s}(1 + a\cos\Omega t)$ and impose on region 2 a timing modulation $\delta(t) = b\sin\Omega t$, with $a, b > 0$ and no second carrier. Then $\dot{A}_2\delta = -A_\mathrm{s}ab\,\Omega\sin^2\Omega t \le 0$, which by (21.5) is the signature of $\Delta f > 0$. The circuit should lower the fish's frequency, as it would for a neighbor above the fish at beat rate $\Omega/2\pi$. Reversing the sign of b should raise it.

23.2 Behavioral Misclassification

The fish's circuit can be described as a classifier. Let the world be in one of three states: a neighbor above, a neighbor below, or no neighbor. Let the circuit's output be one of three responses: lower the frequency, raise it, or leave it alone. A behavioral *misclassification* occurs when the response is the one appropriate to a different state of the world from the actual one. Phantom jamming is a misclassification of the third state as one of the first two.

Definition 23.2 (Misclassification)**.** Let W be a set of world states, Z a set of internal readings, and $\pi : Z \to \mathcal{A}$ a policy. Let $\alpha : W \to \mathcal{A}$ assign to each world state the action appropriate to it, according to a stated standard. The reading $z = \mathsf{D}(\mathsf{M}(s_\mathrm{A}, w))$ is a *misclassification* of w if $\pi(z) \neq \alpha(w)$.

The definition does all its work through α. Without a standard of appropriateness, the fish's shift in response to a phantom stimulus is simply a response, as lawful as the shift in response to a neighbor. With a standard, it becomes a mistake. The rest of the chapter asks where a standard for a fish could come from.

23.3 When Is an Error Really an Error?

Compare three devices that respond wrongly.

A thermometer in the sun reads 5 K above the air temperature. It is in error relative to the use its maker and its reader intend: to report air temperature. Nothing in the thermometer is different when it misreads; the standard is outside it, in the intentions of people.

An adaptive radio hops away from a channel on which an unrelated device produced a brief burst of noise. It is in error relative to its designers' purpose, avoiding persistent interference, and the standard is again external.

A fish lowers its frequency in response to a moving object. The standard that makes this an error, if it is one, is not supplied by any designer. It must come from the fish's own organization or history.

Principle 23.3 (No error without a standard)**.** A physical deviation between a system's response and some feature of its environment counts as an error only relative to a standard that specifies what the response is for. The standard may be external (a designer's intention, a user's purpose) or internal (a function the system has by virtue of its history or its organization).

The principle is a compressed statement of a position that runs through the philosophy of representation: content and correctness conditions depend on function, and function depends on something beyond the current causal facts (Millikan 1989; F. Dretske 1988). It is also a warning against an easy inference. The fact that the circuit "responds to the same thing" in the jamming and phantom cases does not settle whether the phantom response is a mistake. If the thing the circuit is for is "a rotating amplitude and phase trajectory," the phantom response is correct. If it is "a neighbor at a nearby frequency," it is an error.

23.4 False Positives and False Negatives

Signal detection theory gives the error its quantitative form. Suppose the circuit's pooled output is a scalar z whose distribution depends on the world state, and the fish shifts its frequency whenever $|z|$ exceeds a criterion c. A *false positive* is a shift with no neighbor present; a *false negative* is a failure to shift when a neighbor is present. Raising c trades false positives for false negatives.

Model 23.4 (Criterion placement under asymmetric costs)**.** Let p be the probability that a neighbor is present, C_{FP} the cost of a phantom shift (a small change of frequency, some energy, some disruption of communication), and C_{FN} the cost of a missed neighbor (degraded electrolocation for as long as the jamming lasts). For likelihoods $\ell_1(z)$ and $\ell_0(z)$ of the pooled output with and without a neighbor, the expected cost is minimized by shifting whenever

$$\frac{\ell_1(z)}{\ell_0(z)} > \frac{(1-p)\,C_{\mathrm{FP}}}{p\,C_{\mathrm{FN}}}. \tag{23.1}$$

The model is standard Bayesian decision theory and its parameters are illustrative. Its lesson for the fish is qualitative. If phantom shifts are cheap and missed jamming is costly, the optimal criterion is liberal, and a well-designed

circuit will produce phantom responses as a matter of course. Phantom jamming is then the visible cost of a sensible trade-off, not a defect. Whether the costs in fact have this shape for *Eigenmannia* is an empirical question the model poses and does not answer.

23.5 Teleology Versus Mere Mismatch

Four candidate standards could make a fish's response wrong, and they give different verdicts on the phantom case.

Design. The fish has no designer, so this standard is unavailable except as metaphor.

Selection history. Teleosemantic theories fix the function of a biological mechanism by what it was selected to do: the feature of past environments whose presence explains why the mechanism was retained (Millikan 1984; Millikan 1989). If the jamming avoidance circuit was retained because it moved the fish away from neighbors' frequencies, then its function is to respond to neighbors, and a response to a moving object is an error relative to that function, however harmless. On this view phantom jamming is misrepresentation, in the sense developed in Chapter 30.

Current viability. Enactive and autonomy-based theories locate norms in the ongoing self-maintenance of the organism: what matters is whether a response contributes to or detracts from the system's continued existence (Di Paolo 2005; Barandiaran et al. 2009). On this view a phantom shift is an error only if it harms the fish, and a cheap shift may not count as an error at all. Chapter 28 develops the view.

Prediction. In predictive-processing accounts, error is the mismatch between a system's predictions and its input (K. Friston 2010; Clark 2013). The fish's circuit makes no prediction about the cause of the rotating trajectory, so this standard classifies nothing in the phantom case; it becomes relevant only in systems with generative models (Chapter 33).

The four standards disagree, and the disagreement is informative. The fish has settled what it responds to; it has not settled which of our standards applies to it. The empirical facts about the circuit do not by themselves decide whether phantom jamming is error, mismatch, or correct performance. That is the point at which the question of meaning, as distinct from the question of reading, first becomes unavoidable.

23.6 The Beginning of Semantic Questions

Consider a sign-selective neuron in the torus that fires when the neighbor is above the fish. What does its firing mean? Three answers are available.

It means "the joint trajectory of amplitude and differential phase is rotating clockwise." This is what the neuron's firing is lawfully correlated with, in both

the jamming and the phantom cases. On this reading the neuron never errs, because it reports exactly what drives it.

It means "a neighbor is discharging at a slightly higher frequency." This is what the firing usually indicates in the fish's natural environment, and what the downstream behavior is appropriate to. On this reading the phantom case is a false report.

It means "a neighbor above *or* a moving capacitive object *or* any of the other causes that produce the trajectory." This disjunctive content makes the neuron infallible by including every cause of its firing.

The choice among these is the *disjunction problem*, which Fodor used to argue that purely causal and informational theories of content cannot explain how misrepresentation is possible: if content is whatever reliably causes the representation, every cause is part of the content and error disappears (Fodor 1987; Fodor 1990). Teleosemantic theories answer by choosing the second reading: content is fixed by what the mechanism was selected to respond to, so the neuron means "neighbor above" and the phantom case is misrepresentation (Millikan 1989; Neander 2017). Informational theories supplemented by learning histories give related answers (F. Dretske 1988). Chapter 29 surveys the options and Chapter 30 asks what distinguishes misrepresentation from any other physical deviation.

The fish does not solve the disjunction problem. It gives it a concrete home. The neuron, its inputs, its lawful causes, and the behavior it drives are all known; the question that remains is which of those facts fixes what the neuron is about.

23.7 Audit

Canonical phenomenon

Eigenmannia shifts its discharge frequency in response to non-jamming combinations of amplitude and phase modulation, with tuning and dynamics continuous with the true jamming avoidance response (Carlson and Kawasaki 2007).

Minimal model

The sign rule (21.5) applied to a stimulus without a second carrier (Example 23.1); misclassification relative to a standard (Definition 23.2); criterion placement under asymmetric costs (Model 23.4).

What is relational here?

The covariation of amplitude change and differential phase, which the circuit reads whatever its physical cause.

What physically reads it?

The jamming avoidance circuit of Chapter 21. Its reading is the same in the jamming and phantom cases; what differs is the world.

What existing field explains it?

The phantom experiments (Carlson and Kawasaki 2007), signal detection and decision theory, teleosemantics (Millikan 1984; Millikan 1989; Neander 2017), informational semantics (F. Dretske 1988), the disjunction problem (Fodor 1990), and enactive accounts of normativity (Di Paolo 2005).

What remains unexplained?

The mechanism leaves nothing unexplained. What remains open is normative: which standard fixes what the fish's reading is about, and so whether phantom jamming is misrepresentation. The chapter shows that the empirical facts do not settle it.

A discriminating experiment

Measure the fitness or energetic cost of phantom shifts and of missed jamming in the fish's natural habitat, and compare the observed criterion with Model 23.4. A criterion near the cost-optimal one would support treating phantom shifts as the price of a selected trade-off; a criterion far from it would suggest the circuit's function is something other than jamming avoidance alone.

23.8 Notes and References

The phantom jamming experiments are reported by Carlson and Kawasaki (2007). The circuit whose inputs they mimic is described in Heiligenberg (1991a) and Chapter 21. On function and content, Millikan (1984) and Millikan (1989) develop teleosemantics, Neander (2017) defends an informational version, and F. Dretske (1988) grounds content in learning histories. Fodor (1987) and Fodor (1990) pose the disjunction problem for causal and informational theories. Autonomy-based accounts of normativity are developed by Di Paolo (2005) and Barandiaran et al. (2009); predictive-processing accounts of error by K. Friston (2010) and Clark (2013).

23.9 Exercises

Exercise 23.1. Verify the computation in Example 23.1. Then let the timing modulation lead the amplitude modulation by a phase β, $\delta(t) = b\sin(\Omega t + \beta)$, and compute the average of $\dot{A}_2\delta$ over a cycle as a function of β. For which β should the phantom response vanish?

Exercise 23.2. Model an object moving at constant speed along the fish's body as a Gaussian bump of amplitude increase traveling from head to tail, accompanied by a small local timing shift proportional to the bump. Compute $\dot{A}\,\delta$ for two patches and determine the sign of the implied Δf. Does reversing the object's direction reverse the phantom response?

Exercise 23.3. Derive (23.1). For Gaussian likelihoods with equal variance and means separated by d' standard deviations, find the criterion as a function of p and the cost ratio, and plot the false-positive rate against the cost ratio for $p = 0.1$ and $p = 0.5$.

Exercise 23.4. For each of the four standards of Section 23.5, state an observation that would count, under that standard, as showing the phantom shift to be an error. Which of these observations could be made in the laboratory?

Exercise 23.5. Formulate the three candidate contents of Section 23.6 as conditions under which the neuron's firing is correct. For each, list the stimuli that would count as errors. Which content makes the neuron's behavior most informative to an outside observer, and does that settle anything about what the firing means for the fish?

Exercise 23.6. Construct an adaptive radio that exhibits "phantom hopping" in response to a signal that mimics interference. Explain why nobody hesitates to call this an error, and identify which of the four standards licenses the judgment.

24

Convergent Evolution of Reading

Eigenmannia lives in South America and *Gymnarchus* in Africa. Their lineages became electric independently, and their electrosensory brains were built from different ancestral parts. Both emit continuous wave discharges, both suffer jamming when a neighbor's frequency comes close, and both shift their frequencies apart. They do it with the same computational rules (Kawasaki 1993; Kawasaki 1996). They do not do it with the same brain regions: the phase comparison at the heart of the computation takes place in the hindbrain of *Gymnarchus* and in the midbrain of *Eigenmannia* (Kawasaki 1996; Kawasaki and Y. Guo 1996).

This chapter asks what that convergence shows. The thesis has two parts. First, the convergence is what the physics of the problem predicts: a reader with no internal reference must use differential phase, and because the amplitude alone carries no sign, any reader that extracts the sign must combine amplitude with differential phase. The structure of the relation fixes the algorithm and leaves the implementation free. Second, the convergence gives Marr's distinction between computational, algorithmic, and implementational levels an empirical anchor: two independent lineages share the first two and differ in the third, so the shared description is a fact about the animals and not an interpretation imposed on them.

24.1 Two Independent Origins

Muscle-derived electric organs evolved at least six times among fishes, and genomic comparisons of independent lineages show similar developmental and cellular pathways recruited repeatedly (Gallant et al. 2014). Electrogenesis in the African weakly electric fishes (the mormyroids, which include *Gymnarchus*) and in the South American gymnotiforms (which include *Eigenmannia*) arose independently, and molecular dating places the two origins at comparable times (Lavoué et al. 2012). Among the African weakly electric fishes, *Gymnarchus*

niloticus is distinctive in emitting a continuous wave discharge, of about 400 Hz (Kawasaki and Y. Guo 1996); its relatives, the mormyrids, emit pulses. The wave-type discharge, and with it the jamming problem, arose separately on the two continents.

	Eigenmannia	*Gymnarchus*
Continent, lineage	South America, Gymnotiformes	Africa, Mormyroidea
Discharge	continuous wave, a few hundred hertz	continuous wave, about 400 Hz
Jamming avoidance response	yes	yes
Internal reference to pacemaker	not used	not used; no corollary discharge to sensory areas
Computational rules	amplitude with differential phase	the same
Site of phase comparison	midbrain (torus semicircularis)	hindbrain (electrosensory lateral line lobe)

Table 24.1: Two independently evolved jamming avoidance systems. Sources: Heiligenberg (1991a), Kawasaki (1993), Kawasaki (1994), Kawasaki (1996), Kawasaki and Y. Guo (1996).

24.2 The Same Problem

The physics of jamming does not depend on the continent. Any wave-type electrolocating fish that meets a neighbor with a nearby frequency receives on its skin the mixture of Chapter 20: a carrier at its own frequency, modulated in amplitude and phase at the beat rate, with modulation depths that vary over the body according to the neighbor's position. The beat degrades electrolocation in the same way, and the remedy, moving one's own frequency away from the neighbor's, requires the same information, the sign of the frequency difference. The two fishes face one problem, posed by the same relation.

24.3 The Same Computational Rules

Kawasaki (1993) analyzed the jamming avoidance response of *Gymnarchus* behaviorally, using the stimulus paradigms developed for *Eigenmannia*. The fish determines the direction of its frequency shift from the signal mixture alone, without referring to its own pacemaker command. It evaluates the amplitude and phase modulations over different body regions and compares phase between regions, and the rules by which it combines these cues are the rules that *Eigenmannia* uses. Kawasaki concluded that the two genera employ identical computational algorithms, and suggested that both built

their jamming avoidance systems from preexisting electrolocation circuitry. The same suggestion had been made for the gymnotiforms on the evidence of *Sternopygus*, the one gymnotiform genus that lacks a jamming avoidance response and whose electrolocation mechanisms look like the ancestral substrate from which the response evolved (G. Rose, C. Keller, et al. 1987).

Two further facts sharpen the parallel. *Gymnarchus* lacks the corollary discharge pathways that its pulse-type relatives use to gate their electrosensory input: its pacemaker projects only through relay nuclei to the electromotor neurons, and no collaterals reach sensory regions (Kawasaki 1994). It could not use an internal reference if it tried. And its temporal acuity rivals *Eigenmannia*'s: phase-locked neurons in its hindbrain time the signal with a precision of a few microseconds, the behavior responds to time disparities below 1 µs, and the response survives a large jitter imposed on the phase of the whole stimulus because the relevant neurons read only differential phase (Y.-X. Guo and Kawasaki 1997).

24.4 Different Implementations

The anatomy differs where it matters most. In *Eigenmannia*, differential phase is first computed in lamina 6 of the torus semicircularis, in the midbrain, from the timing information relayed by the hindbrain (G. Rose and Heiligenberg 1985a; C. Carr et al. 1986). In *Gymnarchus*, timing information from S-type electroreceptor afferents is compared in the inner cellular layer of the electrosensory lateral line lobe, in the hindbrain, where giant neurons and afferent terminals converge on neurons sensitive to differential phase (Kawasaki and Y. Guo 1996). The same step of the algorithm is carried out one major brain division earlier.

The combination of amplitude and phase also has its own implementation in *Gymnarchus*. Amplitude and phase information travel through separate medullary pathways and converge in the torus semicircularis on combination-sensitive neurons, which become selective for the sign of the frequency difference through nonlinear interaction rather than by summation: responses are facilitated when the amplitude and phase response peaks coincide in time, and suppressed when they are offset (Carlson and Kawasaki 2004). The same study identified neurons sensitive to amplitude *differences* between body regions, which escape an ambiguity that affects neurons combining amplitude with phase at a single orientation of the neighbor.

Kawasaki (1996) summarized the comparison: an identical set of complex computational rules, implemented by different brain structures in independently evolved systems with a similar behavioral function.

24.5 Marr's Levels, Made Empirical

Marr (1982) distinguished three levels at which an information-processing system can be understood. The *computational* level states what is computed and why; the *algorithmic* level states the representation and the procedure; the *implementational* level states how the procedure is realized physically.

Level	*Eigenmannia*	*Gymnarchus*
Computational	sign of Δf, to move f_{s} away from f_{e}	the same
Algorithmic	direction of amplitude change combined with differential phase across body regions; no internal reference	the same
Implementational	phase comparison in torus lamina 6; sign selectivity in deeper torus laminae; push-pull prepacemaker control	phase comparison in the hindbrain ELL; combination-sensitive neurons in the torus

Table 24.2: The jamming avoidance systems at Marr's three levels.

For most systems the division into levels is an analyst's choice. One might describe a circuit's algorithm in several ways, and nothing in the circuit favors one description. Chapter 17 discusses the worry in its general form. The electric fishes turn the choice into a finding. The algorithmic description in Table 24.2 is the description that two independently evolved nervous systems share, and it was established by behavioral experiments that manipulated amplitude and phase independently in each species. A description that picks out what independent lineages have in common, under experimental control, is a description of the animals, and it is the right grain at which to state what they compute.

24.6 Why the Algorithm Converged

Convergent evolution is usually explained by shared constraints: similar problems under similar physics admit few good solutions. In this case the constraint can be stated as a theorem about the signal.

Proposition 24.1 (A sign reader must be joint)**.** *Consider the mixture of Chapter 20 at any number of patches, to first order in the contamination ratios, and a reader with no access to the absolute phase of the carrier. The signals available to such a reader are the patch amplitudes $A_j(t)$ and the differential phases $\delta_{jk}(t) = \varphi_j(t) - \varphi_k(t)$. Then no function of the amplitudes alone, and no function of the differential phases alone, determines* $\operatorname{sgn}(\Delta f)$ *for every initial beat phase ψ_0. A reader that determines the sign must depend jointly on at least one amplitude and at least one differential phase.*

Proof. To first order, $A_j \propto 1 + m_j \cos\psi$ and $\delta_{jk} \propto (m_j - m_k)\sin\psi$, with $\psi = \Delta\omega\, t + \psi_0$. The substitution $(\Delta\omega, \psi_0) \mapsto (-\Delta\omega, -\psi_0)$ leaves every $A_j(t)$ unchanged, so a function of the amplitudes alone gives the same output for opposite signs of Δf. The substitution $(\Delta\omega, \psi_0) \mapsto (-\Delta\omega, \pi - \psi_0)$ maps ψ to $\pi - \psi$ and leaves every $\sin\psi$, and so every $\delta_{jk}(t)$, unchanged; a function of the differential phases alone gives the same output for opposite signs. Theorem 21.3 exhibits a joint function that succeeds. □

At higher order the phase waveform carries a weak trace of the sign (Proposition 21.2), but at the modest contamination ratios where jamming matters most, the trace is small, and a reader relying on it would be slow and fragile. The proposition therefore explains the convergence at the algorithmic level. Any electrolocating fish that lacks an internal reference and must extract the sign of a frequency difference is pushed toward the same kind of solution, because the information is in the joint modulation and nowhere else. Evolution found the solution twice because the physics of the mixture leaves only one good place to look.

The same reasoning explains the absence of an internal reference in both systems. By Theorem 1.5, a reader without access to the absolute phase origin can use only frame-independent variables, the relations internal to the mixture. That condition was satisfied in both lineages by the continuous presence of the fish's own carrier on its skin (Chapter 22), and neither lineage needed the corollary discharge machinery that pulse-type fishes use.

24.7 Multiple Realizability, Carefully

The functionalist tradition in philosophy of mind holds that mental states are defined by their functional roles and can be realized in physically different systems (Putnam 1967; Fodor 1974). The electric fishes are often a tempting illustration: one algorithm, two implementations. The illustration needs care.

Both implementations are vertebrate nervous systems built from homologous cell types, and they differ in which brain region performs a step, not in the kind of material that performs it. What the fishes show is *locus independence* within a shared kind of substrate, a weaker claim than the substrate independence functionalism usually intends. Bechtel and Mundale (1999) argue that neuroscience rarely supports strong multiple realizability, because functional and neural taxonomies usually line up once they are drawn at matching grains. The fishes fit their argument as well as they fit the functionalist's: at the grain of "phase comparison in a vertebrate brain," the two realizations are the same kind of thing; at the grain of "hindbrain versus midbrain," they differ.

The case supports a narrower conclusion that the book needs. A relational computation can be a real property of a nervous system, specifiable independently of its anatomical locus and discoverable by experiment in more than one lineage. Whether the same holds for anything experiential is a separate

question. If phenomenal properties depended only on the algorithmic level, the two fishes would be alike in them, whatever they are; if they depended on implementational details, the fishes might differ. Nothing in the jamming avoidance literature bears on the choice, and Chapter 50 returns to it.

24.8 Audit

Canonical phenomenon

Gymnarchus and *Eigenmannia*, whose electric senses evolved independently, perform jamming avoidance with identical computational rules implemented in different brain divisions.

Minimal model

The first-order modulations $A_j \propto 1 + m_j \cos\psi$ and $\delta_{jk} \propto (m_j - m_k)\sin\psi$; Proposition 24.1; Marr's three levels.

What is relational here?

The sign of Δf, which the physics of the mixture makes available only in the joint behavior of amplitude and differential phase.

What physically reads it?

Differential-phase neurons in the hindbrain of *Gymnarchus* and in the midbrain of *Eigenmannia*, and in each case neurons that combine amplitude and phase to yield sign selectivity.

What existing field explains it?

Comparative neuroethology (Kawasaki 1993; Kawasaki 1996; Kawasaki and Y. Guo 1996; Carlson and Kawasaki 2004; G. J. Rose 2004), the phylogenetics of electrogenesis (Lavoué et al. 2012; Gallant et al. 2014), Marr's levels (Marr 1982), and the philosophical debate on multiple realizability (Putnam 1967; Bechtel and Mundale 1999).

What remains unexplained?

The proposition that a sign reader must be joint is elementary, and the convergence it explains was documented before any such proposition was written down. What the chapter adds is the explicit link between the structure of a relation and the convergence of the algorithms that read it. The general form of that link, whether the structure of a relation constrains the algorithms of every system that reads it, is a candidate for the transferable principle Part XII asks about.

A discriminating experiment

Examine a third, independently evolved wave-type electrolocating system, or an engineered one built to avoid interference without an internal reference. The proposition predicts that any solution will combine amplitude change

with differential phase; a working solution that reads the sign from amplitude alone, or from differential phase alone at low contamination, would refute it.

24.9 Notes and References

The behavioral analysis of the *Gymnarchus* jamming avoidance response is Kawasaki (1993); the absence of corollary discharge pathways is shown in Kawasaki (1994); the hindbrain phase-comparison circuit in Kawasaki and Y. Guo (1996); the temporal precision in Y.-X. Guo and Kawasaki (1997); and the midbrain combination-sensitive neurons in Carlson and Kawasaki (2004). Kawasaki (1996) compares the two genera directly, and G. J. Rose (2004) reviews what electric fish have taught about neural mechanisms and the evolution of behavior. On the independent origins of electrogenesis, see Lavoué et al. (2012) and Gallant et al. (2014); on the origins of the JAR in electrolocation circuitry, G. Rose, C. Keller, et al. (1987). Marr (1982) introduced the three levels. The multiple realizability debate begins with Putnam (1967) and Fodor (1974); Bechtel and Mundale (1999) give the neuroscientific critique.

24.10 Exercises

Exercise 24.1. Verify the two substitutions in the proof of Proposition 24.1. Then show that no single substitution of the form $(\Delta\omega, \psi_0) \mapsto (-\Delta\omega, \psi_0')$ leaves both the amplitudes and the differential phases unchanged, so that a joint reader can succeed.

Exercise 24.2. Extend the analysis to second order in the contamination ratios. Construct a function of the differential phases alone that determines the sign at second order, and estimate how much longer it would need to integrate than the joint rule of Theorem 21.3 to reach the same reliability at $m = 0.1$.

Exercise 24.3. Fill in a table like Table 24.2 for two independently evolved systems of your choice (for example, echolocation in bats and toothed whales, or camera eyes in vertebrates and cephalopods). At which level do they converge, and what physical constraint explains the convergence?

Exercise 24.4. Design an interference-avoidance scheme for a radio that cannot access its own oscillator's phase directly but receives the mixture of its own and an interferer's carrier at two antennas. Use Proposition 24.1 to argue what your scheme must measure. Does your design "converge" with the fish in the sense of this chapter?

Exercise 24.5. State precisely the difference between locus independence and substrate independence. Give an example, from any field, of a computation that exhibits the second and not merely the first, and assess whether the example would support a functionalist argument about mental states.

Exercise 24.6. Suppose a new wave-type species performs jamming avoidance but, unlike *Eigenmannia* and *Gymnarchus*, has a corollary discharge projection to its electrosensory system. State two hypotheses about how it might use the corollary discharge, and design the mimic experiment that distinguishes them.

The weakly electric fish generates the carrier it senses, reads the beat its neighbor imposes on that carrier, extracts the sign of a frequency difference, which only the joint behavior of amplitude and phase carries exactly, and changes its own frequency to change the relation. It does this without consulting a copy of its own motor command: the mixture carries the information. Two lineages that diverged long before either became electric arrived at the same algorithm with different brain regions. And the reading can be wrong. Part VI asks what it takes for such a reading to matter to the system that performs it.

Part VI

From Reading to Meaning

Why does the distinction matter to the system that draws it?

25

The Epistemic Cut

A measuring instrument is a physical system obeying the same laws as the system it measures, and yet its output is a record: a stable, discrete state that stands for something other than itself. The pointer's position is a fact about the pointer, governed by mechanics, and at the same time a report about a voltage. Somewhere in any account of measurement, the description changes register, from equations of motion to statements about what was found. Howard Pattee called that boundary the *epistemic cut*.

The thesis of this chapter is that the epistemic cut, which physics treats as a feature of descriptions, is drawn by living systems inside their own organization. Pattee's contribution was to identify the physical conditions under which such a cut can exist in matter: records that are energy-degenerate and rate-independent, constraints that cannot be integrated away, and, for open-ended evolution, linear heteropolymers that can both store a description and fold into the machinery that reads it. The jamming avoidance response contains a small instance of the same structure, where continuous electrosensory dynamics end in a binary choice between two motor pathways. Seeing the cut in the fish, and first in the cell, turns the reader of Part IV into the measuring organism that the rest of Part VI examines.

25.1 Pattee

Pattee's question, posed in 1969, was how a molecule becomes a message (Pattee 1969). A strand of nucleic acid is a chemical object with a definite structure and definite interactions. It is also, in the cell, a sequence of instructions whose meaning is fixed by the machinery that translates it. Physics describes the first aspect completely and says nothing about the second. For five decades Pattee asked what physical conditions allow a material structure to function as a symbol, and his answers, collected with commentary by Pattee and Rączaszek-Leonardi (2012), form the most sustained attempt to treat measurement, symbols, and control as problems of physics rather than of interpretation (Pattee 1982; Pattee 1995; Pattee 2001).

The problem is a generalization of the one Chapter 17 met in the philosophy of computation. There, the worry was that any physical system could be mapped onto any computation, so that computation seemed to depend on an interpreter. Pattee's version asks how a physical system can contain its own interpreter: how a part of the world can measure another part, record the result, and use the record to control what happens next, without an outside observer supplying the meaning of the record.

25.2 Dynamics Versus Symbols

Pattee's analysis rests on a contrast between two ways in which physical structure can figure in explanation. Laws are universal, inexorable, and rate-dependent: they describe how states change in time, continuously and everywhere, whatever anyone knows about them. Symbols, in the sense Pattee needed, are local structures whose effects do not depend on the rate at which they are read and whose physical alternatives are energetically equivalent.

Definition 25.1 (Rate-independent, energy-degenerate record)**.** A physical structure is a *record* when (i) its state can take any of several alternative configurations of equal or nearly equal free energy, so that physical law does not select among them (*energy degeneracy*), and (ii) the effect of the configuration on other processes does not depend on how fast the configuration is read or on its detailed time course (*rate independence*).

The genome is the paradigm. The sequences of a nucleic acid of given length are, to a first approximation, energetically equivalent: the backbone is the same whatever the order of bases, and the chemistry of polymerization does not favor one order over another. Which sequence a cell carries is therefore not fixed by the laws of chemistry; it is a record of history, fixed by descent and selection. And the sequence controls the cell's dynamics in a rate-independent way: a gene specifies the same amino acid sequence whether it is transcribed quickly or slowly. Pattee described the genotype–phenotype distinction as a primeval epistemic cut that separates energy-degenerate, rate-independent genetic symbols from the rate-dependent dynamics of construction that they control, and argued that the same symbol–matter distinction recurs at every higher level where symbols are related to referents by an arbitrary code (Pattee 2001).

Table 25.1 sets out the contrast.

The word "arbitrary" in the last row does not mean random. It means that the relation between a symbol and what it does is not fixed by the physics of the symbol alone. The same codon could have specified a different amino acid had the translation machinery been built differently, and in a few lineages it does. The relation is fixed by a reader, and the reader is itself a physical structure, which is where the problem of Chapter 26 begins.

Table 25.1: Pattee's contrast between laws and symbols, as used in this chapter.

	Laws and dynamics	**Records and symbols**
Scope	Universal	Local, carried by particular structures
Time	Rate-dependent: continuous change	Rate-independent: effect does not depend on reading speed
Energy	States differ in energy and are selected by it	Alternatives energetically degenerate
Determination	By initial conditions and law	By history (descent, selection, design)
Relation to referent	None: dynamics are not about anything	Arbitrary code, fixed by a reading mechanism
Example	Folding, catalysis, diffusion, a field on the skin	Base sequence, a switch position, a stored bit

25.3 Measurement

Control runs from symbols to dynamics: a rate-independent record constrains a rate-dependent process. Pattee called measurement its converse, the process in which a rate-dependent dynamical state is coded into quiescent symbols (Pattee 2001).

Definition 25.2 (Measurement and control)**.** A *measurement* is a physical process that takes a system whose state changes continuously in time and leaves a record (Definition 25.1) whose configuration depends on that state. A *control* is a physical process in which the configuration of a record constrains the time course of a dynamical process.

Definition 25.3 (Epistemic cut)**.** The *epistemic cut* is the boundary, in a description of a system that measures or controls, between the part described by laws and initial conditions and the part described as records whose configurations stand for something.

Physics has known for a long time that such a boundary is required and that its location is partly a matter of choice. In his analysis of the quantum measuring process, von Neumann divided the world into the system observed, the measuring instrument, and the actual observer, and showed that the boundary between observed and observer can be drawn between the system and the instrument, or between the instrument and the observer, or further along the chain, without changing any prediction (Neumann 1955). What cannot be done is to omit the boundary. A description that treats everything, instrument and observer included, as dynamics contains no result, only a larger system evolving.

Example 25.4 (A thermometer, three ways). A mercury thermometer in a bath can be described as dynamics throughout: heat flows, the mercury expands, the column rises, light scatters from it into an eye. It can be described with a cut between the bath and the thermometer: the bath's temperature is recorded as the column's height. Or the cut can sit between the column and the reader of the scale: the column's height is recorded as a number in a notebook. Each placement gives the same prediction for the number in the notebook. Only the fully dynamical description has no record in it, and so no measurement.

Records are physical, and their physics has consequences. Landauer argued that information is always carried by a physical representation, and that erasing a record has an irreducible thermodynamic cost (Landauer 1991). A measuring device that is to measure again must reset its record, and the reset dissipates energy. The epistemic cut is thus a boundary in description that is marked, in the world, by structures with a definite thermodynamics.

25.4 Constraints

What kind of physical structure can hold a record? Pattee's answer is phrased in the language of constraints: non-integrable constraints are one necessary condition for bridging the epistemic cut by measurement, control, and coding (Pattee 2001).

A constraint restricts the motions available to a system. A *holonomic* constraint restricts its configurations: a bead on a wire obeys $g(\boldsymbol{q}) = 0$ and has one fewer degree of freedom than a free bead. Such a constraint can be eliminated by a change of coordinates, after which the system is simply a smaller system with no memory of the constraint. A *non-holonomic* or non-integrable constraint restricts velocities in a way that cannot be integrated into a restriction on configurations (Landau and Lifshitz 1976).

Example 25.5 (The rolling disk). A disk of radius ρ rolls without slipping on a plane, remaining upright. Its configuration is given by the contact point (x, y), the heading ϑ, and the rotation angle ϕ of the disk about its axle. Rolling without slipping imposes

$$\dot{x} = \rho\,\dot{\phi}\cos\vartheta, \qquad \dot{y} = \rho\,\dot{\phi}\sin\vartheta. \tag{25.1}$$

These are constraints on velocities, and they cannot be integrated into two equations among x, y, ϑ, ϕ: by steering, the disk can reach every configuration of the four coordinates. Two disks that end at the same position and heading can differ in ϕ, and the difference records something about the paths they took.

The example shows why non-integrable constraints matter for records. A holonomic constraint leaves nothing behind: once it is imposed, the reduced system's state is a function of its present coordinates, and its past is forgotten

except through the ordinary dynamics. A non-integrable constraint allows the same local laws to leave the system in any of many configurations, depending on history, and so it can carry a record of which history occurred. Switches, latches, and ratchets have this character, and so, in Pattee's analysis, does every device that measures: its pointer can end in any of several positions that the laws treat alike, and which position it ends in depends on what it interacted with.

Pattee also argued that non-integrable constraints are necessary but not sufficient for biology. Additional properties of heteropolymer constraints are needed for biological evolution (Pattee 2001): a linear sequence of monomers that can store a record in its order, be copied without being read, and fold into a three-dimensional structure whose shape carries out the control the record specifies.

25.5 Control

The logical structure that Pattee's heteropolymers realize was worked out, before the molecular details were known, by von Neumann (Neumann 1966). Von Neumann asked what an automaton would need in order to construct a copy of itself that could, in turn, construct a copy of itself, and to do so in a way that could accumulate changes. His answer separated four parts: a universal constructor A, which builds whatever automaton a description specifies; a copier B, which duplicates any description; a controller C, which coordinates the two; and a description Φ of $A + B + C$. Given Φ, the machine builds $A + B + C$ from the description, copies the description, and attaches the copy to the new machine.

The description is used twice, in two different ways. The constructor *interprets* it: the description's configuration controls what is built. The copier *does not interpret* it: the description is duplicated as a string, without regard to what it says. The two uses correspond to the two sides of the epistemic cut. Interpretation crosses the cut, from record to dynamics. Copying stays on the symbol side, and that is what allows a change in the description (a mutation) to be inherited even when its effect on construction is new.

Pattee took von Neumann's scheme as the logical requirement for open-ended evolution and the genome as its physical realization. A nucleic acid sequence is copied by replication, which does not depend on what the sequence specifies, and read by translation, which does. The translation apparatus that does the reading is itself specified by sequences in the same genome. That last fact is the subject of Chapter 26.

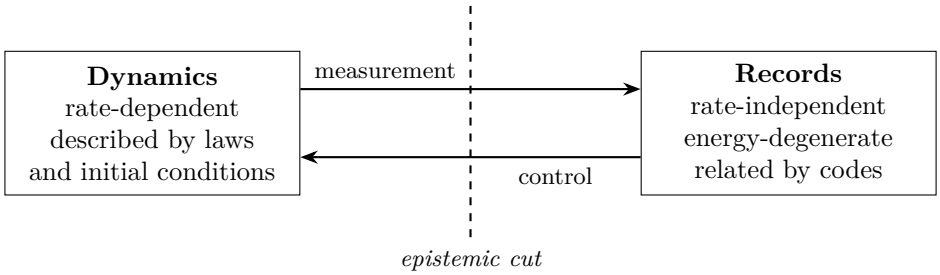


Figure 25.1: The epistemic cut. Measurement codes a dynamical state into a record; control lets a record constrain dynamics. Pattee argued that both directions require non-integrable constraints, and that open-ended evolution requires records that can also be copied without being read (Pattee 2001; Neumann 1966).

25.6 Physical Embodiment of Measurement

For von Neumann, the placement of the cut was a matter of descriptive convenience. For an organism, Pattee argued, it is not. The organism places its own cuts: it makes records at definite places (a genome, a stored synaptic weight, the setting of a regulatory switch) and uses them at definite places, and the boundary between what is treated as dynamics and what is treated as record is fixed by its organization rather than by an observer's choice. In his late work Pattee argued that the first memory-based self-replicating unit, such as a cell, is the most primitive case of a necessary symbol–matter distinction, and that all sensing, detection, and measurement require the subject–object distinction he called the epistemic cut (Pattee 2015).

The jamming avoidance response contains a small instance of this structure. Everything on the input side is rate-dependent dynamics: the fish's own field and its neighbor's superpose continuously on the skin, the receptors fire in time with the discharge, and the modulations of amplitude and phase evolve at the beat rate. The output is a choice between two alternatives, and the alternatives are carried by physically separate pathways. Frequency rises are driven by one pathway through the prepacemaker nucleus to the pacemaker, and frequency falls by another that acts through the sublemniscal prepacemaker nucleus, in a push-pull arrangement with no single population of final recognition units (Metzner 1993b). Which pathway is active is, in the vocabulary of this chapter, a record of the sign of Δf: a discrete configuration, selected by the history of the sensory input and not by the energetics of the pathways, whose effect on the pacemaker is a control. The measurement is embodied in the circuit, and the cut falls where the circuit changes from following the continuous beat to selecting one of two outputs.

Remark 25.6. The fish's record is not quiescent in Pattee's strict sense: it is maintained by neural activity, not by a stable molecular configuration, and it lasts only as long as the stimulus that sets it. The analogy is structural. The

fish shows that a continuous physical relation can end in a discrete, history-selected alternative that controls further dynamics, inside the organism and without an external observer; it does not show that the alternative is a symbol in the full sense that a codon is.

Pattee's language invites a stronger reading that this book declines. He described the epistemic cut as a subject–object distinction and spoke of the phenomena of the first cells (Pattee 2015). The words "subject" and "phenomenon" are used there for a structural role: the side of the cut that makes records and acts on them. Whether anything is experienced on that side is a separate question, and Chapter 43 keeps it separate.

25.7 Audit

Canonical phenomenon

A measuring device, a genome, and the fish's choice between raising and lowering its discharge frequency: in each, continuous dynamics end in a stable alternative that stands for something and controls further dynamics.

Minimal model

Records as energy-degenerate, rate-independent structures (Definition 25.1); measurement and control as the two directions across the epistemic cut (Definitions 25.2 and 25.3); non-integrable constraints as the physical condition for records (Example 25.5); von Neumann's constructor with a description used in two ways.

What is relational here?

The relation between a record and what it records, fixed by the process that made it and the process that reads it; in the fish, the relation between the sign of Δf and the active motor pathway.

What physically reads it?

In the cell, the translation apparatus that interprets sequences; in von Neumann's automaton, the constructor A; in the fish, the push-pull prepacemaker pathways that convert the midbrain's sign-selective activity into a rise or a fall of the pacemaker frequency.

What existing field explains it?

Quantum measurement theory (Neumann 1955), the physics of information (Landauer 1991), the theory of self-reproducing automata (Neumann 1966), and Pattee's physics of symbols (Pattee 1969; Pattee 1982; Pattee 1995; Pattee 2001; Pattee and Rączaszek-Leonardi 2012). Pattee made the chapter's central move: that organisms draw epistemic cuts in their own organization.

What remains unexplained?

Pattee gives necessary conditions for a physical record (degeneracy, rate

independence, non-integrable constraints, heteropolymers) and no sufficient ones. What makes a structure function as a record for a system, rather than merely being describable as one, depends on how the system uses it, which is the question of the next chapter.

A discriminating experiment

In a system proposed to contain an internal measurement, identify the candidate record and test the two defining properties: set the record to each alternative configuration by intervention and check that the controlled dynamics follow the configuration and not the rate or energy of the setting process; then vary the reading rate and check that the effect is unchanged. A structure that fails either test is part of the dynamics, whatever it is called.

25.8 Notes and References

Pattee's early paper on how a molecule becomes a message (Pattee 1969) and his analysis of the symbol–matter problem in the cell (Pattee 1982) lead to the explicit treatment of semantic closure in Pattee (1995) and of the epistemic cut and non-integrable constraints in Pattee (2001). Pattee and Rączaszek-Leonardi (2012) collects the classic papers with commentaries by later authors, and Pattee (2015) is his late statement that the cell is the first case of a phenomenon in his structural sense. Von Neumann's analysis of the measuring process is chapter VI of Neumann (1955); his theory of self-reproduction, edited and completed by Burks, is Neumann (1966). Non-holonomic constraints are treated in standard mechanics texts, including Landau and Lifshitz (1976). Landauer (1991) argues that information is physical and that erasure has a thermodynamic cost. The two-pathway control of the jamming avoidance response is described by Metzner (1993b); the sensory side of the circuit is reviewed in Heiligenberg (1991a).

25.9 Exercises

Exercise 25.1. For each of the following, identify what is rate-dependent dynamics and what, if anything, is a record in the sense of Definition 25.1: a bimetallic thermostat; DNA replication; a punched card read by a loom; the fish's choice between raising and lowering its frequency. Where does the epistemic cut fall in each, and who or what places it?

Exercise 25.2. How many distinct nucleotide sequences of length 100 are there? Explain why the laws of chemistry do not select among them, and what does. Give one physical property that breaks the degeneracy slightly, and explain why it does not undermine the argument.

Exercise 25.3. Show that the constraints (25.1) are non-integrable by exhibiting a closed path in (x, y, ϑ) that returns the disk to its starting position and heading but changes ϕ. What does the change in ϕ record?

Exercise 25.4. Write out the three placements of the cut in Example 25.4 as three descriptions, and show that they predict the same number in the notebook. Which physical assumption about the thermometer makes the predictions agree?

Exercise 25.5. In von Neumann's scheme, suppose the copier B is removed and the constructor is asked to build, from Φ alone, a machine that already contains a description of itself. Show that this leads to a regress, and explain how copying without interpretation removes it. Which cellular process plays the role of B, and which plays the role of A?

Exercise 25.6. Compute the minimum energy $k_{\mathrm{B}} T \ln 2$ required to erase one bit at $310\,\mathrm{K}$. Compare it with the free energy released by hydrolysis of one ATP molecule, taking the latter as roughly $50\,\mathrm{kJ/mol}$ (an illustrative value), and comment on what the comparison says about the cost of resetting records in cells.

Exercise 25.7. Remark 25.6 says the fish's record is not quiescent. Propose a modification of the jamming avoidance circuit (real or hypothetical) that would make the record quiescent in Pattee's sense, and describe the behavioral consequence you would expect.

26

Semantic Closure

A radio receiver reads. It separates one carrier from many, demodulates it, and turns a relation between a received signal and a local oscillator into sound. Its reader exists because an engineer designed it, and its readings serve the purposes of the people who listen. A cell also reads: it senses nutrients, compares concentrations over time, and translates nucleic acid sequences into proteins. Its readers are built by the cell, from records the cell carries, and its readings feed the processes that keep the cell, and its readers, in existence. Pattee called this arrangement *semantic closure* (Pattee 1995).

The thesis of this chapter is that semantic closure is the organizational condition under which a system's readings can be said to be *for* the system: the reader is constructed and maintained by the system whose behavior it regulates, and what it reads contributes to that construction and maintenance. The condition has precise formulations in three traditions, Pattee's physics of symbols, Rosen's closure to efficient causation, and the closure of constraints, and it separates the cell and the fish from the radio even though all three read relations and act on them. It is also the point at which the words this book has been using, "reader," "code," "observer," become dangerous, and the chapter ends by saying how.

26.1 Closure

Several traditions in theoretical biology characterize living organization by some form of closure: a set of components or causes that produce and maintain one another. They differ in what is closed.

Rosen defined an organism as a system that is closed to efficient causation: every process that acts as an efficient cause within the system, every catalyst or machine that makes something happen, is itself produced within the system (Rosen 1991). A factory is not closed in this sense, because its machines are built elsewhere; a cell comes close, because its enzymes are made by processes that other enzymes catalyze.

Montévil and Mossio made the idea operational by distinguishing two causal regimes (Montévil and Mossio 2015). *Processes* are the changes that occur in

open, non-equilibrium thermodynamic conditions: reactions, transport, flows. *Constraints* are entities that act on processes while being conserved, in some respect, at the time scale at which they act. An enzyme is a constraint on the reaction it catalyzes: it changes the reaction's rate and is not consumed at that time scale.

Definition 26.1 (Closure of constraints)**.** A set of constraints $\mathcal{C}$ is *closed* if every constraint in $\mathcal{C}$ depends, for its production or maintenance, on processes constrained by other members of $\mathcal{C}$, and contributes to the production or maintenance of at least one other member of $\mathcal{C}$.

The definition, following Montévil and Mossio (2015), makes closure a relation among constraints, not among molecules, and so allows it to be identified at different scales and to mark boundaries between interacting systems. It says nothing yet about reading. A closed network of catalysts maintains itself; whether anything in it reads anything is a further question.

26.2 Self-Construction

The step from closure to semantic closure requires that some of the closed constraints be readers of records, in the sense of Chapter 25, and that the records specify how the readers are built. Von Neumann's self-reproducing automaton has this structure: a constructor builds machines from descriptions, and the description it reads includes a description of the constructor (Neumann 1966).

The cell realizes it in molecules.

Example 26.2 (The reader of the genetic code). The translation apparatus that reads messenger RNA and builds proteins is made of proteins and RNAs specified by the genome it reads. The assignment of amino acids to codons, the genetic code, is carried out by the aminoacyl-tRNA synthetases, enzymes each of which attaches a particular amino acid to the transfer RNAs that recognize particular codons. The synthetases are themselves proteins encoded in the genome and built by the apparatus they complete. The reader of the genetic records is a product of the genetic records.

Pattee took this as the defining case of semantic closure: the symbols and the machinery that interprets them construct each other (Pattee 1995; Pattee 2001).

The electric fish shows self-construction at a different scale. It builds its own emitter, its electroreceptors, the hindbrain and midbrain circuits that compare amplitude and phase, the diencephalic nuclei that choose a direction, and the motor pathways that drive the pacemaker (Part V). None of them is designed by anyone else. The jamming avoidance response appears in young fish, 1.2 to 1.6 cm long, without training (Viete and Heiligenberg 1991), so the

reader of the beat is assembled by development from the fish's own records before it is ever used.

26.3 Codes

A code, in the sense this chapter needs, is a mapping between two sets of physical structures that is implemented by a reader and not fixed by the physics of either set.

Definition 26.3 (Code)**.** A *code* is a mapping κ from a set of records $\mathcal{S}$ to a set of effects $\mathcal{E}$ such that (i) the mapping is carried out by a physical reader, and (ii) other mappings $\kappa' \neq \kappa$ are physically possible with the same records and the same effects, given a differently built reader.

Condition (ii) is Pattee's arbitrariness (Chapter 25). It is what distinguishes a code from a law: gravity does not map masses to accelerations by a mapping that could have been otherwise, while a synthetase maps a codon to an amino acid by a mapping that a different synthetase would change. The genetic code satisfies both conditions. So does the fish's mapping from the sign of Δf to the direction of its frequency shift: an animal whose midbrain wiring were mirrored would read the same modulations and shift toward its neighbor's frequency, making the jamming worse. That the actual mapping is the adaptive one is a fact about history and selection, not about the physics of fields and neurons.

26.4 Interpretation

The three ingredients (closure, self-construction, codes) combine into a definition.

Definition 26.4 (Semantic closure)**.** A system S exhibits *semantic closure* if

1. S contains records and at least one reader that implements a code (Definition 26.3) from those records, or from relations it measures, to effects;
2. the constraints that constitute the reader are produced and maintained by processes within S that are themselves controlled, directly or indirectly, by records read by readers of S; and
3. the effects of the reader's readings contribute to maintaining the closed set of constraints of S (Definition 26.1), including the reader itself.

Figure 26.1 draws the loop: the system constructs a reader, the reader reads the world or the system's own records, the reading controls action, and the action maintains the system and, through it, the reader.

The definition answers a question that Chapter 17 left open. There, the fish's circuit fixed the mapping from physical states to the computational

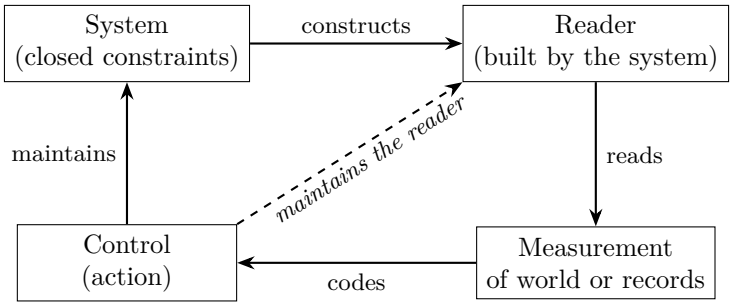


Figure 26.1: Semantic closure. The reader is built by the system it serves, its readings control action, and the action maintains the system and the reader. An engineered receiver has the reading, coding, and control steps without the arrows that construct and maintain.

distinction by what the states did downstream, and an adaptive radio did the same. The difference between them is in conditions (2) and (3). The radio's reader is produced by an external manufacturing process, and nothing the radio reads contributes to maintaining the radio; if its frequency-hopping logic fails, the radio does not repair it. The fish's reader is produced by the fish's own development, and its readings protect the electrolocation on which the fish's foraging and navigation depend (Heiligenberg 1991a).

Proposition 26.5 (Receivers are open)**.** *An engineered receiver whose components are manufactured outside it and whose readings do not contribute to the production or maintenance of its own components fails conditions (2) and (3) of Definition 26.4, whatever its input-output function.*

Proof. Condition (2) requires that the reader's constraints be produced by processes within the system; by hypothesis they are produced outside it. Condition (3) requires that the reader's effects contribute to maintaining the system's constraints; by hypothesis they do not. □

The proposition is definitional, and its interest is in what it rules out: no amount of sophistication in what a receiver reads, or in how it acts on its readings, meets the conditions, because the conditions concern where the reader comes from and what its reading maintains. Table 26.1 applies the three conditions to the three systems of the chapter's opening.

Closure also supplies norms. On the organizational account of Mossio et al. (2009), a trait has a function when it contributes to the maintenance of a closed organization that in turn contributes to producing the trait. Function so defined is grounded in the system's current organization, and so is malfunction: a reading can be correct or incorrect relative to what the organization needs. This is how Chapter 30 will make sense of the fish's phantom jamming, in which the circuit shifts the discharge frequency in response to cues that mimic a neighbor when no neighbor is there (Carlson and Kawasaki 2007).

Table 26.1: The conditions of semantic closure applied to three readers.

Condition	Adaptive radio	Cell	Electric fish
(1) Reader implements a code	Yes	Yes	Yes
(2) Reader built by processes within the system, controlled by its records	No	Yes	Yes
(3) Readings contribute to maintaining the system and the reader	No	Yes	Yes, through the organism
Origin of the code's specific mapping	Design	Descent and selection	Descent and selection

26.5 The Cell as Minimal Observer?

Pattee argued that the first memory-based, self-replicating unit, such as a cell, is the most primitive case of a necessary symbol–matter distinction, and that sensing, detection, and measurement all require the subject–object distinction he called the epistemic cut (Pattee 2015). In the same late paper he held that what is detected is determined by the organism and is potentially functional for it, and that the concept of function has meaning only for living organisms.

Two claims are bundled here and should be separated. The organizational claim is that semantic closure first appears with cells: a cell builds its readers from its records, reads, and uses its readings for its own maintenance, and nothing simpler known to biology does all three. That claim is well supported and is the one this chapter adopts. Bacterial chemotaxis is a clear example of the cell as reader: a bacterium compares the concentration of an attractant now with its concentration a few seconds earlier, a temporal relation, and modulates its swimming accordingly (Macnab and Koshland 1972; Berg and D. A. Brown 1972); the receptors and signaling proteins that perform the comparison are encoded and built by the cell. Chapters 28 and 49 return to it.

The second claim is carried by the vocabulary: "observer," "subject," "phenomenon." Pattee uses these words for a structural role, the side of the epistemic cut that makes records and acts on them. Read that way, the cell is a minimal observer by definition, and nothing follows about whether anything is experienced by it. Read any other way, the claim goes beyond what the organizational evidence shows.

26.6 Where the Metaphor Becomes Dangerous

This book has used "reader" deliberately, as a term for a physical operation that makes a relation dynamically explicit. Semantic closure adds "code,"

"interpretation," and "meaning." Each of these words comes from human practices of reading, and each carries features of those practices that the organizational definitions do not include. Three dangers follow.

Inflation. If every feedback loop counts as reading and every regularity as a code, the words lose their content. A thermostat has a sensor, a switch, and an effect on the temperature it senses, and it has no semantic closure: its sensor is manufactured outside it, and nothing it reads maintains the thermostat. Definition 26.4 blocks inflation by its second and third conditions, and the test of any claim that a system reads should be whether those conditions can be checked.

Phenomenological smuggling. "Observer" and "subject" suggest a point of view, and "meaning" suggests understanding. Semantic closure establishes neither. It establishes that some of a system's physical states have effects that are for the system, in the sense that they contribute to its maintenance, and that the mapping from states to effects is implemented by readers the system builds. Whether such a system has any perspective in the sense of Chapter 43 is a separate question that the definition leaves exactly where it was.

The homunculus. If the cell reads its genome, what reads the reader? The regress that threatened Chapter 18 returns. Semantic closure answers it by closing the loop instead of adding a level: the reader of the genome is itself specified by the genome, and its correct functioning is maintained by the same organization it serves, so no further reader is required for the system to go on. The answer removes the regress in the organizational sense. It does not show that the closed loop *understands* anything, and it should not be taken to.

The discipline the chapter recommends is simple to state. Whenever a text, including this one, says that a system reads, interprets, or means, it should be possible to cash the claim out in terms of the three conditions of Definition 26.4: which reader, built by what, maintaining what. Where the claim cannot be cashed out that way, the word is doing work the evidence does not support.

26.7 Audit

Canonical phenomenon

The cell's translation apparatus, which reads genetic records and is itself specified by them; the fish's jamming avoidance circuit, built by development and serving the fish's electrolocation; the radio, which reads without either property.

Minimal model

Closure of constraints (Definition 26.1); codes as reader-implemented mappings that could have been otherwise (Definition 26.3); semantic closure as the conjunction of reading, self-construction of the reader, and self-

maintenance through reading (Definition 26.4); the openness of engineered receivers (Proposition 26.5).

What is relational here?
The relation between a record (or a measured relation) and its effect, fixed by a reader; and the relation between the reader and the organization that builds it, which closes the loop.

What physically reads it?
In the cell, the translation apparatus and the aminoacyl-tRNA synthetases that implement the genetic code; in the fish, the sensory and motor circuit of the jamming avoidance response; in the radio, its demodulator, which reads but is not part of a closed organization.

What existing field explains it?
Pattee's semantic closure (Pattee 1995; Pattee 2001; Pattee 2015), Rosen's closure to efficient causation (Rosen 1991), the closure of constraints and the organizational account of function (Montévil and Mossio 2015; Mossio et al. 2009), and von Neumann's self-reproducing automata (Neumann 1966). The chapter adds only the application of these criteria to the fish and the radio.

What remains unexplained?
Semantic closure explains how readings can be for a system and how they can be correct or incorrect. It does not explain why a closed reader should have a point of view; the organizational and the phenomenal senses of "observer" remain distinct.

A discriminating experiment
In a synthetic system, build two versions of the same sensor-controller loop: one whose sensor is supplied from outside, and one whose sensor is produced and repaired by the system's own processes under control of its own records. Damage the sensor in both. The closed version should restore its reading and the functions that depend on it; the open version should not. Any claimed difference in what the two systems "read" beyond that repair behavior would need independent evidence.

26.8 Notes and References

Semantic closure is introduced in Pattee (1995) and developed in Pattee (2001); the collected papers with commentary are Pattee and Rączaszek-Leonardi (2012), and Pattee (2015) is Pattee's late statement on the cell as the first case. Rosen (1991) develops the idea of an organism as a system closed to efficient causation, through his theory of metabolism–repair systems. Montévil and Mossio (2015) give the closure-of-constraints formulation, and Mossio et al. (2009) build on it an organizational account of biological function that grounds norms in current organization rather than in selection history alone;

Moreno and Mossio (2015) develop the full framework of biological autonomy. Von Neumann's constructor with a self-describing tape is in Neumann (1966). Bacterial temporal comparison is established by Macnab and Koshland (1972); Berg and D. A. Brown (1972) tracked individual bacteria in three dimensions. The two-pathway control of the jamming avoidance response and its innate development are in Metzner (1993b) and Viete and Heiligenberg (1991).

26.9 Exercises

Exercise 26.1. Apply the three conditions of Definition 26.4 to each of the following, stating which conditions hold and why: a thermostat; a virus inside a host cell; a virus outside any cell; a bacterium performing chemotaxis; a self-driving car that schedules its own maintenance at a service station.

Exercise 26.2. Show that the fish's mapping from the sign of Δf to the direction of its frequency shift satisfies Definition 26.3. Describe the reversed mapping, and use a qualitative argument (or a simple model of two fish that each adjust their frequency) to show what would happen to two fish that both used it.

Exercise 26.3. Consider a minimal chemical network in which catalyst K_1 catalyzes the formation of K_2 from precursors, and K_2 catalyzes the formation of K_1. Is the set $\{K_1, K_2\}$ closed in the sense of Definition 26.1? Add a membrane component whose formation is catalyzed by K_1 and which keeps both catalysts from diffusing away. How does the closed set change? Does anything in the network read anything?

Exercise 26.4. State the homunculus regress for the genetic code as an infinite sequence of readers. Show how semantic closure stops the sequence, and say precisely which feature of the translation apparatus does the stopping.

Exercise 26.5. Using the organizational account of function (Mossio et al. 2009), decide whether each of the following is a malfunction, and for which system: a fish shifting its frequency in response to a phantom jamming stimulus; a radio hopping away from a channel that was in fact free; a synthetase attaching the wrong amino acid to a transfer RNA.

Exercise 26.6. Design the experiment in the audit's discriminating test for a concrete synthetic system (for example, an engineered bacterium with a sensor protein whose expression is controlled by the sensor's own output). Specify the damage, the measurements, and the result that would count against the claim that semantic closure makes a difference to the system's reading.

27

Autopoiesis

Semantic closure asks where a system's readers come from. Autopoiesis asks a prior question: what is the system for which anything can be read? Humberto Maturana and Francisco Varela proposed that what all living systems share, whatever their chemistry, is a kind of organization: a network of processes that produces the components that realize the network, including a boundary that separates the network from its surroundings (F. Varela, H. Maturana, et al. 1974; H. R. Maturana and F. J. Varela 1980). They called it *autopoiesis*, self-production.

The thesis of this chapter is that autopoiesis gives a precise account of self-production and boundary creation, and with it an answer to the question "for whom is a reading made?": for the unity that produces itself. It also explains why the environment of a living system is always a selected environment, an *Umwelt* of the differences that the system's own organization lets it register. What autopoiesis does not explain is why some of those differences matter more than others, or how an organism can do better or worse at staying alive. That gap is why its successors added adaptivity, closure of constraints, and interactive agency, and why Chapter 28 follows.

27.1 Organizational Closure

Maturana and Varela distinguished the *organization* of a system from its *structure* (H. R. Maturana and F. J. Varela 1980). The organization is the set of relations among components that makes the system a member of its class: what any system must preserve to remain the kind of thing it is. The structure is the actual components and relations that realize the organization in a particular case.

Definition 27.1 (Organization and structure)**.** The *organization* of a system is the set of relations among its components that must hold for the system to belong to a given class. Its *structure* is the particular set of components and relations that realizes the organization at a given time.

The distinction lets structure change while organization persists. A bacterium replaces most of its molecules over its lifetime and remains a bacterium; a clock can be rebuilt with new gears and remain a clock. It also makes organization a relational notion in the sense of Chapter 1: an organization is a relation among components, defined on no component alone, and it can be realized by many different sets of components.

A system is *organizationally closed* when the processes that constitute it are conditions for one another, so that the network of processes, taken as a whole, regenerates the conditions of its own operation. Autopoiesis is the case in which the processes are productions of physical components, and the network produces, among other things, the boundary that makes it a unity in space.

27.2 Self-Production

Definition 27.2 (Autopoietic system)**.** A system is *autopoietic* if it is a network of processes of production and transformation of components such that (i) the components, through their interactions, continuously regenerate the network of processes that produced them, and (ii) the components constitute the network as a concrete unity in the space in which they exist, by producing the boundary that delimits it.

The definition paraphrases H. R. Maturana and F. J. Varela (1980); F. Varela, H. Maturana, et al. (1974) turned it into a procedure for deciding whether a given unity is autopoietic. In outline, one asks in turn:

1. whether the unity has identifiable boundaries;
2. whether it has constitutive components;
3. whether it is a mechanistic system, whose components' properties determine its interactions and transformations;
4. whether the components of the boundary constitute the boundary through their own preferential interactions;
5. whether the boundary components are produced by interactions among the unity's components; and
6. whether all other components are produced in the same way, or participate as necessary permanent constituents in producing the others.

A unity that passes all six tests is autopoietic in the space in which its components exist.

Two features of the definition deserve emphasis. First, reproduction is absent. Maturana, Varela, and Uribe insisted that a living system is a whole, autonomous unity that is alive whether or not it reproduces, and they criticized contemporary biology for running together history-dependent processes (evolution, development) with the history-independent organization of the individual (F. Varela, H. Maturana, et al. 1974). Second, nothing in the definition

mentions information, codes, or reading. An autopoietic system produces itself; whether it reads anything is not part of the criterion.

27.3 Boundary Creation

The boundary is where autopoiesis differs from other forms of closure. A closed network of reactions in a well-mixed vessel has no boundary of its own; the vessel supplies one. An autopoietic network produces the components that hold it together.

Model 27.3 (The 1974 computational model)**.** F. Varela, H. Maturana, et al. illustrated the definition with a simulation on a two-dimensional grid (F. Varela, H. Maturana, et al. 1974). The space contains three kinds of element: substrate particles S, a catalyst K, and links L. Three processes act on them:

$$\mathrm{K} + 2\,\mathrm{S} \longrightarrow \mathrm{K} + \mathrm{L} \qquad \text{(production, catalyzed)} \tag{27.1}$$

$$\mathrm{L} + \mathrm{L} \longrightarrow \mathrm{L}{-}\mathrm{L} \qquad \text{(bonding of links into chains)} \tag{27.2}$$

$$\mathrm{L} \longrightarrow 2\,\mathrm{S} \qquad \text{(spontaneous disintegration)} \tag{27.3}$$

Links are produced near the catalyst, bond into chains, and eventually close into a ring that encloses the catalyst. Substrate passes through the ring, while links do not, so production continues inside. When a link in the ring disintegrates, the gap is filled by links produced inside, and the ring is repaired.

The model is a minimal case of Definition 27.2. The ring is produced by the network inside it, and it keeps the network's products concentrated near the catalyst so that production can continue. Neither the ring nor the reactions alone are autopoietic; the organization is the relation between them.

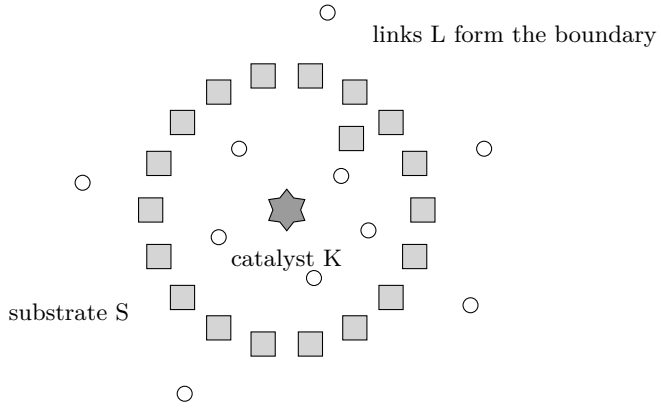


Figure 27.1: The computational model of autopoiesis (schematic). A catalyst produces links from substrate; links bond into a ring that encloses the catalyst; substrate crosses the ring while links do not; decaying links are replaced from inside. After F. Varela, H. Maturana, et al. (1974).

Real cells do the same with lipid membranes, transporters, and the metabolic networks that produce both. Luisi and colleagues showed that much simpler chemical systems can produce their own boundaries: they built micelles that catalyze the formation of more of their own surfactant from a precursor, so that the population of micelles grows autocatalytically (Bachmann et al. 1992), and they placed such systems within a program of *chemical autopoiesis* aimed at minimal models of cellular life (Luisi 2003).

Example 27.4 (Self-production without a reader). A population of autocatalytic micelles produces the molecules that constitute its own boundaries. It carries no records in the sense of Chapter 25: there is no sequence whose alternatives are energy-degenerate and whose configuration specifies how components are built. It therefore illustrates self-production of a boundary without semantic closure. Autopoiesis and semantic closure are distinct properties, and a system can have the first without the second.

27.4 Operational Autonomy

An autopoietic system is closed in its organization and open in its matter and energy: substrate enters, waste leaves, and energy flows through. Maturana and Varela described its relation to the environment as *structural coupling*. The environment perturbs the system, and the system changes its structure in response, but which changes a perturbation triggers is determined by the system's own structure, not by the perturbation (H. R. Maturana and F. J. Varela 1980). In Luisi's summary, the perturbations brought about by the environment are changes selected and triggered by the inner organization of the living system (Luisi 2003).

Definition 27.5 (Structural coupling)**.** A system and its environment are *structurally coupled* when their recurrent interactions produce congruent changes of structure in each, with the changes in the system determined by the system's structure and only triggered by the environment.

In the vocabulary of this book, structural coupling says that D belongs to the system. A perturbation is a mixture $y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})$ arriving at the system's boundary; what becomes a variable for the system is $z = \mathsf{D}(y)$, and D is part of the system's structure. Two systems in the same medium, receiving physically identical mixtures, can make different variables explicit because their readers differ. This is the claim of Chapter 16, stated fifty years earlier for living systems in general.

Maturana and Varela drew a further conclusion: that living systems are cognitive systems, and that living, as a process, is a process of cognition (H. R. Maturana and F. J. Varela 1980). Luisi, a sympathetic reviewer, called "cognition" an arguable term for the selective coupling it names, and reported the resulting view that autopoiesis and cognition are complementary and, in a

way, equivalent aspects of life (Luisi 2003). The enactive program developed the idea into a theory of mind in which cognition is the enactment of a world through sensorimotor coupling (F. J. Varela et al. 1991), and Thompson (2007) defended a strong continuity between life and mind on that basis. The identification of life with cognition is a substantive thesis, contested, and this book does not need it; what it needs is the weaker and well-supported claim that the reader belongs to the system.

27.5 Environment Versus Umwelt

Jakob von Uexküll had made the same point about animals decades before autopoiesis, in a vocabulary that has proved durable. The *Umwelt* of an animal is the part of its surroundings that its receptors and effectors make available to it, a world of perceptual and effector cues that differs from species to species even when the physical environment is shared (Uexküll 2010). His best-known example is the tick, whose world, in his description, is built from a small number of cues: the odor of butyric acid from mammalian skin, the touch of hair, and the warmth of skin. The forest in which the tick waits is full of other physical differences, and none of them is part of its Umwelt.

Definition 27.6 (Umwelt)**.** The *Umwelt* of a system is the set of environmental differences that its readers can make dynamically explicit and that its effectors can act on.

In these terms, the Umwelt is fixed by D and F together. It is relational twice over: it is a relation between organism and environment, and many of its members are themselves relations. The Umwelt of *Eigenmannia* includes the distortions of its own field by nearby objects, and the beat between its own discharge and a neighbor's, including the sign of the frequency difference (Part V). A human swimming in the same water has none of these in its Umwelt, although every one of them is physically present at the human's skin.

The Umwelt also varies between close relatives. *Sternopygus*, a gymnotiform related to *Eigenmannia*, has midbrain neurons that discriminate the sign of the difference frequency between interfering discharges, yet it performs no jamming avoidance response; Rose, Keller and Heiligenberg suggested that in *Sternopygus* these neurons serve the detection of moving objects and conspecifics (G. Rose, C. Keller, et al. 1987). The same relation is read in both fish and enters their Umwelten with different consequences for action.

27.6 Criticisms

Autopoiesis has been criticized from within the tradition that it founded and from outside it. Five criticisms bear on this book.

All or nothing. A system either is autopoietic or it is not; the definition has no degrees. Di Paolo (2005) argued that this makes autopoiesis too coarse to ground sense-making: a system that merely maintains its organization, as long as it can, has no way of doing better or worse, and no basis for treating one environmental difference as more significant than another. He proposed adding *adaptivity*, a system's capacity to regulate its own states and its interactions with respect to the boundaries of its viability, as the property that turns self-production into sense-making. Chapter 28 develops the proposal.

Life and cognition. The identification of life with cognition, discussed above, stretches "cognition" to cover every autopoietic system, and critics have found the stretch unhelpful (Luisi 2003). The question it answers by definition, whether a bacterium cognizes, is better left as a question about specific capacities.

Reproduction and evolution. By excluding reproduction from the definition, autopoiesis sets aside heredity, and with it the open-ended evolution that Pattee and von Neumann took to require records copied without being read (Chapter 25). An autopoietic system without records can persist; it cannot accumulate the kind of heritable variation that builds new readers. The micelles of Example 27.4 are the illustration.

Closure and openness. Autopoietic systems are organizationally closed and thermodynamically open, and early presentations were sometimes read as denying the second. The biological autonomy framework of Moreno and Mossio (2015) integrates the two: constitutive closure (the closure of constraints of Chapter 26) together with interactive autonomy, the regulated exchange with the environment through which an organism acts. Barandiaran et al. (2009) give a companion definition of agency in terms of individuality, normativity, interactional asymmetry, and spatiotemporal organization.

Teleology. Weber and F. J. Varela (2002) argued that autopoiesis grounds a form of intrinsic teleology, in the spirit of Kant's natural purposes: an autopoietic system has a concern for its own continuation that no machine built for external purposes has. Hans Jonas had described the same feature as the organism's needful freedom, the dependence on continual exchange with the world that gives the organism both its independence of any particular matter and its precariousness (Jonas 1966). Whether that concern is more than a description of self-maintenance is precisely the question the phrase "for the system" raises, and the book returns to it in Part X.

27.7 Audit

Canonical phenomenon

A cell that replaces its components while remaining the same cell, and produces the membrane that holds its metabolism together; in simulation, a catalyst that builds a ring of links around itself and repairs it.

Minimal model

Organization and structure (Definition 27.1); autopoiesis as self-production of components including a boundary (Definition 27.2); the 1974 computational model (Model 27.3); structural coupling and the Umwelt (Definitions 27.5 and 27.6).

What is relational here?

The organization itself, a relation among components defined on none of them; the Umwelt, a relation between organism and environment whose members are often relations.

What physically reads it?

Autopoiesis makes the reader part of the system's structure: a perturbation becomes a variable for the system only through the system's own receptors and processes. The criterion itself contains no reader, and a system can be autopoietic without reading anything.

What existing field explains it?

The autopoietic tradition (F. Varela, H. Maturana, et al. 1974; H. R. Maturana and F. J. Varela 1980; F. J. Varela et al. 1991; Thompson 2007), von Uexküll's Umwelt (Uexküll 2010), chemical autopoiesis (Bachmann et al. 1992; Luisi 2003), and the biological autonomy school (Moreno and Mossio 2015; Barandiaran et al. 2009; Di Paolo 2005; Weber and F. J. Varela 2002).

What remains unexplained?

Autopoiesis says for whom a reading is made. It does not say why a particular reading matters, how an organism can do better or worse, or whether the self it produces has a point of view. The first two are the subject of Chapter 28; the third remains open.

A discriminating experiment

Compare two minimal chemical systems that produce their own boundaries, one with and one without a regulatory loop that changes production rates in response to a sensed environmental variable. Perturb the environment toward the edge of viability. Definition 27.2 predicts no difference in kind between them; the adaptivity proposal predicts that only the regulated system will adjust before its organization fails.

27.8 Notes and References

The foundational statements are H. R. Maturana and F. J. Varela (1980), which includes the English version of Maturana and Varela's original essay on autopoiesis, and F. Varela, H. Maturana, et al. (1974), which gives the six-point characterization and the computational model. Luisi (2003) is a critical review by a chemist who worked with Varela, with a survey of chemical autopoiesis; Bachmann et al. (1992) report autocatalytic self-replicating micelles. The

enactive development is F. J. Varela et al. (1991) and Thompson (2007); Weber and F. J. Varela (2002) connect autopoiesis to Kant's natural purposes. Di Paolo (2005) introduces adaptivity; Barandiaran et al. (2009) define agency; Moreno and Mossio (2015) develop biological autonomy as closure of constraints plus interactive agency. Jonas (1966) is the classic philosophical biology of the organism's needful freedom. Von Uexküll's account of the Umwelt, first published in 1934, is available in Uexküll (2010). G. Rose, C. Keller, et al. (1987) report sign-selective midbrain neurons in *Sternopygus*, which lacks a jamming avoidance response.

27.9 Exercises

Exercise 27.1. Apply the six questions of Section 27.2 to a candle flame, a hurricane, a growing crystal, a bacterium, and a virus particle outside a cell. Which pass, which fail, and at which question?

Exercise 27.2. Implement a simplified version of Model 27.3 on a 30×30 grid with random walks for substrate and links, production within one cell of the catalyst, bonding between adjacent links, and disintegration with probability 10^{-3} per link per step (an illustrative value). Report how often a closed ring forms, how long it persists, and what happens when you set the disintegration probability to zero or make links permeable.

Exercise 27.3. Give two different structures that realize the same organization, and one structural change that destroys the organization while leaving most components in place. Use a cell, a clock, or an institution as your example.

Exercise 27.4. List five physical differences present at the skin of a weakly electric fish and at the skin of a human swimmer in the same water. Mark which belong to each one's Umwelt (Definition 27.6), and for each, name the reader that makes it explicit.

Exercise 27.5. Model the number N of autocatalytic micelles with a finite supply P of precursor: $\dot{N} = kNP$ and $\dot{P} = -\alpha kNP$. Solve for $N(t)$ and interpret the parameters. Is the system autopoietic by Definition 27.2? Does it read anything?

Exercise 27.6. Describe a system that is autopoietic but not adaptive in Di Paolo's sense, and an observation that would distinguish it from an adaptive one near the edge of its viability.

Exercise 27.7. Does the jamming avoidance response require its bearer to be autopoietic? Argue both sides: first that an engineered, non-autopoietic device could perform the same computation (Chapter 17); then that the response serves a self for which the jamming of electrolocation is a cost. Which argument concerns the reader and which concerns the reading's significance?

28

Sense-Making

A radio receiver can read the beat between two carriers as well as a fish can, and nothing that happens to the radio is better or worse for the radio. A bacterium swimming in a sugar gradient reads far less, and some of what happens to it is better for it than the alternatives, in a sense that does not depend on anyone's opinion. The difference lies in what the reading is *for*. The bacterium's readings feed the processes that keep the bacterium in existence, and those processes would fail without them.

The thesis of this chapter is that this difference can be stated physically. A distinction becomes information *for a system* when the system's continued existence depends on how it responds to that distinction; the resulting norms are relations between an organization and its environment; and they require no consciousness. The enactivist tradition calls the activity that establishes such norms *sense-making*. We adopt the phenomenon, state it in the book's formal terms, and use bacterial chemotaxis as the minimal case, a case in which every part of the loop has a known molecular implementation.

28.1 Viability

Every living system has variables that must stay within bounds for the system to go on existing: temperature, osmotic balance, energy charge, membrane integrity. Ashby called them *essential variables* and treated the maintenance of their bounds as the defining task of biological regulation (Ashby 1956). The set of states in which all of them are within bounds is the system's viability set.

Definition 28.1 (Viability set)**.** Let a system's state be $s_{\mathrm{A}} \in S$ and let s_{A} evolve by $\dot{s}_{\mathrm{A}} = G(s_{\mathrm{A}}, s_{\mathrm{E}}, a)$, where s_{E} is the environment's state and a the system's action. The *viability set* $V \subset S$ is the set of states in which the system continues to exist as the kind of system it is. The *exit time* from an initial state $s_0 \in V$ under environment history $s_{\mathrm{E}}(\cdot)$ and policy π is

$$T_V = \inf\{\, t > 0 : s_{\mathrm{A}}(t) \notin V \,\}.$$

A rock also has a set of states in which it continues to be a rock, and its dynamics do nothing to keep it there. Living systems differ in a specific way: their existence is *precarious*. Left without their own activity, they leave V, and they remain in it only because what they do keeps them there (Di Paolo 2005; Barandiaran et al. 2009).

Definition 28.2 (Precarious self-maintenance)**.** A system is *precariously self-maintaining* if, for the environments it normally inhabits, T_V is finite when its own activity is suppressed and is much longer when its activity proceeds, and if the processes that keep it in V are themselves among the states that V protects.

The last clause is the one that matters. A thermostat keeps a room within bounds, but the thermostat's own existence does not depend on the room's temperature. A cell keeps its internal chemistry within bounds, and the machinery that does the keeping is made of that chemistry. This is the closure studied in Chapter 27 and Chapter 26, seen from the side of its consequences: once a system's persistence depends on its own activity, outcomes can be ranked by what they do to that persistence (Jonas 1966; Moreno and Mossio 2015).

28.2 Normativity

A ranking of outcomes by their effect on a system's persistence is a norm, and it is a norm that belongs to the system.

Definition 28.3 (System-relative value)**.** For a precariously self-maintaining system with viability set V and policy π, the *value* of an environmental condition e is

$$v(e) = \mathbb{E}\big[\, T_V \mid s_{\mathrm{E}} = e,\ \pi \,\big],$$

or, where survival over a fixed horizon τ is the relevant quantity, $v(e) = \mathbb{P}(T_V > \tau \mid s_{\mathrm{E}} = e, \pi)$. Condition e is *better for the system* than e' when $v(e) > v(e')$.

Three features of the definition deserve comment. First, v is a relational quantity in the sense of Chapter 1. It is defined on the pair (organization, environment) and on neither alone: the same sugar concentration has one value for a bacterium that can metabolize the sugar and another for one that cannot. Second, v is observer-independent in the sense that matters. The observer chooses to compute it, but the observer does not choose V; the boundary of viability is fixed by the system's organization, and crossing it has consequences whether anyone is watching or not. Third, nothing in the definition refers to experience. The ranking is a fact about dynamics, survival, and organization.

The organizational account of biological function makes the same move for parts of organisms. A trait has a function when it contributes to maintaining

the organization that produces and maintains it (Mossio et al. 2009). On that account the flagellar motor has the function of moving the cell because the moving contributes to the persistence of the organization that builds flagellar motors, and a malfunction is a failure relative to that contribution. Function and value are two faces of one relation between a closed organization and the conditions it must meet.

28.3 Adaptive Regulation

Self-maintenance by itself is all or nothing: the system is in V or it is not. Regulation that responds to the environment needs more. The system must be sensitive to where it is heading relative to the boundary of V, and able to change heading.

Definition 28.4 (Adaptivity)**.** Following Di Paolo (2005), a precariously self-maintaining system is *adaptive* if, in some circumstances, it regulates its states and its relation to the environment so that, when its state is sufficiently close to the boundary of V, (i) trajectories that approach the boundary are distinguished from trajectories that recede from it and are acted upon differently, and (ii) as a result, approaching trajectories are turned into receding ones.

Adaptivity is where reading enters the account of normativity. To distinguish approaching trajectories from receding ones, the system must be sensitive to a variable that covaries with the sign of $\dot{v}$, the rate at which its situation is improving or deteriorating. The variable is usually a proxy: a nutrient concentration, a temperature change, a field distortion. The system reads the proxy, and the reading is valuable to the extent that the proxy tracks the sign of $\dot{v}$ in the environments the system inhabits.

We can make the notion of mattering quantitative.

Definition 28.5 (Salience as value of information)**.** Let z be a variable the system could read, and let π_z and $\pi_\varnothing$ be the best available policies that do and do not use z. The *salience* of z for the system is

$$\sigma(z) = \mathbb{E}[v \mid \pi_z] - \mathbb{E}[v \mid \pi_\varnothing].$$

Salience is the decision-theoretic value of information, computed with the system's own value function. It is zero for distinctions that make no difference to how long the system lasts, however large their physical magnitude, and it can be large for tiny ones. Von Uexküll called the subset of the world that is present to an animal its *Umwelt* (Uexküll 2010); in the terms of this chapter, the Umwelt is the set of variables with non-zero salience that the animal has readers for. The tick's Umwelt, in his most famous example, contains butyric acid, warmth, and hair, and little else. A weakly electric fish's Umwelt contains distortions of its own electric field and the beats imposed on that field by neighbors (Part V).

28.4 The Minimal Case: Bacterial Chemotaxis

Escherichia coli swims by rotating its flagella. When the motors turn one way the flagella bundle and the cell runs in a roughly straight line; when they reverse, the bundle flies apart and the cell tumbles, reorienting at random. Adler's capillary assays established that the cells gather where attractants are concentrated (J. Adler 1966). Berg and Brown tracked single cells in three dimensions and found how: runs that carry a cell up a gradient of attractant are extended, so the random walk acquires a bias (Berg and D. A. Brown 1972).

The cell is only a few micrometers long, and one might expect it to compare concentrations at its two ends. Macnab and Koshland showed that it need not. When they mixed cells rapidly into a uniform solution with a higher attractant concentration, tumbling was suppressed for a while and then returned to normal; a sudden decrease had the opposite effect (Macnab and Koshland 1972). The cell compares the concentration it senses now with the concentration it sensed a few moments earlier. The variable it reads is a relation between the present and the recent past.

Model 28.6 (Chemotaxis as temporal comparison)**.** Let $c(t) = c(\boldsymbol{x}(t))$ be the attractant concentration at the cell's position. A slow internal variable $m(t)$ tracks c,

$$\tau\, \dot{m} = c - m, \tag{28.1}$$

and the tumbling rate depends on the difference $u = c - m$,

$$\lambda(t) = \lambda_0 \, \exp\!\big(-g\, u(t)\big), \qquad g > 0. \tag{28.2}$$

While the cell runs, $\boldsymbol{x}$ moves with constant speed along its current heading; at rate λ it tumbles and draws a new heading at random.

The model is a caricature of the molecular mechanism, and it keeps the features that matter here. In the real circuit, receptor clusters bound to attractant suppress a kinase that promotes tumbling, and a slow methylation system resets the receptors' sensitivity so that the kinase returns to its baseline activity at any constant ligand level (Barkai and Leibler 1997). The return to baseline is exact over a wide range of concentrations and robust to large changes in protein levels (Alon et al. 1999), and it is the signature of integral feedback: the methylation level integrates the kinase's deviation from baseline (Yi et al. 2000). Equation (28.1) is the simplest integrator with that property. For a step increase Δc at $t = 0$, $u(t) = \Delta c\, \mathrm{e}^{-t/\tau}$; tumbling drops and recovers (Figure 28.1). For a slow ramp $c(t) = c_0 + kt$, u settles at $k\tau$: the cell's reading approximates $\tau\, \mathrm{d}c/\mathrm{d}t$.

In the notation of the reading loop, every term is present. The mixture M combines the cell's position with the environment's concentration field to produce the receptor occupancy. The reader D is the receptor, kinase, and

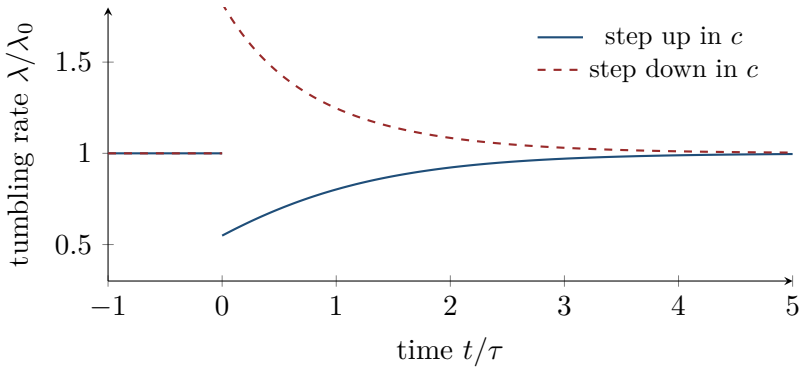


Figure 28.1: Illustrative response of Model 28.6 to a step in attractant concentration at $t = 0$ ($g\Delta c = 0.6$, arbitrary units). The tumbling rate responds to the change and returns exactly to baseline: the cell reads $c(t) - m(t)$, a relation between the present and the remembered past, and not c itself.

methylation complex, which makes $u = c - m$ dynamically explicit. The policy π is the dependence of motor reversals on kinase activity. The intervention F is swimming, which changes $\boldsymbol{x}$ and therefore the next value of c. The cell participates in producing the relation it reads: whether c is rising depends on which way the cell happens to be swimming.

The case is minimal in a second sense. The cell's reading is valuable because attractants are, in the environments the cell's ancestors inhabited, mostly nutrients, so that a rising c usually means a rising v. The reader tracks the proxy, and the proxy is tied to value only statistically.

Example 28.7 (A reader of a proxy). Adler found that some chemicals the cell metabolizes extensively do not attract it, that some it cannot metabolize do attract it, and that mutants unable to respond to an attractant can still metabolize it: metabolism of an attractant is neither required nor sufficient for chemotaxis toward it (J. Adler 1969). The receptor responds to molecular identity. A different architecture ties behavior to value more directly: in *metabolism-based* chemotaxis, motility is modulated by the cell's own metabolic state, and there is evidence that several bacterial species combine both mechanisms (Egbert et al. 2010).

The contrast between a reader of a proxy and a reader of the system's own condition returns in Chapter 29, where it becomes the question of what a reading is about, and in Chapter 30, where it becomes the question of when a reading is wrong.

28.5 Sense-Making and Enactivism

The enactive approach to cognition, introduced by Varela, Thompson, and Rosch, holds that cognition is the activity by which a living system brings

forth a world of significance through its own self-maintaining activity (F. J. Varela et al. 1991). Weber and Varela grounded the idea in the autopoietic account of the living and in Kant's notion of natural purposes (Weber and F. J. Varela 2002); Di Paolo showed that autopoiesis alone yields only all-or-nothing norms and added adaptivity to obtain graded ones (Di Paolo 2005); Barandiaran, Di Paolo, and Rohde defined agency by individuality, normativity, and interactional asymmetry (Barandiaran et al. 2009); Thompson developed the whole into a philosophy of mind organized around the continuity of life and mind (Thompson 2007). Behind the tradition stands Jonas, for whom metabolism is a "needful freedom": an organism is free of any particular matter, because it continuously replaces its matter, and needful of matter in general, because it must keep replacing it (Jonas 1966).

Sense-making, in this tradition, is the organism's evaluation of its encounters in terms of their bearing on its continued existence. The definitions of this chapter reconstruct that idea in the book's vocabulary: viability sets the norm, adaptivity supplies the graded sensitivity, salience says which distinctions matter, and the reading loop says how they are made available. The reconstruction is intentionally deflationary, and it loses nothing that the evidence supports.

What enactivism contributes is an insistence the rest of this book depends on: the meaning a reading has for a system cannot be read off the reading itself. It depends on the organization that uses the reading and on what that organization must do to persist. The fish of Part V reads the sign of a frequency difference with sub-microsecond precision, and the reading matters to the fish because jamming degrades its electrolocation (Kawasaki 1996). The value of the jamming avoidance response is a value relative to another reading: it protects the fish's ability to sense its surroundings, and that ability matters because the fish must find food and avoid obstacles to stay in V.

28.6 Does Normativity Require Consciousness?

No.

The answer needs to be explicit because the vocabulary invites the opposite. Words such as "evaluation," "concern," "significance," and "sense" carry experiential connotations in ordinary speech, and enactivist writing sometimes moves between the organizational and the experiential senses without marking the move. The definitions above use none of those connotations. The viability set, the value function, adaptivity, and salience are defined by dynamics and persistence. A bacterium satisfies all of them, and no evidence suggests that a bacterium experiences anything. The norms are real, in that they are not projections of an observer, and they are silent about experience.

The distinction between intrinsic and derived norms survives the deflation. A thermostat has a norm: the setpoint. The norm is derived, because the thermostat's persistence does not depend on meeting it and the setpoint was

chosen by someone else. A bacterium's norms are intrinsic, because failing them ends the bacterium. Intrinsic normativity, on this account, requires precarious self-maintenance, and it requires nothing more mysterious than that.

A stronger thesis is available and must be kept separate. The *strong continuity thesis* holds that life and mind share a common set of organizational properties, so that the organizational properties distinctive of mind are an enriched version of those fundamental to life (Thompson 2007). If "mind" is read functionally, the thesis is plausible and this book's ladder (Part XI) is one way of spelling it out. If "mind" is read as phenomenal consciousness, the thesis is a claim about where experience begins, and nothing in this chapter bears on it. Part X explains why no current evidence does.

28.7 Audit

Canonical phenomenon

A swimming *E. coli* cell extends its runs when the attractant concentration it senses is rising, and so climbs gradients of nutrients.

Minimal model

Viability set and exit time (Definition 28.1); system-relative value $v(e)$ (Definition 28.3); salience as value of information (Definition 28.5); temporal comparison by integral feedback, $\tau\dot{m} = c - m$, $\lambda = \lambda_0 \mathrm{e}^{-g(c-m)}$ (Model 28.6).

What is relational here?

Two relations. The cell reads $u = c(t) - m(t)$, a relation between the present concentration and a remembered one. The value of any condition, $v(e)$, is a relation between the cell's organization and its environment.

What physically reads it?

The receptor clusters, the kinase they regulate, and the methylation system that adapts them, which together implement integral feedback (Barkai and Leibler 1997; Alon et al. 1999; Yi et al. 2000). The value relation has no dedicated reader in *E. coli*: the cell reads a proxy (molecular identity) whose tie to value is statistical (J. Adler 1969). In metabolism-based chemotaxis the cell's own metabolic state is read instead (Egbert et al. 2010).

What existing field explains it?

The biochemistry and physics of bacterial chemotaxis (J. Adler 1966; Berg and D. A. Brown 1972; Macnab and Koshland 1972); control theory (integral feedback and robust adaptation); Ashby's essential variables (Ashby 1956); the theory of biological autonomy and organizational function (Moreno and Mossio 2015; Mossio et al. 2009); enactivism (F. J. Varela et al. 1991; Weber and F. J. Varela 2002; Di Paolo 2005; Thompson 2007).

What remains unexplained?

Whether intrinsic normativity is anything over and above precarious self-maintenance plus adaptive regulation. We have found no reason to think so. The strong continuity thesis, read phenomenally, is an open question that no evidence in this chapter addresses.

A discriminating experiment

Dissociate the proxy from the value. Present bacteria with a gradient of a chemical that binds the attractant receptor but cannot be metabolized, and with a metabolizable nutrient that binds no receptor, while measuring both chemotactic drift and growth. A receptor-based reader climbs the first gradient at a cost to growth and ignores the second; a metabolism-based reader does the reverse; a mixed architecture shows both, in proportions that can be measured (J. Adler 1969; Egbert et al. 2010).

28.8 Notes and References

Ashby introduced essential variables and treated regulation as the protection of their bounds (Ashby 1956). The philosophical roots of the chapter lie in Jonas (1966) and in the autopoietic theory of H. R. Maturana and F. J. Varela (1980). Weber and F. J. Varela (2002) and Di Paolo (2005) are the key papers on sense-making and adaptivity; Barandiaran et al. (2009) give the most careful definition of minimal agency; Thompson (2007) is the book-length statement, including the strong continuity thesis. The organizational account of function is due to Mossio et al. (2009) and is developed with the theory of autonomy in Moreno and Mossio (2015). Lyon (2006) argues for a "biogenic" approach that starts the study of cognition from the simplest living systems; this chapter follows that strategy without endorsing its conclusions about mind. For von Uexküll's Umwelt, see Uexküll (2010).

On chemotaxis, J. Adler (1966) and J. Adler (1969) established the phenomenon and its receptor basis, Berg and D. A. Brown (1972) the run-and-tumble mechanism, and Macnab and Koshland (1972) temporal sensing. Robust perfect adaptation was predicted by Barkai and Leibler (1997), confirmed by Alon et al. (1999), and identified as integral feedback by Yi et al. (2000). Egbert et al. (2010) develop the alternative of metabolism-based chemotaxis and its connection to enactive theory.

28.9 Exercises

Exercise 28.1. Solve (28.1) for a ramp $c(t) = c_0 + kt$ with $m(0) = c_0$. Show that $u(t) = k\tau(1 - \mathrm{e}^{-t/\tau})$, so that after a transient the cell reads $\tau\, \mathrm{d}c/\mathrm{d}t$. What does the cell's reading ignore, and why is that useful in environments where the absolute concentration varies over several orders of magnitude?

Exercise 28.2. Consider a one-dimensional run-and-tumble walker with speed v_0 whose mean run duration is τ_+ when moving up a gradient and τ_- when moving down, with each tumble choosing the new direction with probability $1/2$. Show that the mean drift velocity is $v_{\mathrm{d}} = v_0(\tau_+ - \tau_-)/(\tau_+ + \tau_-)$. Using (28.2) with $u \approx \pm k\tau$ for slow ramps, express v_{d} in terms of g, k, τ, and v_0 for small $gk\tau$.

Exercise 28.3. A system lives in an environment that is "good" or "bad" with equal probability. In a good environment staying yields $v = 10$ and leaving yields $v = 6$; in a bad one staying yields $v = 2$ and leaving yields $v = 6$ (arbitrary units). A sensor reports the true state with probability q. Compute the salience σ of the sensor reading (Definition 28.5) as a function of q. At what q does the reading stop mattering?

Exercise 28.4. Describe a thermostat and a bacterium in the terms of Definitions 28.1 to 28.3. Identify precisely which clause of Definition 28.2 the thermostat fails. Could a machine be built that satisfies all of them? Describe one.

Exercise 28.5. Design an experiment in the style of J. Adler (1969) to decide whether a newly isolated bacterium's chemotaxis tracks molecular identity, nutritional value, or both. Specify the chemicals, the mutants, and the measurements, and state which pattern of results would support each answer.

Exercise 28.6. The jamming avoidance response protects electrolocation (Kawasaki 1996). Define a value function for a weakly electric fish in terms of electrolocation performance, and write down the salience of the sign of Δf under that value function. Which measurements would be needed to estimate it in the laboratory?

Exercise 28.7. State the strong continuity thesis in two versions, one functional and one phenomenal. For each version, describe an observation that would count against it. Is there an observation that would count against the phenomenal version without also counting against the functional one?

29

Naturalizing Aboutness

Neurons in the midbrain of *Eigenmannia* fire one way when a neighbor's electric organ discharge is above the fish's own frequency and another way when it is below (Heiligenberg and G. Rose 1985; G. Rose and Heiligenberg 1986b). It is natural to say that their firing is *about* the neighbor. It is equally natural to say that it is about the pattern of amplitude and phase modulation on the fish's skin, or about the spikes arriving from the hindbrain, since the firing covaries with each of these at least as well as with the neighbor. Which is it, and what would make one answer right?

The thesis of this chapter is that the naturalistic theories of content each isolate a real ingredient of aboutness: correlation, counterfactual dependence, function, and structural correspondence. Together they show that aboutness can be grounded in physical relations exploited by a consumer inside the system, without an interpreter. None of them removes the indeterminacy the fish exposes: which link in the causal chain, how broad a class of causes, under which description. And none of them, alone or together, says anything about experience.

29.1 Intentionality

Brentano proposed that what distinguishes mental phenomena is their direction upon an object, their "intentional inexistence": a belief is about something, a desire is for something, and the something need not exist (Brentano 1995). Three features of intentionality make it hard to explain. A state can be *about* a distant object that has no current causal influence on it. A state can be *wrong*, representing things as being a way they are not. And a state represents its object under an aspect: the same neighbor can be represented as "higher in frequency than me" or as "the cause of this modulation."

The project of *naturalizing* intentionality is to state, in terms that do not presuppose content, sufficient conditions for a physical state to have a given content. The minimal formal object is a content assignment.

Definition 29.1 (Content assignment)**.** Let z be a variable internal to a system, with values $z_1, z_2, \ldots$. A *content assignment* maps each value z_k to a condition C_k on the world. The state $z = z_k$ is *correct* (veridical) when C_k obtains and *incorrect* otherwise. A theory of content is a rule that selects, from the physical facts about the system and its history, one content assignment or a restricted class of them.

Every theory below is such a rule. They differ in which physical facts they take to fix the assignment.

The fish gives a concrete test case, and Figure 29.1 lays out the candidates. The sign-selective population's state z has two values, which we can call $z_\uparrow$ and $z_\downarrow$. Candidate contents include: the neighbor's frequency is higher (or lower) than mine; the amplitude and differential phase at my skin rotate one way (or the other) in their joint plane; the T-unit and P-unit afferents fire in one temporal relation (or the other); a conspecific that will jam my electrolocation is present. Each is a condition on the world, each covaries with z, and they come apart in the laboratory.

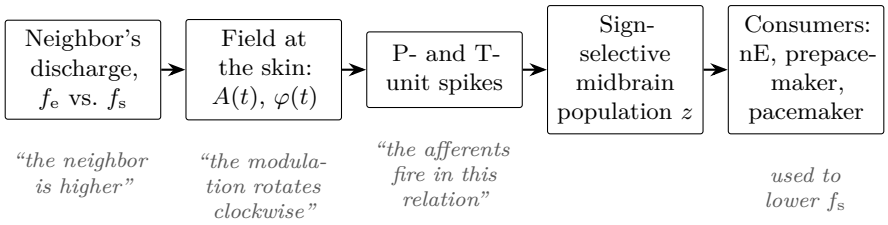


Figure 29.1: The causal chain behind the fish's sign reading, with a candidate content for the reading at each link. Physiology is fixed by Chapter 21; which candidate is the content is fixed by a theory of content.

29.2 Causal Theories and the Disjunction Problem

The simplest proposal is that a state is about whatever causes it. The fish's $z_\downarrow$ is caused by neighbors with higher frequencies, so it means "a neighbor with a higher frequency."

The proposal fails for a reason that recurs throughout the chapter. Tokens of $z_\downarrow$ are also caused by stimuli that are not neighbors at all. Carlson and Kawasaki showed that non-jamming combinations of amplitude and phase modulation, including local modulations, transients, and moving objects, evoke frequency shifts whose tuning, sensitivity, and dynamics are continuous with the jamming avoidance response (Carlson and Kawasaki 2007). If content is whatever causes the state, then the content of $z_\downarrow$ is the disjunction "a higher neighbor *or* one of those other stimuli," and every token is correct. A theory on which a state cannot be wrong is a theory of causation, and misrepresentation disappears from it. Fodor named this the *disjunction problem* (Fodor 1987).

Fodor's own solution appeals to a counterfactual asymmetry (Fodor 1990). The non-neighbor causes produce $z_{\downarrow}$ only because neighbors do; had higher neighbors not caused $z_{\downarrow}$, the other stimuli would not either; the converse dependence does not hold. Content is fixed by the cause on which the others asymmetrically depend. The proposal gets the intuitive verdict in the fish, since the phantom responses are plausibly parasitic on the mechanism that evolved for jamming. Its difficulty is that the counterfactuals it needs are about nearby possible worlds, and it is unclear what physical fact about a circuit settles them.

29.3 Information

Dretske grounded content in information. A signal r carries the information that s is F when the conditional probability that s is F, given r and what the receiver already knows, is one (F. I. Dretske 1981). The theory makes aboutness objective and gives it a quantitative backbone: Shannon's measures say how much a state reduces uncertainty about a variable (Shannon 1948; Cover and Thomas 2006).

It has two well-known limits. The first is error. If carrying information requires a conditional probability of one, a state that is sometimes caused by something else carries no information about its usual cause, and a state that carries information cannot misrepresent. Dretske's later theory adds a history: an indicator acquires a content when learning recruits it to control behavior because of what it indicates, and it keeps that content when it later fires in error (F. Dretske 1988). The addition is a function, and it moves the account toward the teleological theories of the next section.

The second limit is less discussed and matters more for this book. Information does not choose a link in the causal chain. It systematically favors the nearest one.

Proposition 29.2 (Information favors the proximal)**.** *Let $W \to X \to Z$ be a Markov chain: a distal variable W influences an internal state Z only through a proximal variable X. Then $I(Z;W) \le I(Z;X)$, with equality if and only if Z and X are conditionally independent given W.*

Proof. This is the data-processing inequality (Cover and Thomas 2006). By the chain rule, $I(Z;X,W) = I(Z;X) + I(Z;W \mid X) = I(Z;W) + I(Z;X \mid W)$. The Markov property gives $I(Z;W \mid X) = 0$, so $I(Z;W) = I(Z;X) - I(Z;X \mid W) \le I(Z;X)$. □

For the fish, W is the neighbor's frequency relative to the fish's own, X is the modulation pattern at the skin, and Z is the midbrain population's state. By Proposition 29.2, Z carries at least as much information about the skin pattern as about the neighbor, and strictly more whenever anything other than

a neighbor can produce the pattern, which is exactly what phantom jamming shows happens. An information-maximizing rule for content would make the reading about the skin, and by the same argument about the afferent spikes, and then about the reader's own input synapses. Whatever makes the reading about the neighbor, if anything does, is not information alone.

29.4 Teleosemantics

Teleological theories locate content in function. In Millikan's version, a representation sits between a producer and a consumer, and its content is the condition that must obtain for the consumer to perform its proper functions in the historically normal way (Millikan 1984; Millikan 1989). Content is fixed by what the consumer needs, and misrepresentation is a failure of the Normal conditions under which the consumer's functions were selected.

Applied to the fish, the consumer is the chain of diencephalic and medullary nuclei that turn the sign reading into a frequency shift (C. H. Keller et al. 1990; Metzner 1993b). Its function is to move f_{s} away from the neighbor's frequency, which protects the fish's electrolocation from jamming (Kawasaki 1996). That function is served, in the normal case, only when the reading corresponds to the actual sign of Δf. Consumer-based teleosemantics therefore assigns the distal content: $z_{\downarrow}$ means that a neighbor's frequency is higher than the fish's own. The phantom responses come out as errors, since the consumer lowers f_{s} when there is no neighbor to avoid.

Neander's informational teleosemantics assigns content differently (Neander 2017). On her view, a sensory representation has the content C when the sensory system has the function of producing it in response to C, and the relevant C is the property the system was selected for responding to, at the level at which that response function is specified. Her extended treatment of prey recognition in toads, a case that descends from the frog retinal neurons that Lettvin et al. (1959) found responsive to small moving objects, argues that the content of the toad's prey-detecting representation is a visible property of the stimulus, roughly that of a small moving thing of a certain shape, and not "food." Applied to the fish on our reading of her view, the content of $z_{\downarrow}$ sits lower in the chain than Millikan's: something like a joint modulation of amplitude and differential phase in a given sense of rotation, which a neighbor normally produces. On that assignment, many phantom responses are correct readings of the skin that happen to have no neighbor behind them.

The two versions disagree about the same circuit with the same physiology. They disagree because they take different physical facts (the consumer's needs, or the producer's selected response) as content-fixing. Nothing measured in the fish's midbrain adjudicates between them.

29.5 Structural Representation

A different family of theories locates content in a correspondence of structure. A map represents a city because relations among marks on the map mirror relations among places in the city, and a user can reason about the city by manipulating the map (Swoyer 1991). Cummins developed the idea into a general theory of representation as structural similarity (Cummins 1996), and Ramsey argued that structural representations are the clearest case of states that earn the name "representation" in cognitive explanation (Ramsey 2007).

Definition 29.3 (Structural representation)**.** Let Z be a set of internal states with a relation ρ_{int} and W a set of world states with a relation ρ_{ext}. A map $\phi : Z \to W$ is a *structural correspondence* if $\rho_{\text{int}}(z, z') \Rightarrow \rho_{\text{ext}}(\phi(z), \phi(z'))$ for the pairs that matter. The internal states *structurally represent* W when a consumer inside the system exploits ρ_{int} in a way that is successful because the correspondence holds.

The definition is the one place in the theory of content where relations do the semantic work directly: a structural representation represents relations by relations. The hippocampal cognitive map is the standard biological example. Place cells fire at particular locations, and the pattern of which cells are active together or in sequence mirrors which places are near each other or on a path (O'Keefe and Nadel 1978). Downstream processes that exploit co-activity exploit spatial adjacency. Chapter 31 returns to the map, because it can be run when the animal is not at the places it represents.

Example 29.4 (The fish's order relation). The jamming avoidance response exploits a structural correspondence of the simplest kind. The world relation is an order: $f_{\text{e}} > f_{\text{s}}$ or $f_{\text{e}} < f_{\text{s}}$. The internal relation is the difference between the activity of two opposed pathways, the "increase" and "decrease" subdivisions of the nucleus electrosensorius that drive frequency rises and falls (C. H. Keller et al. 1990; Metzner 1993b). The consumer, the push-pull motor control of the pacemaker, exploits the internal relation, and it succeeds because the internal order corresponds to the external one. On a structural theory the content of the reading is the order relation between the two frequencies, and it is relational twice over: a relation between the fish's frequency and its neighbor's, represented by a relation between two pathways.

Structural theories inherit a version of the indeterminacy problem. Structure is cheap, so any set of states corresponds to many targets under some mapping, a point that parallels the triviality arguments against computational implementation in Chapter 17. The constraint that the consumer must *exploit* the correspondence does the work, as it does for implementation.

29.6 Predictive Representation

A fourth approach treats representations as states of a generative model that predicts the system's input. On this view a state is about the hidden cause that the model posits to explain its sensory signals, and its correctness is measured by how well the model's predictions fit (Clark 2013; Hohwy 2013). The approach has an attractive consequence for the distal-proximal problem. A generative model that predicts skin modulations by positing a neighbor, and that predicts how the modulations will change when the fish changes its own frequency, has a variable whose role in the model is to stand for the neighbor. Whether the fish has such a model is an empirical question; the jamming avoidance response can be explained without one (Chapter 21). Generative models are the subject of Chapter 33.

29.7 Three Indeterminacies

The fish makes the classic problems concrete. The *distal-proximal* problem asks which link in the causal chain the state is about. The *disjunction* problem asks how broad the class of causes is that the state represents. The *description* problem asks under which aspect it represents them: the frog's detector responds to flies, to food, and to small dark moving things, which are coextensive in the frog's normal environment and come apart in the laboratory (Lettvin et al. 1959; Neander 2017). Table 29.1 summarizes how each theory handles them.

No row solves all three. The consumer-based theory comes closest, at the price of making content depend on selection history, a fact about the past that no measurement of the present circuit reveals.

29.8 Representation Without a Homunculus

A standing objection to representational explanation is that representations need interpreters, and an interpreter inside the head would need its own representations, and so on. Every theory above answers the objection the same way: the consumer is a mechanism. The fish's nucleus electrosensorius does not understand anything; it responds to the difference between two input streams, and the fish's frequency moves accordingly. The content of the reading is fixed by facts about that mechanism, its history, and its environment, and not by an understanding anywhere.

Ramsey pressed a sharper question: when does calling a state a representation do explanatory work, and when is it a flattering name for a causal mediator (Ramsey 2007)? Shea's answer is that content earns its place when it explains how a system succeeds at the tasks its functions define, by pointing to the exploitable relations (correlations and structural correspondences) that

Theory	Distal or proximal?	Disjunction	Description
Crude causal	Every link is a cause; no choice	Fails: every cause included	No resource
Asymmetric dependence	Whichever cause the others depend on	Solved in principle	Partly, via the counterfactuals
Informational	Favors the proximal (Proposition 29.2)	Fails without a history	No resource
Consumer teleosemantics	Distal: what the consumer needs	Solved: other causes are errors	Fixed by the consumer's function
Informational teleosemantics	Lower: what the producer was selected to respond to	Solved, with a narrower content	Fixed by the selected response
Structural	Whatever the exploited correspondence maps to	Partly	Partly
Predictive	The hidden cause posited by the model	Solved if the model posits one cause	Fixed by the model's variables

Table 29.1: How the main naturalistic theories of content handle the three indeterminacies, applied to the fish's sign reading. The entries summarize each theory's characteristic answer; each has refined versions.

the system's internal processing relies on (Shea 2018). The fish passes that test. Its success at avoiding jamming is explained by citing the correlation between the internal pathway difference and the external frequency order, and the explanation fails where the correlation fails, in phantom jamming.

In this book's vocabulary, the consumer is the reader of Chapter 15 in its third sense, an endogenous controller. Aboutness, on every theory that survives the disjunction problem, requires one.

29.9 What Naturalized Content Does Not Deliver

Brentano proposed intentionality as the mark of the mental. The naturalization project has inverted his proposal. Every theory in this chapter assigns content to states of systems that no one regards as minded: bacterial receptors (Chapter 28), toad prey detectors, fish midbrains, and, on derived rather than original functions, adaptive radios. If these theories are right, aboutness is far more widespread than mind, and possessing it settles nothing about experience. A state can be about the neighbor, correct or incorrect, exploited

by a consumer, and embedded in a self-maintaining organism, and the question of whether there is anything it is like to be in that state remains exactly where it was. Part X takes it up.

29.10 Audit

Canonical phenomenon

Sign-selective neurons in the midbrain of *Eigenmannia* whose activity distinguishes neighbors above the fish's frequency from neighbors below it (Heiligenberg and G. Rose 1985; G. Rose and Heiligenberg 1986b).

Minimal model

Content assignments (Definition 29.1); the data-processing inequality applied to a causal chain (Proposition 29.2); structural correspondence exploited by a consumer (Definition 29.3).

What is relational here?

The content candidate "$f_e > f_s$" is itself a relation, represented by a relation between two opposed pathways (Example 29.4).

What physically reads it?

The consumers: the nucleus electrosensorius and the prepacemaker nuclei that drive the pacemaker in push-pull (C. H. Keller et al. 1990; Metzner 1993b).

What existing field explains it?

The philosophy of intentionality and its naturalization: causal and asymmetric-dependence theories (Fodor 1987; Fodor 1990), informational semantics (F. I. Dretske 1981; F. Dretske 1988), teleosemantics (Millikan 1984; Millikan 1989; Neander 2017), structural representation (Swoyer 1991; Cummins 1996; Ramsey 2007), and varitel semantics (Shea 2018); information theory for the formal point (Cover and Thomas 2006).

What remains unexplained?

Content indeterminacy is not removed by any single theory; the choice among them is partly a choice of explanatory purpose. Nothing in the chapter bears on experience.

A discriminating experiment

Present stimuli that separate the candidate contents: neighbor-mimic stimuli that produce the normal skin pattern, phantom stimuli that produce the skin pattern without a neighbor (Carlson and Kawasaki 2007), and stimuli that present a neighbor's frequency relation through an unusual skin pattern. Measure the exploitable correlations between the nucleus electrosensorius output and each candidate. The theories predict the same physiology and

different correctness conditions; the experiment shows which correlations the consumer's success depends on, which is the input a varitel theory needs.

29.11 Notes and References

Brentano (1995) is the source of the modern concept of intentionality. The naturalization literature begins with F. I. Dretske (1981); Dretske's functional turn is F. Dretske (1988). Fodor's disjunction problem and asymmetric dependence theory are in Fodor (1987) and Fodor (1990). Millikan (1984) and Millikan (1989) are the foundational statements of consumer-based teleosemantics; Neander (2017) is the most developed informational teleosemantics, including the toad case; Lettvin et al. (1959) is the physiological source of that case. On structural representation, see Swoyer (1991), Cummins (1996), and Ramsey (2007); Shea (2018) integrates correlational and structural content into a single framework organized around task functions. Predictive approaches to content are developed in Clark (2013) and Hohwy (2013). The fish physiology used throughout follows Heiligenberg (1991a).

29.12 Exercises

Exercise 29.1. Prove Proposition 29.2 from the chain rule for mutual information, and give an example in which the inequality is strict because a second cause can produce X without W.

Exercise 29.2. A sign-selective neuron reports the sign of Δf correctly with probability 0.9, and the two signs are equally likely. Compute the mutual information between the neuron's report and the sign, in bits. Now suppose phantom stimuli, which occur on 10% of occasions and have no sign, evoke the report $z_{\downarrow}$. Recompute the information the report carries about the sign and about the skin pattern.

Exercise 29.3. For each candidate content in Figure 29.1, describe a laboratory stimulus under which the fish's reading is correct on that content and incorrect on at least one other. Which candidate makes the fewest readings incorrect, and why is that not an argument for it?

Exercise 29.4. Adler found that some non-metabolizable chemicals attract *E. coli* (J. Adler 1969). State the disjunction problem for the bacterial receptor's reading, and give the verdict of consumer teleosemantics, informational teleosemantics, and a structural theory.

Exercise 29.5. Place cells $1, \dots, n$ have place fields centered at $\boldsymbol{p}_1, \dots, \boldsymbol{p}_n$. Define an internal relation on pairs of cells from their co-activity, and show that it corresponds to spatial proximity under stated assumptions about field size. What must a downstream consumer do for the correspondence to count as a structural representation under Definition 29.3?

Exercise 29.6. An adaptive radio switches channels when it detects interference. Apply consumer teleosemantics to the radio, using the designer's intentions as the source of function. Which features of the resulting content are derived, and what would it take for the radio's content to be original?

Exercise 29.7. Apply Ramsey's job-description challenge to the methylation state of the bacterial chemotaxis network in Chapter 28. Is it a representation of the recent past concentration, or a causal mediator that the word "representation" merely flatters? Argue both sides, and say what observation would decide.

30

Misrepresentation

Every physical measurement deviates from what it measures. A voltmeter reads a little high, a thermometer lags a changing temperature, a receptor fires on a thermal fluctuation. Most of these deviations are not errors in any interesting sense: they are what happens when one physical system is coupled to another. Some deviations are errors. The weakly electric fish that lowers its frequency to escape a neighbor who is not there has made a mistake, and the mistake is the fish's.

The thesis of this chapter is that misrepresentation is a deviation between an internal estimate $\hat{R}$ and the world variable R that it stands for, evaluated relative to a consumer whose success depends on their correspondence. Four conditions separate it from mere deviation: a shared mechanism, a consumer that acts on the reading, a cost to acting on it, and a content assignment on which it is incorrect. Signal detection theory shows that a well-designed reader should make some errors, so that misrepresentation is not the same as malfunction. Phantom jamming in *Eigenmannia* is a natural case in which the first two conditions have been established, the third is open, and the fourth depends on the theory of content chosen in Chapter 29. None of the four mentions experience.

30.1 Deviation and Error

We start from the weakest notion.

Definition 30.1 (Physical deviation)**.** Let R be a variable of the world and $\hat{R}$ an internal variable of a system whose value covaries with R. The *deviation* is $\epsilon = \hat{R} - R$, or, for categorical variables, the event $\hat{R} \neq R$.

Deviation is everywhere, and on its own it is not error. The temperature of a stone lying in the sun covaries with the air temperature and deviates from it; nothing has gone wrong. Three further ingredients are needed, and the previous two chapters supplied them. A content assignment (Chapter 29) says which world condition a state stands for, so that there is a fact about when it

is incorrect. A consumer (Chapter 15) acts on the state. And a value function (Chapter 28) ranks the consequences of the consumer's action for the system that performs it.

Definition 30.2 (Misrepresentation)**.** Let a system have an internal variable z with a content assignment $z_k \mapsto C_k$, a consumer that maps z_k to action a_k, and a value function v. A tokening of z_k is a *misrepresentation* when (i) C_k does not obtain, (ii) the consumer produces a_k, the action appropriate if C_k obtained, and (iii) the consequence of a_k, evaluated by v under the actual condition, is worse than the consequence of the action appropriate to the actual condition. When (i) and (ii) hold and (iii) fails, the tokening is *semantically* incorrect and *functionally* harmless.

The definition makes misrepresentation a three-place relation among a reading, the world, and a consumer. Remove the consumer and only deviation remains. Remove the world condition and there is nothing for the reading to be wrong about. Remove the value function and the error has no consequences that belong to the system.

Misrepresentation must also be kept apart from malfunction. A receptor with a mutated binding site malfunctions; it fails to do what receptors of its type were selected to do. A healthy receptor that binds a molecule resembling its normal ligand works exactly as it was selected to work, and it misrepresents because the conditions are abnormal. Millikan's analysis of error rests on this distinction: a representation is false when the conditions that normally explain its success fail to hold, whether or not any mechanism is broken (Millikan 1989). Dretske made the same point with learned indicators that keep their recruited function when they later fire in error (F. Dretske 1988). A system can misrepresent with every part intact, and it can malfunction without misrepresenting, as when a broken consumer ignores a correct reading.

30.2 Kinds of Error

The reading loop fails in several distinct ways, and the distinctions matter because each implicates a different part of the loop (Table 30.1).

The first two are errors of detection and are the subject of the next section. Hallucination and mistaken prediction require internal states that are not driven by the current input, the detached representations of Chapter 31, and they become possible only in systems that model their world. Prediction errors in reafference are the familiar case: a self-produced touch is perceived as less intense and less ticklish than the same touch produced by someone else, and the attenuation is attributed to a prediction of the touch from the motor command (Blakemore et al. 1998). When the prediction is wrong, the attenuation is wrong with it.

Kind	Example	Part of the loop
False detection	A fish shifts its frequency in response to modulations that no neighbor produced (Carlson and Kawasaki 2007)	Reader D
Missed detection	A jamming neighbor too weak or too distant to register	Mixture M and reader D
Hallucination	A state normally caused by a stimulus is generated internally with no stimulus present	Internal model
Mistaken prediction	A forward model predicts the wrong consequence of an action, $\hat{R}(t+\Delta t) \neq R(t+\Delta t)$	Internal model
Inappropriate action	A correct reading drives the wrong response	Policy π

Table 30.1: Kinds of error, classified by the part of the reading loop that fails.

Inappropriate action differs from the rest. The reading is correct and the consumer does the wrong thing with it, so the error is functional without being semantic. The converse also occurs: a false reading that drives a harmless action is semantically wrong and functionally costless. Definition 30.2 requires both kinds of wrongness at once.

30.3 Signal Detection and Optimal Error

A reader that must decide whether a condition holds, from evidence corrupted by noise, cannot avoid errors. It can only choose which errors to make. The standard formal treatment is Bayesian decision theory, which has become a general framework for modeling how neural systems handle uncertainty (Knill and Pouget 2004).

Model 30.3 (A detector with two hypotheses)**.** A reader receives evidence x generated under one of two conditions: H_1 (a jamming neighbor is present) or H_0 (it is not), with prior probabilities p_1 and $p_0 = 1 - p_1$ and likelihoods $p(x \mid H_1)$, $p(x \mid H_0)$. It responds (shifts its frequency) or does not. A false alarm (responding under H_0) costs c_{FA}; a miss (not responding under H_1) costs c_{M}; correct decisions cost nothing.

Proposition 30.4 (The optimal criterion)**.** *In Model 30.3, the rule that minimizes expected cost is to respond exactly when the likelihood ratio exceeds a criterion,*

$$\Lambda(x) = \frac{p(x \mid H_1)}{p(x \mid H_0)} > \beta, \qquad \beta = \frac{p_0\, c_{\mathrm{FA}}}{p_1\, c_{\mathrm{M}}}. \tag{30.1}$$

Proof. For a given x, responding has expected cost $c_{\mathrm{FA}}\, \mathbb{P}(H_0 \mid x)$ and not responding has expected cost $c_{\mathrm{M}}\, \mathbb{P}(H_1 \mid x)$. Respond when the first is smaller:

$c_{\mathrm{FA}}\, p_0\, p(x \mid H_0) < c_{\mathrm{M}}\, p_1\, p(x \mid H_1)$, which rearranges to (30.1). Since the rule minimizes the cost at every x, it minimizes the expected cost. □

The consequence is the reason for this section. Whenever the two likelihoods overlap, every criterion produces both false alarms and misses, and the optimal criterion trades them against each other by their costs and base rates (Figure 30.1). If misses are expensive and false alarms cheap, an optimal reader should accept many false alarms. A detector that never errs in one direction is either facing unrealistically clean evidence or paying too much in the other.

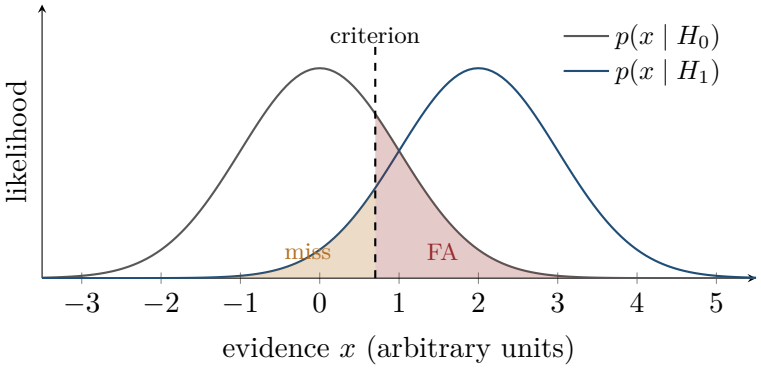


Figure 30.1: Signal detection with overlapping evidence (illustrative: unit-variance Gaussians two standard deviations apart). Any criterion yields false alarms (FA) and misses; Proposition 30.4 places it where their expected costs balance. Some misrepresentation is designed in.

For the fish, the relevant costs are not known. A plausible structure is easy to write down: a missed jamming neighbor degrades electrolocation, which the jamming avoidance response exists to protect (Kawasaki 1996), while a false alarm costs a small, reversible frequency shift. If that structure holds, selection would favor a criterion that tolerates phantom responses. We state this as a model to be tested and not as a finding; nobody has measured the costs.

30.4 Phantom Jamming

Carlson and Kawasaki asked whether stimuli that do not jam *Eigenmannia* can nevertheless evoke frequency shifts by being misperceived as jamming signals (Carlson and Kawasaki 2007). They presented combinations of amplitude and phase modulation that no neighbor of a nearby frequency would produce. The fish shifted their frequencies. The responses to jamming and to non-jamming stimuli had similar tuning, sensitivity, and dynamics, and they formed a smooth continuum rather than two categories. The authors also identified natural situations that can evoke such shifts: localized amplitude modulations, transient changes, and moving objects.

The first result establishes the condition that makes the question of error well posed. One mechanism handles both kinds of stimulus. If the phantom responses came from a separate system, they would be a different behavior with a different content, and calling them errors of the jamming avoidance system would be a category mistake. Because the tuning and dynamics are continuous, the phantom responses are tokenings of the same reader, consumed by the same motor pathways, and the question "are they misrepresentations?" has a definite form.

Its answer depends on the theory of content, as Chapter 29 predicted. On consumer-based teleosemantics, the reading means that a neighbor with a higher (or lower) frequency is present, and phantom responses are false alarms. On informational teleosemantics, the reading may mean that a certain joint modulation of amplitude and differential phase is present at the skin; phantom stimuli produce that modulation, and the readings are correct, while the consumer acts under abnormal conditions. On a crude causal theory, phantom stimuli are among the causes of the reading, and there is no error at all. The physiology is the same on every account.

Condition (iii) of Definition 30.2, the cost, is an empirical question that the existing work leaves open, and it can be stated as a hypothesis.

Hypothesis 30.5 (Phantom shifts are costly)**.** Frequency shifts evoked by phantom-jamming stimuli reduce the fish's electrolocation performance, measured by object-discrimination accuracy or by the precision of active-sensing movements, relative to matched trials in which the same stimuli are presented to a fish whose frequency is held fixed.

If the hypothesis holds, phantom responses are functional errors, and on consumer semantics they are misrepresentations in the full sense of Definition 30.2. If it fails, they are semantic errors without functional cost, of the kind an optimal detector tolerates.

30.5 What Separates Misrepresentation from Deviation

The analysis yields a test that can be applied to any system.

Principle 30.6 (Four conditions for misrepresentation)**.** A deviation $\hat{R} \neq R$ in a physical system is a misrepresentation only if

1. *Shared mechanism*: the deviant response is produced by the same reader that produces correct responses, with continuous tuning and dynamics;
2. *Consumer*: a downstream process acts on the response as it would under the condition the response normally indicates;
3. *Cost*: the resulting action is worse, by the system's own value function, than the action appropriate to the actual condition;

4. *Content*: under a stated theory of content, the response's content is false.

The first three conditions are empirical and can be measured. The fourth is theoretical, and it is where the indeterminacy of Chapter 29 enters. A stone's temperature meets none of the conditions. A healthy thermometer read by nobody meets the first alone. An adaptive radio that mistakes a burst of noise for interference and hops away from a clean channel meets the first three, and it meets the fourth only relative to a content assigned by its designer: its errors are misrepresentations in a derived sense. The fish's phantom responses meet the first two by the evidence of Carlson and Kawasaki (2007), await a test of the third, and meet the fourth on some theories and not others.

Example 30.7 (The bacterial receptor). Adler found that some chemicals that *E. coli* cannot metabolize nevertheless attract it (J. Adler 1969). A cell climbing a gradient of such a chemical meets the first two conditions: the same receptor and motor pathway operate as for a nutrient. It plausibly meets the third, since the climb spends time that could be spent climbing a nutrient gradient. Whether it meets the fourth depends, again, on whether the receptor's reading means "nutrient" (consumer semantics) or "this molecular shape" (a lower, producer-based content).

The pattern in both examples is the same. Physics and physiology settle the first three conditions. The fourth is settled by a choice of theory, and the choice is not arbitrary, but it is not settled by any measurement of the present mechanism either.

30.6 Error Without Experience

Misrepresentation is sometimes treated as a mark of the mental, on the grounds that only minds can be wrong. The four conditions show why that is too quick. A fish's midbrain, a bacterium's receptor array, and a radio's interference detector can each meet some or all of them, and nothing in any condition refers to how anything seems. Error is where representation and semantics become legitimate questions about a physical system, because only a system that can be wrong has states with correctness conditions. It is not where experience begins, and nothing in this chapter locates that point.

The next chapter follows readings that are cut loose from the present stimulus. Detached readings bring new kinds of error, including hallucination and mistaken prediction, and they bring new capacities that no cued reader has.

30.7 Audit

Canonical phenomenon

Phantom jamming: *Eigenmannia* shifts its discharge frequency in response

to combinations of amplitude and phase modulation that no jamming neighbor produced, with tuning, sensitivity, and dynamics continuous with the jamming avoidance response (Carlson and Kawasaki 2007).

Minimal model

Deviation and misrepresentation (Definitions 30.1 and 30.2); Bayes-optimal detection with costs and base rates (Proposition 30.4); four conditions (Principle 30.6).

What is relational here?

Misrepresentation is a three-place relation among a reading, a world condition, and a consumer with a value function. The phantom stimulus reproduces a relation (a joint amplitude-phase modulation) without the relation behind it (a neighbor at a nearby frequency).

What physically reads it?

The fish's sign-reading circuit in the torus semicircularis and nucleus electrosensorius, consumed by the prepacemaker and pacemaker nuclei (C. H. Keller et al. 1990; Metzner 1993b).

What existing field explains it?

Teleosemantic and informational accounts of error (Millikan 1989; F. Dretske 1988; Neander 2017); the disjunction problem (Fodor 1987); Bayesian decision theory as applied to neural systems (Knill and Pouget 2004); the neuroethology of the jamming avoidance response (Heiligenberg 1991a; Carlson and Kawasaki 2007).

What remains unexplained?

Two items. The cost of phantom shifts is unmeasured (Hypothesis 30.5). The content condition depends on an unresolved choice among theories of content. No experiential residue arises.

A discriminating experiment

Test Hypothesis 30.5: present phantom stimuli during an electrolocation task in two conditions, with the fish free to shift its frequency and with the shift blocked pharmacologically at the prepacemaker nuclei, whose transmitters are known (Metzner 1993b). A performance deficit that appears only when the fish shifts identifies the shift as a costly error.

30.8 Notes and References

The problem of misrepresentation is the central test of every naturalistic theory of content. Fodor (1987) and Fodor (1990) state the disjunction problem; F. Dretske (1988) and Millikan (1989) give the historical and teleological answers; Neander (2017) is the fullest recent treatment and distinguishes misrepresentation from malfunction with care. For Bayesian decision theory

in perception and neural computation, see Knill and Pouget (2004). The phantom-jamming experiments are reported in Carlson and Kawasaki (2007); the physiology of the circuit is summarized in Heiligenberg (1991a). Blakemore et al. (1998) is the standard reference for predictive attenuation of self-produced touch. On the receptor basis of bacterial chemotaxis and its independence from metabolism, see J. Adler (1969).

30.9 Exercises

Exercise 30.1. In Model 30.3, let $x \sim \mathcal{N}(0,1)$ under H_0 and $x \sim \mathcal{N}(d',1)$ under H_1. Show that the optimal rule is to respond when $x > x^* = d'/2 + (\ln \beta)/d'$. Compute the false-alarm and miss rates at x^* for $d' = 2$, $p_1 = 0.2$, and $c_{\mathrm{M}}/c_{\mathrm{FA}} \in \{1, 10, 100\}$.

Exercise 30.2. Classify each of the following as physical deviation, malfunction, misrepresentation, or inappropriate action, and say which condition of Principle 30.6 decides the case: (a) a thermometer in a sunlit window reads above the room temperature; (b) a fish with a lesioned prepacemaker nucleus fails to lower its frequency when a higher neighbor appears; (c) a fish lowers its frequency in response to a moving object; (d) a correctly detected predator elicits approach instead of escape.

Exercise 30.3. Design the experiment of the audit in detail. Specify the phantom stimuli, the electrolocation task, the clamping method, the sample size needed to detect a given effect, and the result that would refute Hypothesis 30.5.

Exercise 30.4. Apply Principle 30.6 to Example 30.7. Which measurement would establish the cost condition for a bacterium, and why is it easier to make than the corresponding measurement for the fish?

Exercise 30.5. Self-produced touch is attenuated relative to the same touch produced externally (Blakemore et al. 1998). Suppose a delay is inserted between the motor command and the touch. Describe the resulting error in the vocabulary of Table 30.1, and say which part of the reading loop has failed.

Exercise 30.6. Construct a pair of systems, one engineered and one evolved, that meet conditions (1) to (3) of Principle 30.6 identically. What, if anything, distinguishes their errors under condition (4)? Does the answer change if the engineered system was designed by an evolutionary algorithm?

A radio receiver reads, but its reader exists because an engineer built it. A cell builds and maintains its own readers, and its readings feed the processes that keep it, and them, in existence. Pattee's semantic closure, the autopoietic tradition, and teleosemantics each describe that loop in their own vocabulary. With the loop comes a norm: a reading can succeed or fail relative to the organization that uses it, and misrepresentation becomes a physical possibility. Part VII asks what happens when readings detach from the present stimulus.

Part VII

Detachment and World Models

How can a reading concern what is not present?

31

Representation Without the Present Stimulus

The weakly electric fish reads the sign of its neighbor's frequency difference for as long as the beat lasts. When the neighbor swims away, the modulation disappears and so does the reading. A rat running in a wheel between the trials of a memory task does something the fish's jamming avoidance system does not: sequences of hippocampal activity unfold with no change in what the rat sees or feels, and they differ according to which arm of the maze the rat will choose next (Pastalkova et al. 2008). The rat's reading concerns a place it is not in and a choice it has not made.

The thesis of this chapter is that this difference, *detachment*, is a graded and measurable property of a physical reader. A detached state carries information about a world variable beyond what the reader's current input carries. Known mechanisms implement it: integration, forward models, emulators, and internally generated sequences. With detachment, a reader stops merely tracking a relation and starts holding one, which brings new capacities and new ways of being wrong.

31.1 Cued and Detached Representations

Gärdenfors distinguished *cued* representations, which stand for something present in the current external situation, from *detached* ones, which stand for something that is not present and may not exist (Gärdenfors 1996). An animal that tracks a visible prey item uses a cued representation; an animal that keeps searching behind the rock where the prey disappeared uses a detached one. Clark and Toribio made a related distinction for problems. Some tasks can be solved by systems that respond to the current input; others are "representation-hungry" because they require reasoning about the absent, the distant, or the counterfactual, or selective sensitivity to abstract properties that the input does not carry directly (Clark and Toribio 1994).

The distinction can be made quantitative. Let $R(s)$ be a world variable at

time s, $Y(t)$ the reader's input at time t, and $Z(t)$ its internal state.

Definition 31.1 (Detachment)**.** The *detachment* of a reader's state with respect to R at lag $s - t$ is

$$\mathcal{D}(t, s) = I\big(Z(t);\, R(s) \mid Y(t)\big),$$

the information the internal state carries about the world at time s beyond what the current input carries. A reader is *cued* with respect to R when $\mathcal{D}(t, s) \approx 0$ for all $s \neq t$, and *detached* to the extent that $\mathcal{D}(t, s) > 0$ for $s < t$ (memory) or $s > t$ (prediction).

The definition is ours, and it formalizes the distinction rather than discovering it. It has one immediate consequence, which fixes where detachment can come from.

Proposition 31.2 (Memoryless readers are cued)**.** *If the internal state is a function of the current input alone, $Z(t) = f(Y(t))$, then $\mathcal{D}(t, s) = 0$ for every s.*

Proof. Given $Y(t)$, the state $Z(t) = f(Y(t))$ is fixed, so $H(Z(t) \mid Y(t)) = 0$, and conditional mutual information is bounded by conditional entropy: $\mathcal{D}(t, s) \leq H(Z(t) \mid Y(t)) = 0$. □

Detachment therefore requires internal dynamics: a state that depends on past inputs, on the system's own past states, or on its motor commands. It is a property of a reader with a history.

31.2 Memory: The Simplest Detachment

The simplest reader with a history is a leaky integrator. The bacterial chemotaxis network of Chapter 28 contains one: a slow methylation state that tracks the recent concentration, so that the cell can compare the present with the remembered past. That memory is a detached representation with a short horizon. It carries information about the concentration a few moments ago, which the present input no longer carries.

Model 31.3 (Leaky memory)**.** A world variable R, Gaussian with variance σ_R^2, drives the input until time t_0 and then disappears. At t_0 the internal state is $Z(t_0) = R + n_0$, with independent Gaussian noise of variance σ_0^2. Afterwards the state relaxes with time constant τ and is perturbed by white noise of intensity q:

$$\dot{Z} = -Z/\tau + \xi(t), \qquad \mathbb{E}[\xi(t)\xi(t')] = q\,\delta(t - t').$$

Solving the linear equation gives $Z(t_0+\Delta t) = (R+n_0)\,\mathrm{e}^{-\Delta t/\tau} + \eta$, where η is Gaussian with variance $\frac{1}{2}q\tau\,(1-\mathrm{e}^{-2\Delta t/\tau})$. The information the state retains about R is that of a Gaussian channel:

$$I(Z(t_0+\Delta t);\,R) = \frac{1}{2}\log_2\!\left(1+\frac{\sigma_R^2\,\mathrm{e}^{-2\Delta t/\tau}}{\sigma_0^2\,\mathrm{e}^{-2\Delta t/\tau}+\frac{1}{2}q\tau\,(1-\mathrm{e}^{-2\Delta t/\tau})}\right). \tag{31.1}$$

The information decays monotonically to zero (Figure 31.1). The horizon over which the state remains useful grows with τ and shrinks with the noise intensity q. A slow, quiet integrator holds a relation for a long time; a fast or noisy one forgets it.

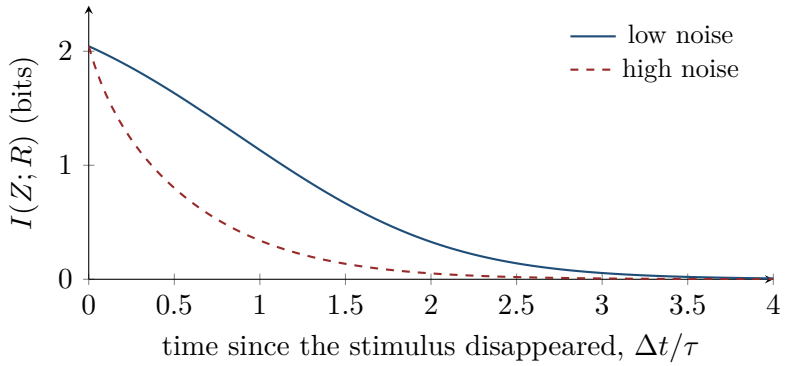


Figure 31.1: Information retained by the leaky memory of Model 31.3 after its stimulus disappears, from (31.1) with $\sigma_R^2/\sigma_0^2 = 16$ and two noise levels ($\frac{1}{2}q\tau/\sigma_0^2 = 0.5$ and 4; illustrative values).

Memory in this sense is cheap and nearly universal among readers with internal dynamics. It is the first step of detachment, and it already changes what a reader can read: a memoryless reader can respond only to relations present in its current input, while a reader with memory can respond to relations between its current input and its past, as the bacterium does.

31.3 Prediction and Forward Models

The second step runs the other way in time. A reader that predicts carries information about $R(t+\Delta t)$ beyond what $Y(t)$ carries, which is possible only if it has a model of how the world, or its own body, will change.

Motor control provides the best-studied case. When a person moves an arm in the dark, errors in the estimated hand position accumulate over the movement in a way that is well described by a model combining a prediction from the outgoing motor command with delayed and noisy sensory feedback, the structure of a Kalman filter (Wolpert et al. 1995). The component that maps the current state and the motor command to a predicted next state is a *forward model*; its partner, which maps a desired next state to the command

that would produce it, is an *inverse model*; both are central to theories of cerebellar motor control (Kawato 1999).

Model 31.4 (Forward model with sensory correction)**.** Let the plant (body and world) evolve as $R_{t+1} = AR_t + Bu_t + w_t$ with motor command u_t, and let the sensory input be $Y_t = CR_t + v_t$. The reader maintains an estimate

$$\hat{R}_{t+1} = A\hat{R}_t + Bu_t + K_{t+1}(Y_{t+1} - C(A\hat{R}_t + Bu_t)), \tag{31.2}$$

where $A\hat{R}_t + Bu_t$ is the forward model's prediction, the bracket is the prediction error, and the gain K weighs prediction against evidence. With feedback removed ($K = 0$), the estimate is propagated by the forward model alone.

The prediction $A\hat{R}_t + Bu_t$ is the book's decisive transition in its simplest form: from a reading of $R(t)$ to a reading of $\hat{R}(t + \Delta t)$, a variable about a state of the world that does not yet exist. Figure 31.2 shows the loop.

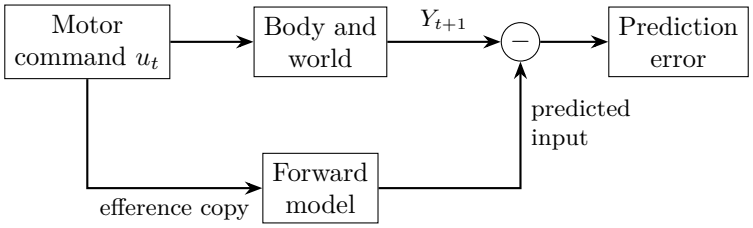


Figure 31.2: A forward model driven by an efference copy of the motor command predicts the next input; the comparison with the actual input yields a prediction error that corrects the estimate (Model 31.4). Run with the lower branch alone, the model produces detached predictions.

Example 31.5 (Two electric fishes, two kinds of reading). Weakly electric fishes span the range from cued to predictive reading. The jamming avoidance response of *Eigenmannia* is cued: it is computed from the current electrosensory mixture without any internal reference to the fish's own motor command (Heiligenberg 1991a). Pulse-type mormyrid fishes, by contrast, receive a corollary discharge of each electric organ command, and cerebellum-like circuits in their electrosensory lobe learn a "negative image" of the sensory consequences of the fish's own discharge and subtract it from the input (C. C. Bell 1981; C. Bell et al. 1997; Requarth and Sawtell 2011). The negative image is a learned prediction of reafference, triggered by the motor command: a detached representation of the fish's own expected input.

Chapter 22 examines the contrast in detail.

31.4 Emulation

A forward model that can be driven by motor commands can also be driven by motor commands that are never executed. Craik proposed in 1943 that an

organism carrying a small-scale model of external reality and of its own possible actions could try out alternatives before committing to one (Craik 1943). Grush turned the proposal into a theory of representation (Grush 2004). An *emulator* is a component that mimics the input–output behavior of the body and environment; run online, it supplies the predictions of Model 31.4; run offline, driven by efference copies with the actual movement inhibited, it produces motor and perceptual imagery. On the emulation theory, a representation is an emulator state that the system uses in place of the state of the target.

Emulation is the step at which detachment becomes deliberate use. A leaky memory carries a past relation because it cannot help it; a forward model predicts because prediction improves online control; an emulator run offline carries relations that the system has chosen to explore and that may never be realized. The capacities are graded, and the mechanisms overlap, since the same forward model serves prediction online and emulation offline.

31.5 Cognitive Maps and Internally Generated Sequences

Tolman argued from maze experiments that rats learn something like a map of their environment, which lets them take novel routes and learn about places without being rewarded there (Tolman 1948). The discovery of place cells in the hippocampus gave the proposal a physiological basis: cells that fire when the animal is at particular places, forming, as a population, a structural representation of space in the sense of Chapter 29 (O'Keefe and Nadel 1978).

A map is a candidate for detachment only if it can be read when the animal is not at the places it represents. The rat in the wheel shows that it can. In the experiment of Pastalkova and colleagues, rats ran in a wheel during the delay period of a memory task, with the environment and the body's movement held nearly constant. Hippocampal neurons fired in sequences that spanned the delay, the sequences were generated internally rather than by changing cues, and they differed on trials that ended in different choices, so that the sequence predicted the arm the rat would enter (Pastalkova et al. 2008). In the terms of Definition 31.1, the hippocampal state carried information about a future world state ($R(s)$, the chosen arm) that the current input did not carry. Chapter 32 follows the same circuits into planning, where the represented states include ones the animal considers and does not choose.

31.6 A Scale of Detachment

The systems of this and earlier chapters can be ordered by how their states relate to their current input (Table 31.1).

The scale is ordinal and deliberately rough. It orders mechanisms by the source of the extra information: past input, learned regularities of the system's

System	What the state carries beyond the current input	Kind
Jamming avoidance, *Eigenmannia*	Nothing beyond the reader's integration time	Cued
Bacterial chemotaxis	The recent concentration	Memory
Mormyrid negative image	The expected sensory consequence of the fish's own discharge	Learned prediction
Motor forward model	The next state of the limb, given the command	Prediction
Motor emulator run offline	Consequences of unexecuted actions	Emulation
Hippocampal delay sequences	The upcoming choice, with constant cues	Internally generated

Table 31.1: A scale of detachment. Each row adds a way in which an internal state can carry information about a world variable that the current input does not.

own actions, models of the body and world, and self-generated dynamics. It does not order them by anything else, and in particular it says nothing about whether any of them is accompanied by experience. Engineered Kalman filters navigate aircraft and position telephones using exactly the structure of Model 31.4.

31.7 What Detachment Adds to the Reader Problem

Three consequences follow for the book's argument.

First, the reader's input now includes the reader. A detached state depends on the system's own past states and motor commands as well as on the mixture $\mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})$, so the reading loop acquires an internal branch. This is the first form of the self-reading of Part IX.

Second, a relation can be held without being present. The fish can use the sign of Δf only while the beat lasts; a system with a detached representation can carry a relation (this place is north of that one, this movement will bring the hand there) into situations where no current input carries it. Relational reading stops being bound to the current configuration.

Third, detachment multiplies the ways of being wrong. A cued reader can misread its input. A detached reader can also mistake its own state for input, which is one way to describe hallucination, and it can predict wrongly, which is the prediction error of Model 31.4 and the mistaken prediction of Chapter 30. The errors that only detached readers make are the price of the capacities that only detached readers have.

31.8 Audit

Canonical phenomenon

Hippocampal cell assemblies in rats running in a wheel generate sequences internally, with constant external cues, that predict the arm the rat will choose after the delay (Pastalkova et al. 2008).

Minimal model

The detachment index $\mathcal{D}(t, s) = I(Z(t); R(s) \mid Y(t))$ (Definition 31.1); memoryless readers are cued (Proposition 31.2); information decay in a leaky memory (31.1); a forward model with sensory correction (31.2).

What is relational here?

A relation across time: between the system's present state and a past or future state of the world, and in prediction between a present estimate and a state that does not yet exist.

What physically reads it?

Integrators (the methylation memory of chemotaxis), forward models and emulators (cerebellar circuits in motor control (Kawato 1999)), learned negative images in mormyrid electrosensory lobes (C. Bell et al. 1997), and hippocampal circuits that generate sequences (Pastalkova et al. 2008).

What existing field explains it?

Estimation and control theory (Kalman filtering and internal models), motor neuroscience (Wolpert et al. 1995; Kawato 1999), cognitive science of representation (Gärdenfors 1996; Clark and Toribio 1994; Grush 2004; Craik 1943), and hippocampal physiology (Tolman 1948; O'Keefe and Nadel 1978; Pastalkova et al. 2008). Together they explain the phenomena of the chapter.

What remains unexplained?

No mechanistic residue. A conceptual question remains: whether detachment is one continuum, as Definition 31.1 treats it, or whether offline generation (emulation, internally generated sequences) is a distinct kind that a single index hides.

A discriminating experiment

Measure $\mathcal{D}(t, s)$ directly after cue removal in systems at different points of Table 31.1. For the fish, remove a jamming neighbor abruptly and decode the prior sign of Δf from midbrain population activity over the following seconds: the account predicts that the information vanishes within the reader's integration time. For the rat, decode the upcoming choice from delay-period activity with cues held constant and compare with a shuffled-choice null model.

31.9 Notes and References

The cued-detached distinction is from Gärdenfors (1996); the notion of representation-hungry problems is from Clark and Toribio (1994). Craik (1943) is the classic source of the internal-model idea in psychology, and Tolman (1948) of the cognitive map; O'Keefe and Nadel (1978) connected the map to the hippocampus. For internal models in motor control, see Wolpert et al. (1995) and Kawato (1999); for the emulation theory of representation, Grush (2004). Internally generated hippocampal sequences are reported by Pastalkova et al. (2008). The mormyrid negative image is described in C. C. Bell (1981), C. Bell et al. (1997), and Requarth and Sawtell (2011). The information-theoretic tools used in Definition 31.1 and (31.1) are standard; see Cover and Thomas (2006).

31.10 Exercises

Exercise 31.1. Derive (31.1) from Model 31.3. For $\sigma_R^2/\sigma_0^2 = 16$ and $\tfrac{1}{2}q\tau/\sigma_0^2 = 0.5$, find the time at which the retained information falls to 0.1 bits. How does that horizon scale with τ when q is fixed?

Exercise 31.2. Prove Proposition 31.2, and give an example of a reader whose state depends on past inputs but for which $\mathcal{D}(t,s)$ is still zero for all $s \neq t$. What property of the world variable makes this possible?

Exercise 31.3. Simulate Model 31.4 in one dimension with $A = 1$, $B = \Delta t$, $C = 1$, a constant velocity command, and sensory feedback removed after half the movement. Plot the estimate error against time with and without feedback, and compare the qualitative pattern with the one reported by Wolpert et al. (1995).

Exercise 31.4. For each row of Table 31.1, state which lags $s-t$ have $\mathcal{D}(t,s) > 0$ and where the extra information comes from. Add two rows from outside biology.

Exercise 31.5. Design the fish experiment of the audit: specify how to remove the neighbor abruptly, which neurons to record, how to decode the prior sign of Δf, and what time course of decoding accuracy the cued account predicts. What result would show that the jamming avoidance system has a memory?

Exercise 31.6. Specify an emulator for a two-joint arm, following Grush (2004). Describe the inputs it takes when run online and offline, and state what kind of error it can make offline that it cannot make online.

32

Counterfactuals

We describe the electric fish counterfactually: had the neighbor's frequency been higher, the fish would have lowered its own. The description is true, and it is what makes the fish's circuit an implemented computation rather than an interpretation imposed from outside (Chapter 17). The fish, however, does not represent the counterfactual. It responds to the frequency difference that is actually present. A rat pausing at a fork in a maze, with hippocampal activity sweeping ahead along paths it has not taken, is doing something different: its brain is evaluating what would happen along routes that may never be run.

The thesis of this chapter is that counterfactual *representation*, the capacity to evaluate the consequences of states and actions that are not actual, must be kept apart from counterfactual *description*, which applies to every implemented computation. The capacity has an exact formal treatment in structural causal models and possible-worlds semantics. Animals implement parts of it in prospective control, deliberation, and look-ahead. Control theory explains why a regulator that performs well must contain a model of what it regulates, which is one reason such capacities evolve. Nothing in any of this bears on experience.

32.1 What Might Be

Lewis gave counterfactuals a semantics in terms of possible worlds: "if A were the case, C would be the case" is true when some world in which A and C both hold is closer to the actual world than any world in which A holds and C does not (Lewis 1973). The account makes counterfactuals depend on a similarity ordering of worlds, and it leaves the ordering to be fixed by context.

Pearl replaced the similarity ordering with a model of mechanisms (Pearl 2009).

Definition 32.1 (Structural causal model)**.** A *structural causal model* consists of exogenous variables U with a distribution $\mathbb{P}(U)$, endogenous variables $V_1, \dots, V_n$, and one structural equation for each endogenous variable, $V_i = f_i(\mathrm{pa}_i, U_i)$, where pa_i are the variables that directly cause V_i. An *intervention*

$\mathrm{do}(X{=}x)$ replaces the equation for X by the constant x and leaves the others unchanged.

Pearl's causal hierarchy distinguishes three kinds of question. Association asks about $\mathbb{P}(y \mid x)$: what do we expect to see, having seen x? Intervention asks about $\mathbb{P}(y \mid \mathrm{do}(x))$: what would happen if we set X to x? Counterfactuals ask about a particular case that has already happened: given that we saw $X = x'$ and $Y = y'$, what would Y have been had X been x?

Definition 32.2 (Counterfactual evaluation)**.** Given evidence e, the value of Y had X been x, written $Y_x \mid e$, is computed in three steps: (i) *abduction*, updating $\mathbb{P}(U)$ to $\mathbb{P}(U \mid e)$; (ii) *action*, applying $\mathrm{do}(X{=}x)$; (iii) *prediction*, computing Y in the modified model with the updated exogenous distribution.

The three steps make explicit what a counterfactual requires. A system that evaluates counterfactuals needs a model of mechanisms (the equations), a way to condition on what actually happened (abduction), and a way to modify the model without modifying the world (action). The last requirement is where counterfactual representation parts company with prediction. A forward model predicts the consequence of the command about to be issued; a counterfactual evaluation predicts the consequence of a command that will not be issued, in a situation fixed by what has already happened.

Example 32.3 (The jamming avoidance response as a causal model)*.* Write the fish's own frequency and its neighbor's as exogenous, $f_{\mathrm{s}} = U_{\mathrm{s}}$ and $f_{\mathrm{e}} = U_{\mathrm{e}}$. The frequency difference is $\Delta f = f_{\mathrm{e}} - f_{\mathrm{s}}$; the midbrain reading is $z = \mathrm{sgn}(\Delta f)$, corrupted with small probability; the frequency shift is $a = -\kappa z$ with $\kappa > 0$ (Figure 32.1). Suppose we observed a neighbor below the fish and a rise in the fish's frequency. Had the neighbor been above, the model says, the fish would have lowered its frequency; and had we clamped the reading at $z = +1$ by stimulating the relevant pathway, the fish would have lowered its frequency whatever the neighbor did. The second statement is an intervention on an internal variable. The electrophysiology of Chapter 21 and the lesion and stimulation experiments on the motor pathways (C. H. Keller et al. 1990; Metzner 1993b) are real interventions of exactly this kind. Every one of these statements is made by us, about the fish. The fish's own circuit contains no variable for the neighbor it did not have.

The distinction the example draws, between a system that can be described by counterfactuals and a system that evaluates them, is the subject of the rest of the chapter. The first is true of every implemented computation. The second is a specific capacity with specific requirements.

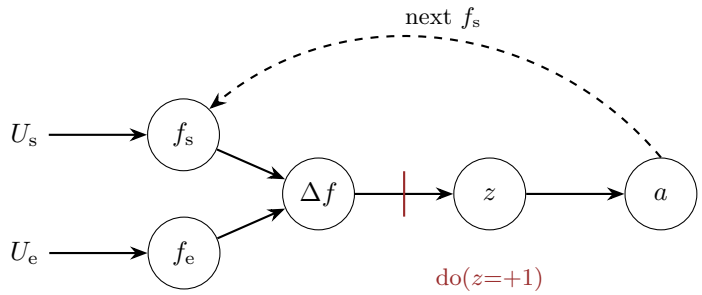


Figure 32.1: A structural causal model of the jamming avoidance response (Example 32.3). The red bar marks the arrow removed by an intervention on the midbrain reading; the dashed arrow closes the loop through the fish's next frequency.

32.2 Prospective Control

The weakest form of acting on what is not yet actual is prospective control: acting on a predicted state rather than on a current error. The forward models of Chapter 31 support it. A reaching movement can be corrected before sensory feedback arrives because the forward model's estimate runs ahead of the delayed feedback (Wolpert et al. 1995; Kawato 1999). Prospective control uses one prediction, the consequence of the action being taken. It evaluates no alternatives, and it involves no abduction about a past that might have gone otherwise. It is the ground floor on which counterfactual capacities are built, and it is common: any controller with a predictive component performs it.

32.3 Planning in a Model

Planning adds alternatives.

Definition 32.4 (Planning)**.** A system *plans* when it selects an action sequence by generating the predicted outcomes of several alternative sequences in an internal model and comparing them, where most of the alternatives are never executed.

Planning is counterfactual in the interventional sense of Pearl's hierarchy: it asks what would happen if the system did this rather than that, from the current state. It needs a model that can be run faster than the world, or at least run without the world.

Machine learning supplies a clean existence proof that the capacity requires no more than a learned generative model and a controller. In one of Ha and Schmidhuber's experiments, an agent first learned a compressed model of its environment's observations and dynamics, then trained its controller entirely inside that model, and finally transferred the controller to the actual environment (Ha and Schmidhuber 2018). The agent's model generated sequences

of observations that never occurred, and the controller learned from them. Whatever else the result shows, it shows that evaluating unexecuted actions in a learned model is an engineering capacity of systems no one supposes to be experiencing anything.

32.4 Deliberation in the Hippocampus

Rats at a difficult choice point often pause and orient back and forth between the options before committing, a behavior called vicarious trial and error (Redish 2016). Johnson and Redish decoded the rat's represented position from ensembles of hippocampal CA3 neurons during such pauses and found that the decoded position transiently swept forward of the animal along the paths ahead of it (Johnson and Redish 2007). The represented locations were places the rat was not at, on routes it might or might not take.

The sweeps meet the definition of planning only partly. They show that the hippocampus generates representations of unexecuted paths at the moment a choice is made. They do not by themselves show that the outcomes along those paths are evaluated and compared, which is what Definition 32.4 requires; that part of the process involves downstream circuits whose roles are still being worked out (Redish 2016). Together with the internally generated delay sequences of Chapter 31 (Pastalkova et al. 2008), they show that the mammalian hippocampus produces detached, forward-directed representations of the kind planning needs.

32.5 Mental Time Travel

Human memory is constructive: remembering an episode reassembles it from parts, and the reassembly is error-prone. Schacter, Addis, and Buckner argued that the constructive character of memory is a feature and not a flaw, because the same core brain network that supports remembering the past also supports imagining the future, recombining elements of past experience into simulations of events that have not happened (Schacter et al. 2007). Suddendorf and Corballis called the combined capacity *mental time travel* and argued that its future-directed form, foresight, may be uniquely human, since the evidence for it in other animals admits simpler explanations (Suddendorf and Corballis 2007).

The claim of uniqueness is contested; evidence of deliberation in rodents, including the sweeps of the previous section, bears on it (Redish 2016). The dispute concerns how rich an animal's model of the future is: whether it represents particular future episodes, located in time and involving itself, or only generic expectations about what follows what. That is a question about capacity, and it can be settled by behavioral and physiological evidence. It is not, as sometimes presented, a question about experience.

32.6 Why Controllers Need Models

Two results from control theory explain why internal models, and with them counterfactual capacities, should be expected in systems that regulate well.

Conant and Ashby proved that, in a formal setting in which a regulator acts to minimize the variability of outcomes, the simplest optimal regulator is one whose actions are a function of the state of the system it regulates; they summarized the result as "every good regulator of a system must be a model of that system" (Conant and Ross Ashby 1970). Read literally, the theorem says that the optimal regulator's behavior mirrors the regulated system's states. Whether that mirroring deserves the name "model" in the richer sense of this chapter is a matter of definition, and the theorem does not require the regulator to evaluate alternatives.

The internal model principle of Francis and Wonham is sharper (Francis and Wonham 1976). A linear controller that regulates robustly, driving the error to zero for a whole class of disturbance or reference signals and continuing to do so when the plant's parameters change, must contain a model of the dynamics that generate those signals. The simplest instance is familiar.

Proposition 32.5 (Integral action for step disturbances)**.** *Consider a stable linear feedback loop with plant transfer function $P(s)$ and controller $C(s)$, subject to a step disturbance of size d entering at the plant output. The steady-state error is $e_\infty = d/(1 + P(0)C(0))$. It vanishes for every d only if the loop gain is infinite at $s = 0$, that is, only if the loop contains an integrator; if the integrator is in the controller, the error vanishes whatever the plant's parameters, provided the loop remains stable.*

Proof. The error is $E(s) = \frac{1}{1+P(s)C(s)} \cdot \frac{d}{s}$. By the final value theorem, which applies because the loop is stable, $e_\infty = \lim_{s\to 0} sE(s) = d/(1 + P(0)C(0))$. This is zero for all d exactly when $P(0)C(0)$ is infinite, which for rational transfer functions means a pole at $s = 0$. A pole of $C(s)$ at the origin gives an infinite loop gain for any plant with $P(0) \neq 0$, so the conclusion holds for every such plant. □

A step is the output of the simplest signal generator, an integrator with no input, and the integrator in the controller is a copy of it: a model of constant signals. The bacterium of Chapter 28 has exactly this structure. Its chemotaxis network returns to baseline after any step in attractant concentration, robustly to large changes in its protein levels (Alon et al. 1999), and Yi and colleagues showed that the robustness requires integral feedback, which the methylation system implements (Yi et al. 2000). In the sense of the internal model principle, the bacterium carries a model of one class of environmental changes. It uses the model to cancel them, which is prediction of the most rudimentary kind, and it evaluates no alternatives.

The two results are the prior art for a claim sometimes presented as new: that adaptive systems must model their environments. Control theory established a precise version of the claim decades ago, and it established the limits of that version too. It says that a good regulator contains a model of the signals it must cancel or track. It does not say that the model represents alternatives, or that it represents anything as absent, and so it stops short of counterfactual representation.

32.7 Counterfactual Capacity Without Experience

The capacities of this chapter form a sequence: a controller with an internal model of disturbances (the bacterium), a controller that predicts the consequences of its own commands (the reaching arm), a system that generates forward sequences along unexecuted paths (the rat's hippocampus), a system that plans in a learned model (the trained agent), and a system that constructs particular future episodes involving itself (the human, on Suddendorf and Corballis's account). Each step adds a requirement from Definition 32.2: a model, then alternatives, then abduction about the particular case, then a self located in the imagined episode.

Two of the systems on the list are machines, and nothing in the requirements distinguishes them from the animals. Counterfactual capacity is a functional property with a graded structure, and it can be measured by behavior and physiology. The agent-centered variables of Part VIII build on it, since a variable of the form "what would happen if I did this" is counterfactual and indexed to the system at once. Whether any of it is accompanied by experience is the question of Part X, and the capacities in this chapter give no purchase on it.

32.8 Audit

Canonical phenomenon

Rats pausing at a choice point, with hippocampal CA3 representations sweeping forward along the paths ahead (Johnson and Redish 2007; Redish 2016).

Minimal model

Structural causal models and counterfactual evaluation by abduction, action, and prediction (Definitions 32.1 and 32.2); planning as model-based comparison of unexecuted alternatives (Definition 32.4); integral action as the internal model of step disturbances (Proposition 32.5).

What is relational here?

A relation between an action and its consequence in a situation that is not

actual: the system represents "the world relative to an action I might take," a variable that is both counterfactual and indexed to the system.

What physically reads it?

Hippocampal circuits that generate forward sweeps and delay-period sequences (Johnson and Redish 2007; Pastalkova et al. 2008), and downstream decision circuits whose roles are still being established (Redish 2016); in machines, learned world models and controllers (Ha and Schmidhuber 2018); in bacteria, the methylation integrator (Yi et al. 2000).

What existing field explains it?

Possible-worlds semantics (Lewis 1973), structural causal models (Pearl 2009), the good-regulator theorem and the internal model principle (Conant and Ross Ashby 1970; Francis and Wonham 1976), the neuroscience of deliberation (Redish 2016; Johnson and Redish 2007), the cognitive science of prospection (Schacter et al. 2007; Suddendorf and Corballis 2007), and model-based machine learning (Ha and Schmidhuber 2018).

What remains unexplained?

The extent of future-directed mental time travel in nonhuman animals is contested (Suddendorf and Corballis 2007). No experiential residue arises.

A discriminating experiment

Separate model-based from cached control by changing the value of an outcome without letting the animal re-experience the actions that lead to it. A system that evaluates unexecuted actions in a model changes its choice at once; a system that caches the values of past actions does not. Recording hippocampal sweeps before the first post-revaluation choice tests whether the sweeps carry the evaluation or only the paths.

32.9 Notes and References

Lewis (1973) is the classic semantics of counterfactuals; Pearl (2009) develops structural causal models, the causal hierarchy, and the three-step evaluation of counterfactuals. On deliberation in rats, Redish (2016) reviews vicarious trial and error and its neural basis, and Johnson and Redish (2007) reports the forward sweeps in CA3; Pastalkova et al. (2008) reports internally generated delay sequences. Schacter et al. (2007) present the prospective function of constructive memory; Suddendorf and Corballis (2007) argue for the distinctiveness of human foresight. The good-regulator theorem is Conant and Ross Ashby (1970) and the internal model principle Francis and Wonham (1976); the application to bacterial chemotaxis is Yi et al. (2000). Ha and Schmidhuber (2018) is an accessible demonstration of planning inside a learned world model.

32.10 Exercises

Exercise 32.1. In the model $X = U_X$, $Y = 2X + U_Y$, we observe $X = 1$ and $Y = 3$. Compute Y had X been 2, following the three steps of Definition 32.2. Compare with the interventional expectation $\mathbb{E}[Y \mid \mathrm{do}(X{=}2)]$ when U_Y has mean zero, and explain the difference.

Exercise 32.2. For the model of Example 32.3, with a reading that reports the correct sign with probability 0.95, compute $\mathbb{P}(a < 0 \mid \mathrm{do}(f_{\mathrm{e}} = f_{\mathrm{s}} + 5\,\mathrm{Hz}))$ and $\mathbb{P}(a < 0 \mid \mathrm{do}(z = +1))$. Which of these could an experimenter measure, and how?

Exercise 32.3. Extend Proposition 32.5 to a ramp disturbance $d(t) = kt$. Show that zero steady-state error requires a double pole at $s = 0$ in the loop gain, and interpret the result as an internal model of ramps. What would a bacterium need in order to reject ramps in attractant concentration exactly, and is there any reason it should?

Exercise 32.4. Give three systems that can be described by true counterfactuals but do not represent counterfactuals, and one that does. For the last, identify the model, the alternatives it evaluates, and the evidence that it evaluates them.

Exercise 32.5. Design the revaluation experiment of the audit for a species of your choice. Specify the outcome, the method of changing its value, the controls, and the prediction of each hypothesis.

Exercise 32.6. Construct a counterfactual about the jamming avoidance response that Lewis's closest-world semantics and Pearl's structural surgery might evaluate differently, for instance one that reasons backward from an effect to a cause ("had the fish lowered its frequency, the neighbor would have been higher"). Which evaluation fits the way experimenters use the model?

33

Generative Models

The fullest form of detachment is a model that can produce the input it expects. A forward model predicts the consequence of an action (Chapter 31); a planning system evaluates actions it will not take (Chapter 32); a *generative model* goes further and explains the input itself, positing hidden causes whose predicted consequences are compared with what arrives. Perception, on this view, is inference to the causes of sensation, and action is a way of making sensation match what the model predicts.

The thesis of this chapter is that this family of theories (Bayesian perception, predictive coding, and active inference) explains a broad range of perceptual and motor phenomena and gives the reading loop its most complete computational form, including agent-centered variables. Whether its most general statement, the free-energy principle, is an empirical theory or a modeling framework is disputed. None of these theories explains why inference should be experienced, although several of their proponents have proposed links, and we set out what the theories predict and where the explanatory gap remains.

33.1 Perception as Inference

Helmholtz argued that perception is the result of unconscious inference: the visual system arrives at the most probable cause of its sensations by a process that resembles reasoning but is not experienced as such (Helmholtz 1867). The modern form of the idea is Bayesian. A hidden cause h produces sensory data y with likelihood $p(y \mid h)$; the system's prior over causes is $p(h)$; the posterior is

$$p(h \mid y) = \frac{p(y \mid h)\, p(h)}{p(y)}. \tag{33.1}$$

The pair $p(y \mid h)$ and $p(h)$ is the generative model: it says how causes produce data and how probable the causes are. Knill and Pouget argued that the brain represents sensory information probabilistically and that much of perception and sensorimotor control can be understood as approximately optimal inference of this kind (Knill and Pouget 2004).

The clearest quantitative success of the approach is cue combination.

Model 33.1 (Gaussian cue combination)**.** Two sensory cues report a hidden quantity h with independent Gaussian noise, $y_i = h + n_i$, $n_i \sim \mathcal{N}(0, \sigma_i^2)$, and the prior over h is flat over the relevant range. The posterior is Gaussian with mean and variance

$$\hat{h} = \frac{\sigma_2^2\, y_1 + \sigma_1^2\, y_2}{\sigma_1^2 + \sigma_2^2}, \qquad \sigma_{\hat{h}}^2 = \frac{\sigma_1^2\, \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \le \min(\sigma_1^2, \sigma_2^2). \tag{33.2}$$

Each cue is weighted by its reliability, the inverse of its variance, and the combined estimate is more reliable than either cue alone.

Example 33.2 (Seeing and feeling the size of an object). Ernst and Banks asked people to judge the height of a ridge that they saw and touched at once, while degrading the visual cue by adding noise. As the visual cue became less reliable, judgments shifted toward the haptic cue, and the weights and the precision of the combined judgments were close to those predicted by (33.2) from the reliabilities of the cues measured separately (Ernst and Banks 2002). The model's parameters were measured independently and the combination was predicted, which is the pattern of a quantitative theory.

33.2 Predictive Coding

Bayesian inference says what the answer should be. Predictive coding proposes how a hierarchy of neurons could compute it.

Rao and Ballard built a hierarchical network in which each level sends to the level below a prediction of that level's activity, and each level sends upward only the residual error between its activity and the prediction (Rao and Ballard 1999). Trained on natural images, the network's units developed receptive fields resembling those of visual cortex, and some of them showed "extra-classical" effects such as end-stopping, a suppression of response when a line extends beyond the receptive field. In the model the suppression has a reason: a long, predictable line is predicted from above, so less error is left to report.

Model 33.3 (A predictive-coding hierarchy)**.** At each level, a vector of causes $\boldsymbol{r}$ generates a prediction $f(\boldsymbol{U}\boldsymbol{r})$ of the activity $\boldsymbol{y}$ at the level below. Error units carry $\boldsymbol{\varepsilon} = \boldsymbol{y} - f(\boldsymbol{U}\boldsymbol{r})$. The causes are updated to reduce the precision-weighted error, together with a penalty from the level above, and the generative weights $\boldsymbol{U}$ are learned slowly in the same direction:

$$\dot{\boldsymbol{r}} \propto \boldsymbol{U}^{\top}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_{\text{above}}, \qquad \Delta\boldsymbol{U} \propto \boldsymbol{\varepsilon}\,\boldsymbol{r}^{\top}$$

(for the linear case $f(x) = x$; nonlinear versions include the derivative of f).

Figure 33.1 shows the arrangement. Predictions descend and errors ascend. The feedforward signal is a relation, the difference between what arrived and

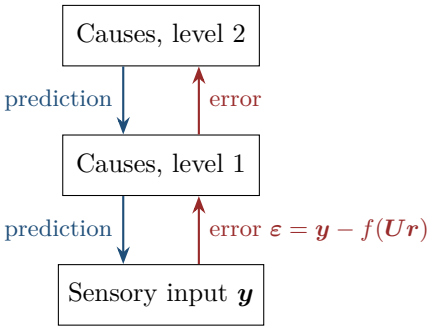


Figure 33.1: A two-level predictive-coding hierarchy (Model 33.3). Top-down connections carry predictions; bottom-up connections carry the residual error. The ascending signal is a relation between input and expectation.

what was expected, and the feedback signal is the model's current account of the causes.

Clark and Hohwy developed predictive coding into a general account of perception, attention, and action, on which the brain is above all a prediction machine and the contents of perception are the brain's best current hypotheses about the causes of its input (Clark 2013; Hohwy 2013). The general account goes well beyond what has been demonstrated physiologically; whether cortex contains distinct prediction and error units, and which cells they would be, remains an active research question.

33.3 Free Energy and Active Inference

Friston generalized the scheme into a principle: self-organizing systems that persist must, on average, minimize a quantity called variational free energy, which bounds the improbability of their sensory states (K. Friston 2010).

Definition 33.4 (Variational free energy)**.** For a generative model $p(y, h)$ and any distribution $q(h)$ over hidden causes, the *variational free energy* of data y is

$$\mathcal{F}[q] = \mathbb{E}_q\big[\ln q(h) - \ln p(y, h)\big].$$

Proposition 33.5 (Free energy bounds surprise)**.** $\mathcal{F}[q] = D_{\mathrm{KL}}(q(h) \,\|\, p(h \mid y)) - \ln p(y) \geq -\ln p(y)$*, with equality if and only if* $q(h) = p(h \mid y)$*.*

Proof. Write $p(y, h) = p(h \mid y)\, p(y)$. Then $\mathcal{F}[q] = \mathbb{E}_q[\ln q(h) - \ln p(h \mid y)] - \ln p(y) = D_{\mathrm{KL}}(q \,\|\, p(\cdot \mid y)) - \ln p(y)$. The Kullback–Leibler divergence is non-negative and vanishes exactly when the two distributions coincide (Cover and Thomas 2006). □

The proposition gives the scheme its structure. Minimizing $\mathcal{F}$ with respect to q is approximate Bayesian inference: q approaches the posterior. Minimizing

$\mathcal{F}$ with respect to the model's parameters is learning. And since $-\ln p(y)$ measures how surprising the data are under the model, a system can also reduce $\mathcal{F}$ by changing y: by acting so that its sensory states become the ones its model expects. That last move is *active inference*, developed at book length by Parr, Pezzulo, and Friston (Parr et al. 2022). In it, action and perception serve one objective, and the distinction between them becomes a distinction between two ways of reducing the same quantity.

The status of the principle is contested. Its proponents present it as a first principle from which the properties of living and cognitive systems follow. Critics argue that, stated generally enough to apply to every persisting system, it becomes a mathematical framework that can redescribe almost any behavior and so makes no empirical claim by itself, and that the step from the formalism to claims about mechanisms needs separate evidence (Colombo and Wright 2021; Andrews 2021). The book's method takes no side in the dispute and asks of any particular application what it predicts that the prior art does not.

33.4 Hidden States and Precision

A generative model's variables are its hidden states: the posited causes that explain its input. Their role connects this chapter to the theory of content of Chapter 29. On a predictive account, a state is about the hidden cause that it stands for in the model, and it is correct to the extent that the model's predictions succeed; the distal-proximal problem is answered by the structure of the model, since a model that explains skin modulations by positing a neighbor has a variable for the neighbor.

The errors that drive inference are weighted by their expected reliability, their *precision*, the inverse variance that appears in (33.2). Precision weighting is how the scheme decides which errors to trust, and active-inference theories identify attention with the optimization of precision (Parr et al. 2022). The claim unifies a great deal, and it also shows the scheme's flexibility: most effects that an alternative theory would attribute to a separate attentional mechanism can be attributed to precision, which makes the attribution hard to test.

33.5 Action as Inference

In active inference, a movement is produced by predicting its proprioceptive consequences and letting reflexes cancel the resulting error; the body moves until it matches the prediction. The same machinery handles reafference. Self-produced touch is predicted from the motor command and therefore attenuated, which is why it tickles less than the same touch produced by someone else (Blakemore et al. 1998). In the vocabulary of Chapter 22, the model supplies the expected reafference and the error carries the exafference.

Example 33.6 (Rewriting the jamming avoidance response). The jamming avoidance response can be rewritten as active inference. Give the fish a generative model in which beat rates below some value are improbable, so that a slow beat produces prediction error; the fish then acts, changing f_{s}, until the beat is fast enough to be expected. The rewrite reproduces the response, and it attaches the response to a general principle. Whether it adds anything depends on what it predicts that Heiligenberg's algorithm, which reads the sign of Δf from the joint modulation of amplitude and differential phase without any model of the neighbor (Heiligenberg 1991a), does not. If it predicts nothing new, it is a redescription, and by the method of this book redescriptions earn no credit. Exercise 33.4 asks the reader to look for a prediction that would separate the two.

33.6 Allostasis and Interoceptive Inference

The generative model need not be about the external world. Sterling contrasted homeostasis, which corrects deviations after they occur, with *allostasis*, regulation that anticipates needs and adjusts physiology before the deviation arrives, and argued that efficient regulation is predictive in this sense (Sterling 2012). The viability of Chapter 28 is then maintained by prediction as well as by correction.

Seth extended predictive processing to the interior of the body: on his account, the brain's predictions about interoceptive signals (heartbeat, breathing, visceral state) underlie emotion and the sense of being an embodied self (Seth 2013). The proposal connects the generative-model framework to the self-models of Chapter 36 and Chapter 41, and it is one of the routes by which proponents of the framework approach consciousness.

33.7 What Generative Models Explain, and What They Do Not

The framework's explanatory record is substantial. It predicts quantitative cue combination (Ernst and Banks 2002); it accounts for classes of receptive-field effects in terms of prediction (Rao and Ballard 1999); it explains the attenuation of self-produced sensation (Blakemore et al. 1998); it unifies perception, learning, and action under a single objective (K. Friston 2010; Parr et al. 2022); and it produces agent-centered variables without extra machinery, because a model that predicts the consequences of the agent's own actions has variables indexed to that agent. It is, of the theories in this book, the most complete computational statement of the reading loop.

It does not explain why any of this inference is experienced. Three observations show why the gap is real. First, Helmholtz's inferences were unconscious

by his own description (Helmholtz 1867), and most of the inference the framework attributes to the brain proceeds without any reportable experience of it. Second, engineered systems perform the same computations: a Kalman filter is exact Bayesian inference for a linear Gaussian model (Chapter 31), and variational inference runs in ordinary software. Third, the proposals within the tradition that do address experience tie it to particular features of the model rather than to inference as such: to the precision and depth of the hierarchy, to interoceptive self-modeling (Seth 2013), or to the dynamics of competing hypotheses, as in Hohwy's treatment of binocular rivalry (Hohwy 2013).

Those proposals are valuable, and they are testable to the extent that they say which conscious contents occur when. Explaining why one hypothesis rather than another is experienced, when both are being entertained, is an achievement. It leaves open why any hypothesis is experienced at all. Chapter 43 states that gap precisely and Chapter 44 compares the theories that claim to cross it, including those built on predictive processing.

33.8 Audit

Canonical phenomenon

People combine a seen and a felt estimate of an object's size, weighting each by its reliability, close to the statistically optimal combination (Ernst and Banks 2002).

Minimal model

Bayesian inference (33.1) and Gaussian cue combination (33.2); a predictive-coding hierarchy (Model 33.3); variational free energy as a bound on surprise (Proposition 33.5).

What is relational here?

The prediction error, a difference between what the model expects and what arrives, is a relation between the system's model and its input; the model's hidden states are posited causes, relata that the system supplies.

What physically reads it?

In predictive-coding models, error units whose activity is the difference between input and top-down prediction, and cause units updated by the errors (Rao and Ballard 1999). Whether and how cortex implements them is under investigation.

What existing field explains it?

Bayesian perception (Helmholtz 1867; Knill and Pouget 2004), predictive coding (Rao and Ballard 1999; Clark 2013; Hohwy 2013), the free-energy principle and active inference (K. Friston 2010; Parr et al. 2022), estimation theory (the Kalman filter as exact Bayesian inference), and the physiology of allostasis (Sterling 2012). They explain the phenomena of the chapter.

What remains unexplained?

The empirical status of the free-energy principle is disputed (Colombo and Wright 2021; Andrews 2021). The framework does not explain why inference is experienced; its proposals about consciousness concern which contents are experienced.

A discriminating experiment

Hold the input fixed and manipulate expectation. Predictive coding predicts a response when an expected stimulus is omitted, since the omission is itself a prediction error, and it predicts that responses to expected stimuli are reduced even when expectation is dissociated from repetition, with effects that scale with the precision assigned to the expectation. A feedforward account without prediction predicts no response to an omission, and an account based on adaptation predicts reduction by repetition but not by expectation. Cue-combination designs that measure each cue's reliability separately and predict the combined weights (Example 33.2) extend the test from perception to any modality.

33.9 Notes and References

Helmholtz (1867) is the historical source of perception as unconscious inference. Knill and Pouget (2004) is the standard statement of the Bayesian brain hypothesis; Ernst and Banks (2002) is the classic demonstration of near-optimal cue combination. Rao and Ballard (1999) introduced hierarchical predictive coding in its modern form; Clark (2013) and Hohwy (2013) develop it into a general theory of mind. The free-energy principle is set out in K. Friston (2010) and active inference in Parr et al. (2022); Colombo and Wright (2021) and Andrews (2021) are careful critiques of the principle's status. Sterling (2012) is the source for allostasis as predictive regulation, and Seth (2013) for interoceptive inference. For the information-theoretic identities used in Proposition 33.5, see Cover and Thomas (2006).

33.10 Exercises

Exercise 33.1. Derive (33.2) by multiplying the two Gaussian likelihoods. For $\sigma_1 = 1$ and $\sigma_2 = 2$, compute the weights and the variance of the combined estimate, and the variance reduction relative to the better cue. What result would show that an observer is combining the cues without weighting them by reliability?

Exercise 33.2. Prove Proposition 33.5. Then show that minimizing $\mathcal{F}$ over Gaussian q for the model of Model 33.1 recovers the exact posterior.

Exercise 33.3. Simulate a one-level linear version of Model 33.3 with a fixed $\boldsymbol{U}$ and a Gaussian prior on $\boldsymbol{r}$. Show that the equilibrium $\boldsymbol{r}$ is the posterior mean, and plot the error signal over time after the input changes abruptly.

Exercise 33.4. Write the jamming avoidance response as active inference, following Example 33.6. Find one prediction the rewrite makes that the sign-reading algorithm of Chapter 21 does not, for instance about the fish's response to a beat whose rate is held constant while its sign is varied, or about the time course of the response. If you find none, state what follows for the rewrite's explanatory standing.

Exercise 33.5. Show that the update of the Kalman filter in Chapter 31 is the posterior mean for a Gaussian prior (the prediction) and a Gaussian likelihood (the measurement). Which quantity plays the role of precision?

Exercise 33.6. State the free-energy principle in a form that makes an empirical prediction about a particular system, and a form that makes none. For the first, describe an observation that would count against it.

Exercise 33.7. A regulated variable is disturbed by a load that arrives on a known daily schedule. Compare a purely reactive proportional controller with one that adds a feedforward term anticipating the load. Compute the mean squared deviation of each and relate the result to Sterling's distinction between homeostasis and allostasis.

Memory, prediction, emulation, and generative models let a system carry relations that are not currently present in its input. The reader is no longer bound to the current mixture; it can hold a relation that might obtain. Bayesian inference, predictive processing, and active inference formalize this, and they explain a great deal about perception and action. Part VIII asks what makes the variables of such a model centered on the system that holds it.

Part VIII

From World-Centered to Agent-Centered Variables

Why are the relevant variables centered on this organism?

34

Here

> We propose that seeing is a way of acting. It is a particular way of exploring the environment.
>
> J. Kevin O'Regan and Alva Noë (2001)

A variable is agent-centered when the system that reads it is one of its relata. The distance from my hand to the cup, the angle of a light relative to my line of gaze, and the distortion of my own electric field by a nearby stone are all of this kind: change the agent and the variable changes, even if the world stands still. Coordinates with the agent at the origin are the simplest agent-centered variables, and nervous systems build and transform them continuously.

The thesis of this chapter is that "here" names such a coordinate system, and that the mechanisms constructing it are well understood and owe nothing to consciousness. Active sensing produces agent-centered variables by construction, before any self-model exists, because the environmental variable is measured as a modulation of a signal the agent itself emits. Location alone does not produce them: a rock has a location and no variable centered on it is read by anything. The chapter ends with a minimal, functional definition of a perspective, cheap enough that a thermostat satisfies it, and uses that cheapness to mark how much of the work of Part X is still to be done.

34.1 Coordinates with an Origin

A physical scene can be described in two ways. An *allocentric* description gives the positions of things in a frame anchored to the world: the landmark is at these coordinates, the nest at those. An *egocentric* description gives the positions of things relative to the agent: the landmark is two meters ahead and slightly to the left. Both are relational in the sense of Chapter 1. The difference lies in which relata they involve.

Definition 34.1 (Pose and egocentric coordinates)**.** Let features of a scene have positions $\boldsymbol{p}_j \in \mathbb{R}^d$ in a world frame, and let the agent's *pose* be the pair

$(\boldsymbol{R}, \boldsymbol{t})$, a rotation $\boldsymbol{R} \in SO(d)$ giving its orientation and a vector $\boldsymbol{t} \in \mathbb{R}^d$ giving its position. The *egocentric coordinates* of feature j are

$$\boldsymbol{q}_j = \boldsymbol{R}^\top (\boldsymbol{p}_j - \boldsymbol{t}). \tag{34.1}$$

A variable is *allocentric* if it is a function of the world positions $\boldsymbol{p}_j$ alone, and *egocentric* if it is a function of the egocentric coordinates $\boldsymbol{q}_j$.

The two kinds of variable respond differently to two kinds of motion.

Lemma 34.2 (What moves what)**.** *Egocentric and allocentric variables respond to motion as follows.*

1. *Under a rigid motion applied to the whole scene with the agent included ($\boldsymbol{p}_j \mapsto \boldsymbol{S}\boldsymbol{p}_j + \boldsymbol{u}$, $\boldsymbol{t} \mapsto \boldsymbol{S}\boldsymbol{t} + \boldsymbol{u}$, $\boldsymbol{R} \mapsto \boldsymbol{S}\boldsymbol{R}$ for a rotation $\boldsymbol{S}$ and translation $\boldsymbol{u}$), every egocentric variable is unchanged.*
2. *Under a motion of the agent alone, every allocentric variable is unchanged, and egocentric variables change.*

Proof. For the first claim, $\boldsymbol{q}_j' = (\boldsymbol{S}\boldsymbol{R})^\top (\boldsymbol{S}\boldsymbol{p}_j + \boldsymbol{u} - \boldsymbol{S}\boldsymbol{t} - \boldsymbol{u}) = \boldsymbol{R}^\top \boldsymbol{S}^\top \boldsymbol{S}(\boldsymbol{p}_j - \boldsymbol{t}) = \boldsymbol{q}_j$, since $\boldsymbol{S}^\top \boldsymbol{S} = \boldsymbol{I}$. The second claim follows because the $\boldsymbol{p}_j$ do not depend on the pose, while $\boldsymbol{q}_j$ does. □

The lemma is the spatial version of the invariance argument of Chapter 1. An agent with no access to the world frame, no map and no satellite fix, cannot read absolute positions at all. What it can read are relations: relations among features, and relations between features and itself. Egocentric coordinates are the second kind, relational observables in which the agent is one of the relata.

It is tempting to treat the egocentric frame as a degraded version of the allocentric one, a view from somewhere where a view from nowhere would be better. For action the order is reversed. A reach is executed by a hand from where the hand is, and a saccade moves the eyes from where they point. A motor command must eventually be expressed in the coordinates of the effector that carries it out, and those coordinates are egocentric by necessity. An allocentric map is useful to an acting system only after the system has located itself in it.

34.2 Chains of Frames

There is no single egocentric frame. A point of light falls on the retina at a retinal position; the eye sits in the head at some angle; the head sits on the body at another; the hand is somewhere else again. Each stage defines its own origin, and moving between them requires a posture signal. In one dimension, with angles measured in the horizontal plane, the chain reads

$$x_{\text{head}} = x_{\text{ret}} + e, \qquad x_{\text{body}} = x_{\text{head}} + h, \qquad x_{\text{hand}} = x_{\text{body}} - b, \tag{34.2}$$

where x_{ret} is the target's position on the retina, e the eye-in-head angle, h the head-on-body angle, and b the hand's position in body coordinates. Each equation is a relation between the target and one part of the agent, computed by adding a relation between two parts of the agent. The target's position relative to the hand is a relation whose relata are the target and the hand, and the nervous system reaches it through a chain of relations among eye, head, body, and limb.

Example 34.3 (One retinal image, two targets). Let a target fall 5° to the left of the fovea, $x_{\text{ret}} = -5°$, with angles to the right counted as positive. If the eyes are turned 20° to the right in the head, $e = +20°$ and the target lies 15° to the right of the head's midline. If the eyes are turned 20° to the left, the identical retinal image belongs to a target 25° to the left of the midline. The retina receives the same input in both cases. Everything that distinguishes the two head-centered relations is carried by the posture signal e.

The arithmetic is trivial. The physical problem is that the quantities in (34.2) are not scalars on wires. Retinal position is carried by a population of neurons with receptive fields tiling the retina, and eye position by neurons whose firing varies with gaze. The question is how a population code for x_{ret} and a population code for e can be combined into something from which a downstream structure can read a function of $x_{\text{ret}} + e$.

34.3 Gain Fields

Andersen, Essick, et al. (1985) recorded from single neurons in the inferior parietal lobule of behaving monkeys and found that the visual sensitivity of their retinotopic receptive fields changed systematically with the angle of gaze. Much of the activity could be described as the product of a gain factor, a function of eye position, and the response profile of the visual receptive field, an operation that produces an eye-position-dependent tuning for locations in head-centered space (Andersen, Essick, et al. 1985). Such a response is called a *gain field*.

Zipser and Andersen (1988) trained a neural network by back-propagation to compute the head-centered location of a stimulus from its retinal location and the position of the eyes. The hidden units of the trained network acquired response properties resembling those of the parietal neurons, which suggested that the gain fields are what a population computing this transformation would be expected to show. Andersen, Snyder, et al. (1997) review the broader evidence: sensory signals from several modalities and efference copies of motor commands converge in posterior parietal cortex and are combined by a gain mechanism into distributed representations usable for converting sensory locations into motor coordinates. Pouget and Sejnowski (1997) gave the population a precise computational role. If each neuron's response is approximately the product of a Gaussian function of retinal location and a

sigmoid function of eye position, the population forms a set of basis functions from which a downstream structure can compute an arbitrary sensorimotor transformation by a direct linear projection, and can represent a location in several reference frames at once.

Why a product? The answer connects this chapter to the physics of Part I.

Theorem 34.4 (Additive separability)**.** *Let $F(x, e)$ be twice continuously differentiable on a rectangle. Then F can be written as $F(x, e) = a(x) + b(e)$ for some functions a and b if and only if $\partial^2 F/\partial x\,\partial e = 0$ throughout the rectangle.*

Proof. If $F = a(x) + b(e)$, the mixed partial derivative is zero. Conversely, if the mixed partial vanishes, then $\partial F/\partial x$ does not depend on e; call it $a'(x)$. Integrating in x gives $F(x, e) = a(x) + b(e)$, where $b(e)$ is the constant of integration, which may depend on e. □

Now let a downstream neuron be tuned to a location in head-centered space, with a Gaussian tuning curve

$$F(x_{\mathrm{ret}}, e) = \exp\!\left(-\frac{(x_{\mathrm{ret}} + e - k)^2}{2\sigma^2}\right). \tag{34.3}$$

Its mixed partial derivative equals the second derivative of the Gaussian at $x_{\mathrm{ret}} + e$, which is nonzero almost everywhere. By Theorem 34.4, no linear readout of units tuned to retinal position alone, added to units tuned to eye position alone, can produce this tuning curve. The readout needs units whose responses depend jointly on both variables, and a product is the simplest joint dependence. The gain field is the parietal counterpart of the square-law detector of Chapter 4. In both cases a relation that is implicit in two separately coded quantities becomes available to a linear stage only after a nonlinear, multiplicative operation has combined them.

Model 34.5 (Gain-field population)**.** Let unit (i, j) respond as

$$f_{ij}(x_{\mathrm{ret}}, e) = \exp\!\left(-\frac{(x_{\mathrm{ret}} - c_i)^2}{2\sigma^2}\right) \frac{1}{1 + \mathrm{e}^{-(e - d_j)/s}}, \tag{34.4}$$

with preferred retinal locations c_i and eye-position thresholds d_j spread over their ranges. A downstream unit reads $z = \sum_{ij} w_{ij} f_{ij}$. With suitable weights w_{ij} the readout approximates head-centered tunings such as (34.3), retinal tunings, or other smooth functions of (x_{ret}, e) (Pouget and Sejnowski 1997). The model is a stylized description of recorded responses (Andersen, Essick, et al. 1985), and its parameters are illustrative.

34.4 Allocentric Maps and a Located Self

Allocentric representations exist too. Hippocampal place cells fire when an animal occupies a particular region of its environment, and O'Keefe and Nadel (1978) proposed that the hippocampus supports a cognitive map in a world-anchored frame. Grid cells of the medial entorhinal cortex fire at the vertices of a triangular lattice spanning the environment (Hafting et al. 2005); how they do so is the subject of Chapter 47. Burgess (2006) reviews the evidence that egocentric and allocentric representations exist in parallel and combine according to the task, with reliance on the allocentric kind growing with the amount of movement between learning and recall, the number of objects to be remembered, and the size, familiarity, and structure of the environment.

An allocentric map still contains an agent-indexed variable: the agent's own position in it. A place cell fires when the animal is in the cell's place field. The map is anchored to the world, and the variable that makes the map usable for action is the relation between the map and the animal. We therefore need a broader term than "egocentric."

Definition 34.6 (Agent-indexed variable)**.** A variable is *agent-indexed* if the state of the agent that reads it (its pose, its emitted signals, its body, or its internal state) is one of the variable's relata.

Egocentric coordinates are agent-indexed, and so is self-location in an allocentric map. The distance between two landmarks is not. The rest of Part VIII adds temporal (Chapter 35), bodily (Chapter 36), and linguistic (Chapter 37) members to the class.

34.5 Active Sensing Builds the Frame

Most discussions of egocentric space start from vision, where light arrives from sources the animal does not control. Active sensing reverses the arrangement. M. E. Nelson and MacIver (2006) define it by the use of self-generated energy to probe the environment, as in echolocation by bats and dolphins and active electrolocation by weakly electric fish, and note that such animals can control the probe's intensity, direction, timing, and spectrum.

The electric fish shows what this does to the measurement variable. The fish's electric organ sets up a field around its body, and electroreceptors in the skin register the field where it crosses the skin. An object with conductivity different from the water distorts the field and casts an *electric image* on the skin, and Lissmann and Machin (1958) showed that the fish locates objects by exactly this distortion of its self-generated field. The images are shaped by the fish's own body and field as much as by the object (Emde 1999; Caputi and Budelli 2006).

Model 34.7 (Probe-relative measurement)**.** Let $s(\boldsymbol{x}, t)$ be the field the fish's own discharge produces at skin location $\boldsymbol{x}$ in the absence of objects, and let nearby objects multiply it by a local modulation $1 + \varepsilon(\boldsymbol{x}, t)$:

$$y(\boldsymbol{x}, t) = s(\boldsymbol{x}, t)\,(1 + \varepsilon(\boldsymbol{x}, t)). \tag{34.5}$$

The receptors respond to y. The environmental variable ε is defined only relative to the probe s, and s depends on the fish's discharge amplitude, its posture, and its position.

Two consequences follow directly from (34.5). First, for a fixed environment, anything the fish does to its own field (bending its body, changing its discharge) changes what the receptors receive. The input is a joint function of the agent's action and the world, which is the defining property of reafference that Chapter 22 develops. Second, the variable the fish needs, the location and nature of the object, is carried by ε, a modulation of the fish's own signal. The measurement variable has the form *world relative to my emission*. It is agent-indexed by construction, before any question of a self-model arises, and it is agent-indexed in a stronger sense than retinal position: the agent contributes one of the relata physically, as the carrier on which the world is written.

The jamming avoidance response of Part V reads a variable of the same form. The frequency difference $\Delta f = f_{\mathrm{e}} - f_{\mathrm{s}}$ between a neighbor's discharge and the fish's own is a relation with the fish as one relatum, and the fish's response changes that relatum: it raises f_{s} when the neighbor is lower and lowers it when the neighbor is higher (Metzner 1993b; Heiligenberg 1991a). The fish reads "the neighbor relative to me" and acts on the "me."

The same point appears in vision once movement is taken into account. In the experiment of Held and Hein (1963), kittens that moved actively through a patterned environment developed normal visually guided behavior, while littermates carried passively through the same environment in a gondola, and so exposed to comparable visual stimulation, did not. O'Regan and Noë (2001) built a theory of vision on the idea that seeing consists in mastery of the laws relating one's movements to the resulting changes in sensory input, which they call sensorimotor contingencies. We do not need their stronger claims about visual consciousness here. We need only the observation behind them: a sensory variable indexed to the agent's own action carries information that the same stimulus, received passively, does not.

34.6 Why Location Is Not Perspective

A rock is located. Its distance to the nearest tree is a well-defined relation with the rock as a relatum, so by Definition 34.6 the rock has agent-indexed variables in the thinnest sense. Nothing in the rock reads them. Location supplies the relation; it supplies no reader.

Three conditions separate a rock from the systems of this chapter.

Definition 34.8 (Minimal functional perspective)**.** A system has a *minimal functional perspective* when

1. some of the variables relevant to it are agent-indexed;
2. those variables are read by physical processes internal to the system, in the sense of Chapter 15; and
3. the readings enter the system's own action or maintenance, closing a reading loop.

Table 34.1 applies the definition. A thermometer reads the temperature at its own location, so it meets the first two conditions, and nothing it reads changes what it does. A thermostat meets all three: it reads the temperature where it is and switches a heater that changes that temperature. By this definition a thermostat has a minimal functional perspective.

	Rock	Thermometer	Thermostat	Fish	Mammal
Agent-indexed variables exist	yes	yes	yes	yes	yes
Read internally	no	yes	yes	yes	yes
Reading enters own action	no	no	yes	yes	yes
Agent supplies the probe	no	no	no	yes	partly
Several frames combined	no	no	no	unknown	yes

Table 34.1: The conditions of Definition 34.8 and two further properties, applied to five systems. "Unknown" marks a property whose presence has not, to our knowledge, been tested.

The result is deliberate. A definition that a thermostat satisfies has isolated a real distinction (rocks fail it) while showing how little that distinction contains. The richer notions of the following chapters add temporal centering (Chapter 35), a classification of signals as self-produced (Chapter 36), indexical content that governs action (Chapter 37), and projective structure (Chapter 38). None of them, alone or together, is yet a point of view in the sense of there being something it is like to occupy it. Chapter 43 states that difference, and the rest of the book keeps the two apart.

34.7 Audit

Canonical phenomenon

Parietal neurons in the monkey respond to a spot of light at a retinal location with a strength set by where the eyes point; the skin of a weakly electric fish registers the distortion of the fish's own field by nearby objects.

Minimal model

Egocentric coordinates $\boldsymbol{q}_j = \boldsymbol{R}^{\top}(\boldsymbol{p}_j - \boldsymbol{t})$ and Lemma 34.2; the additive-

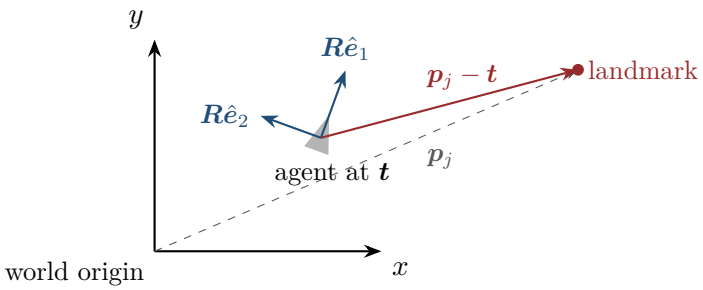


Figure 34.1: Allocentric and egocentric descriptions of one landmark. The dashed vector $\boldsymbol{p}_j$ gives the landmark in the world frame. The solid vector, expressed in the agent's own axes, gives the egocentric coordinates $\boldsymbol{q}_j = \boldsymbol{R}^\top(\boldsymbol{p}_j - \boldsymbol{t})$, which move whenever the agent moves.

separability theorem (Theorem 34.4) and the gain-field population of Model 34.5; probe-relative measurement $y = s(1 + \varepsilon)$ (Model 34.7).

What is relational here?

The position of a feature relative to the agent; head-centered location as the relation between retinal position and eye position; the modulation ε of the fish's field, a relation between world and probe.

What physically reads it?

Parietal neurons whose responses multiply a retinal tuning by an eye-position gain, and downstream populations that read head-centered or effector-centered location from them by linear projection. In the fish, tuberous electroreceptors reading the modulation of the self-generated field.

What existing field explains it?

Sensorimotor and computational neuroscience of coordinate transformations (Andersen, Essick, et al. 1985; Zipser and Andersen 1988; Andersen, Snyder, et al. 1997; Pouget and Sejnowski 1997); the neuroscience of spatial memory (O'Keefe and Nadel 1978; Burgess 2006); the biology of active sensing (M. E. Nelson and MacIver 2006; Lissmann and Machin 1958); and the sensorimotor account of perception (O'Regan and Noë 2001). Together they explain the mechanisms of the chapter.

What remains unexplained?

Nothing in the mechanisms. What remains is a question of what the word "perspective" should mean. In its minimal functional sense it is cheap enough for a thermostat; whether any agent-centered frame is a point of view in the phenomenal sense is untouched by anything here, and is the subject of Chapter 43.

A discriminating experiment

Dissociate agent-indexed from world-anchored coding by moving the agent

while holding the world fixed, and moving the world together with the agent, in a virtual environment; a unit or behavior that follows Lemma 34.2(1) codes egocentrically, one that follows (2) codes allocentrically. For the fish, the corresponding test holds the object fixed and changes the fish's own field, and asks whether behavior follows the receptor input y or the modulation ε.

34.8 Notes and References

Gain fields were reported by Andersen, Essick, et al. (1985); the back-propagation model of Zipser and Andersen (1988) showed that a network trained to compute head-centered location develops them; Andersen, Snyder, et al. (1997) review multimodal convergence in posterior parietal cortex; and Pouget and Sejnowski (1997) develop the basis-function interpretation, including the representation of one location in several frames at once. The cognitive-map theory of the hippocampus is O'Keefe and Nadel (1978); grid cells were described by Hafting et al. (2005); Burgess (2006) reviews how egocentric and allocentric representations combine.

Active sensing is surveyed by M. E. Nelson and MacIver (2006). The founding experiments on electrolocation are Lissmann and Machin (1958); electric images are analyzed by Emde (1999) and Caputi and Budelli (2006), and the monograph of Heiligenberg (1991a) covers the jamming avoidance response. The kitten carousel is Held and Hein (1963). The sensorimotor theory of O'Regan and Noë (2001) and its development in Noë (2004) make stronger claims about perceptual experience than this chapter needs; Gibson (1979) is the classic ecological account of perception as the pickup of information by an active observer.

34.9 Exercises

Exercise 34.1. Write Lemma 34.2 out explicitly in the plane, with $\boldsymbol{R}$ and $\boldsymbol{S}$ as 2×2 rotation matrices. Then show that the distance between two landmarks is invariant under both kinds of motion, and explain why it is nevertheless not agent-indexed.

Exercise 34.2. Compute $\partial^2 F / \partial x_{\mathrm{ret}}\, \partial e$ for the head-centered tuning (34.3) and find where it vanishes. Use Theorem 34.4 to show that no function of the form $a(x_{\mathrm{ret}}) + b(e)$ reproduces (34.3) on any rectangle containing its peak.

Exercise 34.3. Simulate the population of Model 34.5 with 20 retinal centers and 10 eye-position thresholds. Fit linear readouts by least squares to a head-centered Gaussian and to a retinal Gaussian of the same width, and compare their errors. Repeat with the eye-position factor replaced by a constant, and explain the result with Theorem 34.4.

Exercise 34.4. In Model 34.7, suppose the fish's discharge amplitude, and hence s, drifts slowly. Show that the ratio of the input at two nearby skin locations is insensitive to the drift when ε is small, and discuss what this implies about which relations the fish should read. Which relatum does such a ratio discard?

Exercise 34.5. Apply Definition 34.8 to a bat, a satellite-navigation app on a phone, a sunflower, and an automatic door. For each, name the agent-indexed variable, the internal reader if any, and the action if any. Which of them supplies its own probe?

Exercise 34.6. Design a navigation experiment in which egocentric and allocentric strategies predict different errors, using one of the variables that Burgess (2006) identifies as shifting reliance between them (movement between learning and recall, number of objects, or size of the environment). State the prediction of each strategy.

Exercise 34.7. In the carousel experiment of Held and Hein (1963), list the sensory variables available to the passively carried kitten and those available only to the active one. Express the difference in the notation of the reading loop, identifying which term of $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$ the passive kitten lacks.

35

Now

> In short, the practically cognized present is no knife-edge, but a saddle-back, with a certain breadth of its own on which we sit perched, and from which we look in two directions into time.
>
> William James, *The Principles of Psychology* (1890)

Every physical reader integrates its input over some interval, so every reader has a present of finite width. A variable is temporally centered when one of its relata is the reader's own present: the time since the last flash, the time until contact, the recent average, the expected next value. Such variables are the temporal counterpart of the egocentric coordinates of Chapter 34.

The thesis of this chapter is that "now" names three different things, which the literature on time perception often runs together. The first is the integration window that every reader has. The second is temporal indexing, the use of variables centered on the reader's present, which every predictive system performs. The third is a represented present: a variable that stands for the present as such and can be compared with represented pasts and futures. Psychophysics shows that the human present is assembled over windows of tens of milliseconds to a few seconds and that the time assigned to an event can be fixed after the event. None of this requires, or explains, that the present is experienced.

35.1 Every Reader Has a Window

A linear reader turns an input $y(t)$ into an output

$$z(t) = \int_0^\infty h(\tau)\, y(t-\tau)\, \mathrm{d}\tau, \tag{35.1}$$

where the kernel h weights inputs by their age τ. Nonlinear readers have windows too, but the linear case fixes the idea.

Definition 35.1 (Functional present)**.** The *functional present* of a reader at time t is the interval of input times $[t-\tau_1, t-\tau_0]$ that carries a stated fraction $1-\eta$ of the total weight $\int_0^\infty |h(\tau)|\,\mathrm{d}\tau$, where $\tau_0 \geq 0$ is the reader's latency. Its width $T = \tau_1 - \tau_0$ is the *window*.

For a leaky integrator, $h(\tau) = \tau_c^{-1}\mathrm{e}^{-\tau/\tau_c}$, a fraction $1-\mathrm{e}^{-T/\tau_c}$ of the weight falls within the most recent T, so the window that carries 95 percent of the weight is $T \approx 3\tau_c$. The reader's present is a smeared interval ending at its latency, and the smear is set by physics (membrane time constants, synaptic filtering, the dynamics of a population) before any question of experience arises.

A window alone does not determine what the reader can say about time. Temporal order is a relation between two events, and reading it requires a comparison.

Model 35.2 (Reading order from decaying traces)**.** Let events of two kinds, A and B, occur at times t_A and t_B and drive separate leaky traces of equal gain a and time constant τ_c. At a readout time t after both events,

$$z_A(t) = a\,\mathrm{e}^{-(t-t_A)/\tau_c}, \qquad z_B(t) = a\,\mathrm{e}^{-(t-t_B)/\tau_c}, \tag{35.2}$$

and $z_A > z_B$ exactly when A is the more recent event. With independent noise of standard deviation σ on each trace and a small separation $\Delta = t_A - t_B$, the difference is approximately $a\,\mathrm{e}^{-(t-t_B)/\tau_c}\,\Delta/\tau_c$, so the order becomes readable when

$$|\Delta| \gtrsim \frac{\sigma\,\tau_c}{a}\,\mathrm{e}^{(t-t_B)/\tau_c}. \tag{35.3}$$

The parameters are illustrative.

The model makes two points that survive its simplicity. The order of two events is a relation read by comparing two states of the reader, and the threshold for reading it depends on noise, gain, time constant, and the delay before the comparison, not on the window alone. Human psychophysics finds such a threshold: Pöppel (1997) places the order threshold near 30 milliseconds and attributes it to functional system states of that duration, below which successive events are not ordered. The model does not explain the value; it shows what kind of quantity it is.

35.2 The Specious Present

William James took the phrase "specious present" from E. R. Clay and argued that the present we are aware of has duration: a saddle-back with a breadth of its own, from which we look backward and forward in time (W. James 1890, vol. 1, 609). Husserl gave the structure a fuller analysis. In his lectures on internal time consciousness, each moment of experience contains a primal

impression of what is now, a *retention* of what has just elapsed, held as just-past rather than remembered, and a *protention* of what is about to come (Husserl 1991).

Both descriptions are phenomenological, and we keep them as data about experience rather than as mechanisms. They have a functional reading, and the reading is instructive. A reader with a kernel (35.1) carries, in its current state, a decaying trace of its recent input: that trace plays the role of retention. A reader that also predicts, forming an estimate $\hat{y}(t+\Delta \mid y_{\le t})$ of its input a short time ahead, carries an anticipation of the immediate future: that plays the role of protention. The current input plays the role of the primal impression. The structure Husserl described in experience is the structure any predictive filter has in its state.

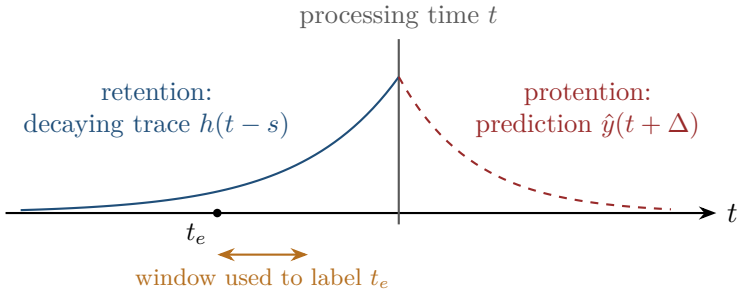


Figure 35.1: A reader's present. The current state holds a decaying trace of the recent past (solid) and a prediction of the near future (dashed). An event at t_e may be assigned its time using input that arrived after it (bracket), so the time a system assigns to an event is fixed only after the event, when its processing is complete.

Psychophysics has measured the widths. Pöppel (1997) proposes a hierarchy: system states of about 30 milliseconds that define successiveness and identify basic events, and an automatic, presemantic integration of successive events into perceptual units of about 3 seconds, which he uses as a pragmatic definition of the subjective present. Wittmann (2011) distinguishes three levels: a *functional moment* in the millisecond range, below which temporal order is not perceived; an *experienced moment* based on integration over up to a few seconds; and *mental presence*, sustained by working memory over multiple seconds. The numbers differ in detail between authors and tasks, and the pattern is stable: a hierarchy of windows, each of which is a property of a reader.

35.3 Postdiction and Time Labels

In the flash-lag illusion, a brief flash presented in line with a moving object appears to lag behind it. Eagleman and Sejnowski (2000) tested two accounts, extrapolation of the moving object's position and a latency advantage for moving objects, and found both inconsistent with their data. They proposed

instead that awareness is *postdictive*: the percept attributed to the time of the flash is a function of events in roughly the 80 milliseconds after the flash. Herzog et al. (2016) go further. They propose that features of objects, including temporal features such as duration, are analyzed unconsciously and quasi-continuously with high temporal resolution, coded as quantitative labels, and rendered conscious together at discrete moments, sometimes hundreds of milliseconds after the stimuli. VanRullen and Koch (2003) review the older and recurring idea that perception operates in discrete epochs and suggest that cortical and thalamocortical oscillations could provide their basis. The discrete-perception hypothesis remains contested, and we do not rely on it. We rely only on a distinction that all of these proposals require.

Definition 35.3 (Three times)**.** For an event registered by a reader, distinguish its *event time* t_e, when it occurred; its *processing time* t_p, when the reader's state encoding it is complete; and its *represented time* $\hat{\tau}$, the time the encoding assigns to the event.

Proposition 35.4 (A labeled present lags)**.** *If a reader assigns $\hat{\tau}$ using input from the interval $[t_e, t_e + W]$, then $t_p \geq t_e + W$.*

Proof. The encoding cannot be complete before the last input it uses has arrived. □

The proposition is trivial, and its consequence is not. A system that labels events accurately in time cannot do so at the moment they occur. Its represented present is assembled after the physical present has passed, and what it represents as "now" is an event time $\hat{\tau}$ assigned from a later processing time. The flash-lag illusion follows from a particular way of assigning $\hat{\tau}$: averaging the moving object's position over a window after the flash places it ahead of the flash. Exercise 35.3 works the arithmetic.

35.4 Temporally Centered Variables

Definition 35.5 (Temporally centered variable)**.** A variable is *temporally centered* if one of its relata is the reader's own processing time: for example, the time elapsed since an event, $t - t_e$, or the expected time until one, $\hat{t}_{\text{next}} - t$.

If every event time and the reader's clock are shifted by the same constant, temporally centered variables do not change and absolute times do. This is the temporal case of the invariance argument of Chapter 1: a system with no access to an external clock can read only temporal relations, among events and between events and its own present. The leaky trace of Model 35.2 is already a temporally centered variable, since its value encodes the time since the event that drove it. So is the state of any system that predicts: the prediction concerns the input a fixed interval after the reader's own present. Predictive

processing (Chapter 33) makes such variables central to perception; for the argument of this chapter it is enough that they are ubiquitous.

Example 35.6 (Time to contact). An animal approaching a wall at constant speed v from distance d will reach it after d/v. The quantity d/v is a relation between the wall and the animal's motion, and it is centered twice: on the animal's position, because d is measured from the animal, and on the animal's present, because the time runs from now. A controller that brakes when d/v falls below a threshold acts on a variable that is agent-indexed in space and in time, and that becomes undefined if the animal stops.

35.5 The Fish's Now

The electric fish shows why a relation can require a present of finite width. In the standard model of Chapter 21 (see also Chapter 20), the sign of the frequency difference $\Delta f = f_\mathrm{e} - f_\mathrm{s}$ is carried by the joint behavior of two modulations at two patches of skin with contamination ratios $m_2 > m_1$: the amplitude $A_2(t)$ at the more contaminated patch and the differential phase $\delta(t) = \varphi_2(t) - \varphi_1(t)$. For small m,

$$\dot{A}_2(t)\,\delta(t) \approx -A_\mathrm{s}\, m_2(m_2 - m_1)\,\Delta\omega\, \sin^2\psi(t), \qquad \psi(t) = \Delta\omega\, t + \psi_0, \quad (35.4)$$

so the product has the sign of $-\Delta\omega$ at every instant except the two per beat cycle at which it vanishes.

Two features of (35.4) bear on time. The variable involves a rate, $\dot{A}_2$, and a rate is a relation between states at neighboring times: no instantaneous snapshot contains it. Any physical reader of the sign therefore has a window. Its magnitude also depends on where in the beat cycle the window falls. A window much shorter than a beat cycle placed near $\psi \approx 0$ or π sees almost nothing, while a window spanning a full cycle averages to $-\frac{1}{2}A_\mathrm{s}m_2(m_2-m_1)\Delta\omega$ regardless of phase. In the presence of receptor noise, a reader that must be reliable whatever the phase needs a window comparable to a beat period. Sign-selective neurons in the midbrain of *Eigenmannia* respond best to beat rates of about 4 to 8 hertz (G. Rose and Heiligenberg 1986b), which corresponds to beat periods of 125 to 250 milliseconds. The fish's reading of the relation has a present on that scale, and its response to what it reads, the slow rise or fall of its discharge frequency (C. H. Keller et al. 1990), unfolds over a longer one. These are inferences from the model and the cited tuning, not measurements of an integration window.

The loop that the reading closes has a present of its own. Madhav et al. (2013) held the jamming stimulus at a fixed frequency offset from the fish's own discharge, so that the relation stayed where the experimenters put it however the fish moved its frequency, and found that a first-order linear system with a delay fitted the behavior near equilibrium. The delay is a temporal property of the whole reading loop, from the mixture on the skin to the change in the

pacemaker. Whatever the fish does about the relation it reads, it does about a relation that held a short time before.

35.6 Does a System Require a Modeled Now?

The three senses of "now" can now be separated.

1. *Integration.* Every physical reader has a functional present (Definition 35.1). This holds of a thermometer, a membrane, and a fish, and costs nothing.
2. *Temporal indexing.* A reader uses temporally centered variables (Definition 35.5). Every system with decaying traces or predictions does this; the fish does it at least in the trivial sense of its traces and in the rate-based reading of (35.4).
3. *A represented present.* A system holds a variable that stands for the present *as* present, one that can be compared with represented past and future events. Human mental time travel, the capacity to re-experience the past and pre-experience the future, requires this (Suddendorf and Corballis 2007), and remembering and imagining draw on shared machinery (Schacter et al. 2007). This grade presupposes detached representation, the subject of Chapters 31 and 32.

The answer to the section's question depends on the grade. Every system has the first. Anything that acts on predictions has the second. The third is rarer, and nothing in the fish's jamming avoidance response requires it.

What none of the three grades supplies is the experienced present that James and Husserl described. A predictive filter has retention and protention in its state and is not thereby aware of a present. The windows that psychophysics measures are windows of processing, and a window of processing can be measured in a system nobody takes to be conscious. The step from a temporally centered variable to a present that is lived through is the step that Chapter 43 isolates, and this chapter does not take it.

35.7 Audit

Canonical phenomenon

In the flash-lag illusion a flash aligned with a moving object appears to lag it, and the percept attributed to the flash depends on what happens in the following 80 milliseconds or so (Eagleman and Sejnowski 2000).

Minimal model

A reader's kernel and functional present (Definition 35.1); order read by comparing decaying traces (Model 35.2); event, processing, and represented times (Definition 35.3) and the lag of a labeled present (Proposition 35.4).

What is relational here?

Temporal order, a relation between events; time since and time until, relations between events and the reader's present; in the fish, a rate of amplitude change combined with a differential phase.

What physically reads it?

Neural integration windows and comparisons of decaying traces; in the fish, midbrain neurons tuned to beat rates of a few hertz, which read the sign of Δf over a window comparable to a beat period.

What existing field explains it?

The psychophysics of temporal integration and order (Pöppel 1997; Wittmann 2011), postdiction and discrete perception (Eagleman and Sejnowski 2000; Herzog et al. 2016; VanRullen and Koch 2003), the phenomenology of time consciousness (W. James 1890; Husserl 1991), the cognitive neuroscience of mental time travel (Suddendorf and Corballis 2007; Schacter et al. 2007), and filtering and prediction in signal processing.

What remains unexplained?

The mechanisms are explained or explicable. What remains is the gap between the third grade, a represented present, and an experienced one. The literature on time perception measures processing windows; whether any window is lived through is not among the things it measures.

A discriminating experiment

Dissociate represented time from processing time. Present an event and, at controlled delays after it, a second stimulus that should change the time a postdictive system assigns to the event but not the time a predictive system assigns. In humans, report of the event's time distinguishes the two; in animals, a timed behavioral response can do the same. For the fish, clamp the beat so that its phase at the window's onset is controlled, and ask whether the reliability of the response depends on phase for short stimulus durations, as (35.4) predicts.

35.8 Notes and References

The specious present is discussed by W. James (1890, vol. 1, ch. 15), who credits the phrase to E. R. Clay. Husserl (1991) develops retention, primal impression, and protention. The hierarchy of temporal windows is proposed by Pöppel (1997) and refined by Wittmann (2011). Postdiction in the flash-lag illusion is argued by Eagleman and Sejnowski (2000); the two-stage model of unconscious feature integration followed by discrete conscious rendering is Herzog et al. (2016); VanRullen and Koch (2003) review discrete perception. Mental time travel and its evolution are discussed by Suddendorf and Corballis (2007), and the shared basis of remembering and imagining by Schacter et al.

(2007). The fish model follows Chapter 21; beat-rate tuning is from G. Rose and Heiligenberg (1986b), and the slow frequency shifts evoked by stimulating the nucleus electrosensorius from C. H. Keller et al. (1990).

35.9 Exercises

Exercise 35.1. For the leaky integrator $h(\tau) = \tau_c^{-1} \mathrm{e}^{-\tau/\tau_c}$, find the window T that carries a fraction $1 - \eta$ of the weight, and evaluate it for $\eta = 0.05$ and $\eta = 0.01$. Repeat for the kernel $h(\tau) = \tau\, \mathrm{e}^{-\tau/\tau_c}/\tau_c^2$, which has a delayed peak, and compare the latencies.

Exercise 35.2. Derive (35.3) from (35.2). How does the order threshold change if the comparison is made later, at larger $t - t_B$? Suggest a mechanism that would make the threshold independent of the comparison time.

Exercise 35.3. Model the flash-lag illusion as follows: the position attributed to a moving object at the time of a flash is its average position over the window $[t_f, t_f + W]$ after the flash. Show that the object appears displaced ahead of the flash by $vW/2$ for an object moving at constant speed v. Evaluate the displacement for $W = 80\,\mathrm{ms}$ and $v = 10\,^\circ/\mathrm{s}$, and state which features of the illusion the toy model cannot capture.

Exercise 35.4. Using (35.4), compute the average of $\dot{A}_2\,\delta$ over a window of length W that starts at beat phase ψ_0. Add independent white noise to the product and find, as a function of W and ψ_0, the probability that the sign of the windowed average is wrong. At what window length does the error probability become nearly independent of ψ_0?

Exercise 35.5. Assign each of the following to the grades of Section 35.6, with reasons: a smoke detector with a delay, a tracking filter that predicts a target's next position, an alarm clock, and a person planning a trip. Which of them hold a variable for the present as present?

Exercise 35.6. Show that shifting every event time and the reader's clock by a constant leaves every temporally centered variable unchanged. Then describe a system whose behavior depends on absolute time and identify the external relatum (a clock, the sun) that it must be reading.

36

Me

> This system constitutes a representation of 'the material me', and might provide a foundation for subjective feelings, emotion and self-awareness.
>
> A. D. Craig (2002)

A system acting in a world has to keep track of which changes in its input it caused, which parts of the world are its own body, which events are its own actions, and what state its own interior is in. Each of these is a reading of a relation with the system as one relatum: predicted input against actual input, seen touch against felt touch, intended outcome against actual outcome, current bodily state against expected bodily state.

The thesis of this chapter is that "me," at the level where neuroscience can study it, is a cluster of such self-specifying readings. Each has a known mechanism, each can be dissociated from the others, and each can be fooled by an intervention on the relation it reads. Philosophers call the cluster the *minimal self* and distinguish it from a *narrative self* that extends over time (Gallagher 2000). The minimal self, understood this way, is functional throughout. The electric fish has part of it, and nothing in it, alone or assembled, is yet a subject.

36.1 Self and Non-Self as Source Attribution

When an animal moves, its sensory input changes for two reasons: because the world changed and because the animal moved. Von Holst and Mittelstaedt named the self-produced component *reafference* and the externally produced component *exafference*, and Sperry proposed that the nervous system uses a copy of the motor command, a corollary discharge, to predict and discount the former (Holst and Mittelstaedt 1950; Sperry 1950). Chapter 22 develops the history and the electric-fish case. Here we need the general form of the problem, which is one of source attribution.

Model 36.1 (Source attribution with a forward model)**.** Let the input be

$$y = \Gamma(a) + e + n, \tag{36.1}$$

where $\Gamma(a)$ is the sensory consequence of the system's own command a, e is an external contribution, and n is noise with standard deviation σ. A forward model driven by a copy of the command predicts $\hat{\Gamma}(a)$, and the system attributes a change to an external source when the residual $r = y - \hat{\Gamma}(a)$ exceeds a criterion, $|r| > \kappa$.

The residual is a relation between two signals, the prediction and the input, and the attribution is a reading of that relation. The reading can fail in a specific way.

Proposition 36.2 (Misattribution from a biased prediction)**.** *In Model 36.1, suppose the input is entirely self-produced ($e = 0$), the noise is Gaussian, and the forward model is biased, $\hat{\Gamma}(a) = \Gamma(a) + b$. Then the probability that the self-produced input is attributed to an external source is*

$$P_{\text{mis}}(b) = 1 - \left[\Phi_N\!\left(\frac{\kappa + b}{\sigma}\right) - \Phi_N\!\left(\frac{-\kappa + b}{\sigma}\right)\right], \tag{36.2}$$

where Φ_N is the standard normal distribution function. P_{mis} is smallest at $b = 0$ and increases with $|b|$.

Proof. With $e = 0$, $r = n - b$, which is Gaussian with mean $-b$ and standard deviation σ. The probability that $|r| \le \kappa$ is $\Phi_N((\kappa + b)/\sigma) - \Phi_N((-\kappa + b)/\sigma)$, and P_{mis} is its complement. The bracket is the probability mass of a unit Gaussian in an interval of fixed length $2\kappa/\sigma$ centered at b/σ, which is largest when the interval is centered at zero and decreases as it moves away. □

The model is a sketch, and its moral is general: a system that separates self from non-self by prediction will misattribute its own products to the world in proportion to the error of its self-prediction. Two lines of evidence fit this form. A self-produced touch is felt as less ticklish than the same touch produced externally; in functional imaging, somatosensory cortex responded more to externally produced stimulation, and the pattern of cerebellar activity suggested that the cerebellum predicts the sensory consequences of movements and supplies the signal that cancels the response to self-generated stimulation (Blakemore et al. 1998). Frith et al. (2000) propose a framework in which awareness of action draws on two sources, sensory signals from skin and muscles and the stream of motor commands, and in which patients with delusions of control can no longer properly link their intentions to their actions. In the terms of the model, their own commands' consequences no longer fall within the criterion of self-produced input.

36.2 Interoception and the Material Me

A second self-specifying reading concerns the body's interior. Craig (2002) describes an afferent system in primates and humans that represents the physiological condition of the body, its temperature, pain, itch, muscular and visceral state, and proposes that this system constitutes a representation of "the material me" and a possible foundation for feelings, emotion, and self-awareness. Seth (2013) recasts interoception in predictive terms: emotions arise from actively inferred generative models of the causes of interoceptive signals, and the same machinery bears on body ownership and conscious selfhood. Damasio (1999) had earlier built a theory of the self on neural maps of the body's state.

On the reading of this book, interoception reads a relation between the body's current state and the state it is expected or regulated to be in. A deviation from a set point is a relation; so is a mismatch between a predicted and an observed heartbeat. Regulation that anticipates needs rather than only correcting errors has been called allostasis (Sterling 2012), and an allostatic controller acts on exactly such predicted relations. The variable is agent-indexed in the sense of Chapter 34, because the agent's own interior is one of its relata, and it closes a loop in the most literal way, since what is read is used to keep the reader alive. Chapter 28 returns to that loop as the origin of norms.

36.3 Proprioception and the Body Model

Sherrington (1906) introduced the term *proprioceptive* for the receptors that signal the body's own movements and positions, from muscles, tendons, and joints. Proprioception supplies the posture signals of the coordinate chains of Chapter 34: the eye-in-head, head-on-body, and hand-in-body terms that turn a retinal position into a reachable one. The positions it reports are relations among parts of the agent, which makes them the most thoroughly agent-indexed variables the nervous system reads. They are also corrigible. In the rubber hand illusion, discussed below, participants' judgments of where their own hidden hand is drift toward a rubber hand they see being stroked (Botvinick and J. Cohen 1998). A proprioceptive reading can be pulled by vision because the brain combines several readings of the same relation.

36.4 Agency

The sense of agency is the experience of controlling one's own actions and, through them, the course of events in the world (Haggard 2017). Its measurable counterpart is a classification: this outcome was produced by my action. A comparator account makes the classification a reading of the match between

the predicted and the actual consequences of an action (Frith et al. 2000). Good matches are attributed to oneself; poor matches are not.

Gallagher (2000) separates two components of the minimal self that ordinary action runs together. When I move my arm, I am both the one whose arm moves (ownership) and the one who moves it (agency). When someone else moves my arm, ownership remains and agency is absent. The two readings rest on different relations. Ownership concerns whether the moving thing belongs to my body; agency concerns whether its movement matches what I intended. Pathology dissociates them: Gallagher (2000) discusses schizophrenia as a case in which the sense of agency may be disrupted while the sense of ownership is preserved.

36.5 Ownership

Botvinick and J. Cohen (1998) placed a rubber hand in front of participants whose own hand was hidden from view, and stroked the two hands with brushes. When the strokes were synchronous, participants reported feeling the touch where they saw the brush touch the rubber hand, and their judgments of the position of their own hand drifted toward the rubber one. The illusion depends on the timing relation between the seen and the felt strokes: asynchronous stroking weakens or abolishes it (Botvinick and J. Cohen 1998; Tsakiris 2010).

The experiment is a relational intervention in the sense of Chapter 1. The visual stimulus and the tactile stimulus are held fixed, and only their temporal relation is changed. Because ownership changes with the relation and not with either stimulus, the relation is what the ownership system reads.

Tsakiris (2010) proposes a model with three stages: a stored model of the body that tests whether an object could be part of it at all; on-line representations of the body's anatomy and posture that modulate the integration of visual and tactile signals and recalibrate their coordinate systems; and the referral of touch to the seen object, which gives rise to the experience of ownership. The model assigns the stages to a network including the right temporoparietal junction, secondary somatosensory cortex, posterior parietal and ventral premotor cortices, and the right posterior insula. The first stage explains why the illusion fails for objects that could not be a hand: the relation is read only for candidates the body model admits.

Model 36.3 (Ownership as a gated reading of coincidence)**.** Let v_k and u_k be the times of seen and felt strokes, and let

$$C = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}(|v_k - u_k| < W) \tag{36.3}$$

be the fraction of felt strokes accompanied by a seen stroke within a window W. An ownership index

$$O = g \cdot \frac{1}{1 + \mathrm{e}^{-\beta(C - C_0)}} \tag{36.4}$$

combines the coincidence reading with a gate $g \in [0, 1]$ supplied by the body model, which is near zero for objects that cannot be part of the body. The functional form and the parameters W, β, C_0 are illustrative.

The model says nothing new about the illusion. It makes the structure explicit: ownership is a reading of a temporal relation between two sensory streams, gated by a model of the body, and manipulating the relation manipulates the reading.

Example 36.4 (Three manipulations). Hold the brushes, the rubber hand, and the participant fixed, and vary three things in turn. Delaying the felt stroke relative to the seen one lowers C and, in the model, ownership. Replacing the rubber hand by an object that could not be a hand sets the gate g near zero whatever the timing. Rotating the rubber hand into a posture the participant's own hand could not adopt lowers g through the posture stage of Tsakiris (2010). Only the first is an intervention on the relation; the other two are interventions on the reader. Chapter 58 makes the distinction a method.

36.6 Minimal and Narrative Self

Table 36.1 collects the readings of this chapter. Each has the system as a relatum, each has a candidate mechanism, and each can be dissociated from the others by an illness or an illusion.

Reading	Relation read	Candidate mechanism	Known failure
Self-produced input	predicted vs. actual input	forward model, corollary discharge	tickle attenuation lost; misattribution
Interior state	current vs. expected bodily state	interoceptive afferents, inference	(see Chapter 28)
Posture	position of parts relative to the body	proprioceptors	drift toward a rubber hand
Agency	intended vs. actual outcome	comparator	delusions of control
Ownership	synchrony of seen and felt touch, gated by a body model	multisensory integration	rubber hand illusion

Table 36.1: Self-specifying readings. Sources in the text: Blakemore et al. (1998), Frith et al. (2000), Craig (2002), Seth (2013), Sherrington (1906), Botvinick and J. Cohen (1998), Tsakiris (2010), and Haggard (2017).

Gallagher (2000) calls the self given in such readings the minimal self, a self without temporal extension, and contrasts it with the narrative self,

which involves personal identity and continuity across time and draws on episodic memory and on the interpretive activity he discusses in connection with Gazzaniga's left-hemisphere "interpreter." The narrative self is a model of the system extended over time, and Chapter 41 treats it with the self-model theories. The minimal self, on our reading, is not a thing at all. It is a set of variables with the system as a relatum, read by different mechanisms, which normally agree and can be made to disagree.

36.7 The Fish's Me

The electric fish reads variables with itself as a relatum at every step. Its electric image is a modulation of its own field (Chapter 34), and the frequency difference that drives the jamming avoidance response, $\Delta f = f_{\mathrm{e}} - f_{\mathrm{s}}$, is its neighbor's frequency relative to its own. Whether it also separates self from non-self by prediction depends on the fish.

Pulse-type mormyrid fishes do. A corollary discharge from the electromotor command reaches cerebellum-like electrosensory structures, where it is modified by reafferent input so that a plastic "negative image" of the predictable consequences of the fish's own discharge is subtracted from the sensory response (C. C. Bell 1981; C. Bell et al. 1997; Requarth and Sawtell 2011). That is Model 36.1 implemented in tissue.

The wave-type *Gymnarchus* does not. It has no neuronal projections from its pacemaker to its sensory areas, so no corollary discharge is available to gate its electrosensory input (Kawasaki 1994), and it determines the direction of its jamming avoidance response from the signal mixture alone (Kawasaki 1993). *Eigenmannia*, likewise, needs no internal reference to its own pacemaker (Heiligenberg 1991a). The jamming avoidance response never has to classify an input as self-produced, because the variable it reads is a relation in which the self is already one of the terms. The wave-type fish has "me" as a relatum without "me" as a predicted source. It shows that agent-indexed reading does not require self-prediction, and the mormyrid shows that self-prediction evolves where the task needs it.

36.8 What "Me" Does Not Yet Include

The readings of this chapter are functional. A robot with a forward model, proprioceptive sensors, a comparator for its own actions, a multisensory check on which of the objects in its camera image is its own arm, and a controller for its battery charge would have all five rows of Table 36.1. Engineers build such systems for ordinary reasons, and nobody supposes that each one thereby hosts someone to whom its arm feels like its own.

What the chapter has shown is how much of the minimal self can be stated as readings of relations, and how precisely each reading can be manipulated. It

has not shown why, in us, these readings are accompanied by a sense that the body is mine. The illusions of this chapter change that sense by changing the relations; they do not explain why there is a sense to change. Chapter 41 asks whether a model of the self as a whole adds what is missing, and Chapter 43 states the gap in its general form.

36.9 Audit

Canonical phenomenon
Synchronous stroking of a visible rubber hand and a hidden real hand makes the rubber hand feel like one's own; asynchronous stroking does not (Botvinick and J. Cohen 1998).

Minimal model
Source attribution by a forward model (Model 36.1) and misattribution from a biased prediction (Proposition 36.2); ownership as a gated reading of coincidence (Model 36.3).

What is relational here?
Predicted against actual input; current against expected bodily state; positions of body parts relative to the body; intended against actual outcome; the synchrony of seen and felt touch.

What physically reads it?
Forward models and corollary discharge (cerebellum; mormyrid cerebellum-like electrosensory structures); interoceptive afferent systems and their cortical targets; proprioceptors; comparators for action; the multisensory network of Tsakiris (2010).

What existing field explains it?
Motor control and the reafference principle (Holst and Mittelstaedt 1950; Sperry 1950; Blakemore et al. 1998; Frith et al. 2000); the neuroscience of interoception (Craig 2002; Seth 2013); research on agency and ownership (Haggard 2017; Botvinick and J. Cohen 1998; Tsakiris 2010); the philosophy of the minimal and narrative self (Gallagher 2000); electrosensory neurobiology (C. Bell et al. 1997; Requarth and Sawtell 2011; Kawasaki 1994).

What remains unexplained?
The mechanisms of the minimal self are explained in outline and some in detail. The residue is the sense of mineness itself: the illusions show that it tracks the relations these mechanisms read, and they do not show why tracking them is accompanied by a sense of anything.

A discriminating experiment
Separate the reading of the relation from the reading of the body model.

Vary visuotactile delay in fine steps with a plausible and an implausible rubber hand, and measure proprioceptive drift and reported ownership. Model 36.3 predicts that delay and plausibility act multiplicatively, through C and g; a purely additive effect would count against a gating account.

36.10 Notes and References

The distinction between minimal and narrative self, and between the senses of agency and ownership, is drawn by Gallagher (2000). Reafference and corollary discharge go back to Holst and Mittelstaedt (1950) and Sperry (1950); Crapse and Sommer (2008) survey corollary discharge across the animal kingdom. The tickle study is Blakemore et al. (1998), and the framework for disorders of the awareness of action is Frith et al. (2000). Interoception is reviewed by Craig (2002) and recast in predictive terms by Seth (2013); Damasio (1999) develops a body-based theory of the self, and Sterling (2012) the idea of allostasis. Sherrington (1906) introduced proprioception. The rubber hand illusion is Botvinick and J. Cohen (1998); Tsakiris (2010) gives the neurocognitive model of body ownership used here, and Haggard (2017) reviews the sense of agency. The mormyrid negative image and its plasticity are described by C. C. Bell (1981), C. Bell et al. (1997), and Requarth and Sawtell (2011); the absence of corollary discharge in *Gymnarchus* by Kawasaki (1994).

36.11 Exercises

Exercise 36.1. Plot $P_{\mathrm{mis}}(b)$ from (36.2) against b/σ for criteria $\kappa = \sigma$, 2σ, and 3σ. For which criterion is the system most sensitive to a small bias, and what does it pay in failures to detect a real external contribution?

Exercise 36.2. Extend Model 36.1 so that the forward model's prediction arrives at a fixed time after the command. If an experimenter inserts a delay between a person's movement and the touch it produces, how should the attenuation of the self-produced touch change with the delay, according to the model? State a result that would count against the forward-model account.

Exercise 36.3. Simulate Model 36.3 with seen and felt strokes at one per second, the felt strokes delayed by d plus Gaussian jitter of 50 ms, and $W = 200$ ms. Plot the ownership index against d and find the delay at which it falls to half its maximum. Which parameter of the model sets that delay?

Exercise 36.4. Classify each of the following by the presence of ownership and agency in Gallagher's sense: a voluntary reach, the same arm lifted by another person, a tendon reflex, and a movement of a cursor on a screen that follows one's hand with a variable delay. For each, name the relation that the relevant reading would compare.

Exercise 36.5. Describe the rubber hand illusion in the notation of the reading loop $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$. What is the mixture, what does the reader output, and is the loop closed? Propose a version of the experiment in which the participant's own movements can change the relation being read.

Exercise 36.6. Explain, using the model of Chapter 21, why a reader of the sign of Δf needs no prediction of the fish's own discharge, while a pulse-type fish detecting a small object near its body benefits from subtracting a negative image of its own discharge. What property of the variable being read makes the difference?

37

The Essential Indexical

> I once followed a trail of sugar on a supermarket floor, pushing my cart down the aisle on one side of a tall counter and back the aisle on the other, seeking the shopper with the torn sack to tell him he was making a mess. With each trip around the counter, the trail became thicker. But I seemed unable to catch up. Finally it dawned on me. I was the shopper I was trying to catch.
>
> John Perry, "The Problem of the Essential Indexical" (1979)

Two beliefs can agree on every fact about the world and differ in what they make their holder do. "The shopper with the torn sack is making a mess" and "I am making a mess" were true together all along in Perry's supermarket, and only the second made him stop and fix the sack. The difference is carried by an indexical, a word such as "I," "here," or "now," whose contribution depends on who uses it, where, and when.

The thesis of this chapter is that indexical content is computationally consequential. Philosophy has a precise account of it: Perry's analysis of belief states, Castañeda's quasi-indicators, Kaplan's distinction between character and content, and Lewis's centered worlds. Formalized with centered worlds, the account yields a simple result: two credal states with the same beliefs about which world is actual can recommend different actions, because actions are performed from a center. An indexical is a self-locator, and a controller needs one before any description of the world can move it. The electric fish has a self-locator without language, because the variable it reads is indexed to itself by construction. None of this is yet a subject.

37.1 The Messy Shopper

Perry's example opens his paper (Perry 1979, p. 3). What changed when it dawned on him that he was the shopper? Not his beliefs about the supermarket: he believed all along that someone with a torn sack was spilling sugar. Not, in any obvious way, his beliefs about John Perry: he might have believed that

John Perry was in the store, or even, told so by a helpful stranger, that John Perry was the one with the torn sack, and still gone on chasing, if he failed to realize that he was John Perry. What changed is a belief he could express only with "I."

Perry's second example makes the same point with time. A professor who wants to attend a department meeting on time, and who believes correctly that it begins at noon, sits motionless in his office at noon; suddenly he starts up. He believed all along that the meeting starts at noon. He has come to believe that it starts *now* (Perry 1979).

The problem Perry drew from such cases is that the belief which explains the action cannot be captured by any description that leaves the indexical out. Replace "I" by "the shopper with the torn sack," "John Perry," or "the only bearded philosopher in the store," and the explanation of the stopping fails unless we add a further belief that links the description back to "I." Perry's own response was to separate the *state* of believing, classified by the role it plays in a person's cognition and action, from the *object* of belief, the proposition believed. Two people in the same belief state (each thinking "I am making a mess") believe different propositions; one person can believe the same proposition in different states (as "John Perry is making a mess" and as "I am making a mess"), and only one of those states stops the cart.

37.2 Quasi-Indicators

Attributing such beliefs to others requires a device of its own. Suppose someone learns that the editor of a certain journal is rich, without knowing that he himself is that editor. The report "he believes that he is rich" is ambiguous. On one reading it says, correctly, that he believes of a certain person, who happens to be him, that the person is rich. On the other it says, incorrectly, that he believes he himself is rich, a belief he would express with "I am rich."

Castañeda (1966) identified the second use of "he" (often written *he**, the *quasi-indicator*) as a distinct logical device, irreducible to names or descriptions of the believer. It is how we attribute to others the first-person beliefs that Perry showed to be essential for explaining their actions. The quasi-indicator is the third-person trace of the indexical: it reports that a belief is held from the believer's own center.

37.3 Character and Content

Kaplan (1989) gave indexicals a formal semantics. A *context* of use fixes an agent, a time, a place, and a world, $c = \langle a_c, t_c, p_c, w_c \rangle$, and an indexical sentence has two levels of meaning.

Definition 37.1 (Character and content)**.** The *content* of a sentence in a context is what it says there: a proposition, which we take to be the set of worlds (at the context's time) in which it is true. The *character* of a sentence is the rule that takes each context to the content the sentence has in it:

$$\chi : c \longmapsto \text{content in } c.$$

For "I am hungry," the character is fixed by the meaning of the words, and the content varies with the speaker: in a context whose agent is Perry, the content is the proposition that Perry is hungry at t_c. Two speakers who say "I am hungry" share a character and express different contents. Perry saying "I am hungry" and someone else saying "Perry is hungry" express the same content with different characters. Kaplan suggested that character corresponds more closely than content to what the first-person case needs, the cognitive significance of a thought, and on this reading Perry's belief states are sorted by character.

The two levels come apart in a way that bears directly on this book. The sentence "I am here now" is true in every context in which it is used, since the speaker is always where and when the utterance is, yet its content in any context is contingent: the speaker might have been elsewhere. Truth in every context of use is a property of the character. It holds because contexts are centered on their agents, and it is exactly the structure of the agent-indexed variables of Chapters 34 and 35.

37.4 Centered Worlds

Lewis (1979) proposed that the objects of belief are properties, which the believer *self-ascribes*. To believe that I am hungry is to ascribe hunger to myself. The proposal has an equivalent formulation in terms of possible worlds with a designated individual and time.

Definition 37.2 (Centered world)**.** Let W be a set of possible worlds. A *centered world* is a triple (w, a, t) in which $w \in W$ is a world, a is an individual existing in w, and t is a time. Write $\mathcal{C}$ for the set of centered worlds.

Definition 37.3 (De dicto and de se content)**.** A *de dicto* content is a set of worlds $P \subseteq W$; its centered image is $\{(w, a, t) \in \mathcal{C} : w \in P\}$. A *de se* content is any set of centered worlds $S \subseteq \mathcal{C}$. A de se content is *irreducibly self-locating* if it is not the centered image of any de dicto content.

"I am hungry" is the set of centered worlds whose center is hungry at its time, and it is irreducibly self-locating whenever some world contains both hungry and sated individuals. Lewis's central example is two gods who inhabit a world they know completely: they know every proposition true in it, and so every de dicto fact. One lives on the tallest mountain and throws down

manna; the other lives on the coldest mountain and throws down thunderbolts. Neither knows which of the two he is (Lewis 1979). Their ignorance is not ignorance of any proposition in the de dicto sense, since there is none left to learn. It is ignorance of where they are in the world they know.

Proposition 37.4 (Self-locating uncertainty)**.** *Let* Cr *be a probability distribution on* $\mathcal{C}$ *and* Cr_W *its marginal on worlds. There are distributions for which* Cr_W *is concentrated on a single world while* Cr *is spread over several centers in it.*

Proof. Take a world w containing two individuals a_1, a_2 at time t, and let Cr assign probability $\frac{1}{2}$ to each of (w, a_1, t) and (w, a_2, t). Then Cr_W gives probability 1 to w. □

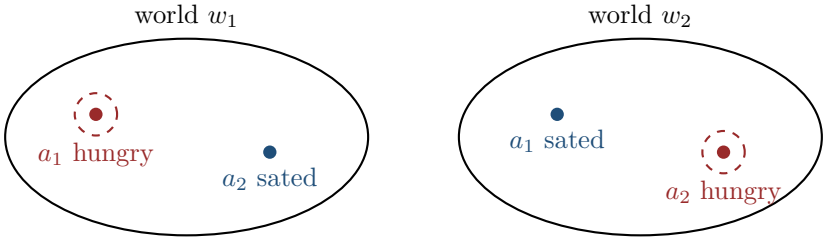


Figure 37.1: Centered worlds. A de dicto content selects whole worlds; a de se content selects centers. In each world one individual is hungry, so "someone is hungry" is true in both worlds, while "I am hungry" holds only at the centers marked. Knowing which world is actual does not settle which center one occupies.

37.5 Why the Center Matters for Action

Actions are performed by someone, somewhere, at some time. The expected value of an action therefore depends on where the actor is, and the following result makes the dependence formal.

Proposition 37.5 (Action requires the center)**.** *Let an agent choose among actions whose outcomes depend on the centered world* (w, a, t)*, because each action is executed by the center* a *at* t*, and let it choose by maximizing expected utility under a credence* Cr *on* $\mathcal{C}$*. Two credences with the same marginal* Cr_W *can recommend different actions.*

Proof. Take a single world w containing a hungry individual a_1 and a sated individual a_2 at t, and two actions, *eat* and *wait*. Let eating yield utility $+1$ at a hungry center and -1 at a sated one, and waiting yield 0. A credence certain of (w, a_1, t) recommends eating; a credence certain of (w, a_2, t) recommends waiting. Both have $\mathrm{Cr}_W(w) = 1$. □

A credence that splits evenly between the two centers gives both actions expected utility zero, and so gives no reason to act at all. That is the position of Lewis's gods, stated as a decision problem: complete knowledge of the world, and no basis for doing anything in it. It is also Perry's position in the supermarket before the trail of sugar dawned on him. He had enough de dicto information to know that someone should stop and fix a sack, and no self-locating information connecting that someone to his own hands.

The proposition is elementary, and it locates the essential indexical precisely. The indexical is essential because action is centered. A description of the world, however complete, recommends nothing until it is joined to the location of the one who will act.

37.6 Indexicals as Computational Variables

In the notation of this book, a system acts through a policy $a = \pi(z)$ on a read variable z (Chapter 15). For the policy to be implementable, z must already be expressed relative to the effectors that will execute a: the cup's position relative to the hand, the time until contact, the neighbor's frequency relative to one's own. A system that holds only world-anchored variables must first compute agent-indexed ones from them, and to do so it needs its own location in the world model.

Definition 37.6 (Self-locator)**.** A *self-locator* for a system with a world model is a map Λ from world-anchored variables to agent-indexed ones whose parameters are the system's own state. In the spatial case of Chapter 34, with the pose $(\boldsymbol{R}, \boldsymbol{t})$ as parameters,

$$\Lambda_{(\boldsymbol{R},\boldsymbol{t})}(\boldsymbol{p}_j) = \boldsymbol{R}^{\top}(\boldsymbol{p}_j - \boldsymbol{t}). \tag{37.1}$$

An indexical belief, in these terms, is a belief whose content includes the value of the self-locator's parameters, or a variable already expressed through them. "I am making a mess" supplies Perry's policy with a mess located at his own cart. "The meeting starts now" supplies the professor's policy with a time until the meeting of zero. Without the self-locator, a complete world model is inert for control, which is Proposition 37.5 in engineering terms: a mobile robot that loses its own position on a perfect map cannot plan a single step, and localization is the first thing it must recover.

Example 37.7 (The fish's indexical). The jamming avoidance response obeys a rule stated entirely in an agent-indexed variable: raise one's own discharge frequency when $\Delta f = f_{\mathrm{e}} - f_{\mathrm{s}} < 0$ and lower it when $\Delta f > 0$ (Metzner 1993b). Consider a de dicto description of the same situation: "one fish discharges at 400 Hz and another at 403 Hz." It contains every fact needed and determines no response, because it does not say which fish is reading it. To act on it, a fish would need the further self-locating fact "I am the one at 400 Hz." The

real fish never needs that fact. What reaches its receptors is the beat of its own field with its neighbor's (Chapter 20), and the sign of Δf it extracts is the neighbor's frequency *relative to its own* from the start (Heiligenberg 1991a). Active sensing supplies the self-locator for free: the variable is indexed to the reader because the reader contributes one of its relata.

The example shows that indexicality in the functional sense does not need language, belief, or a self-model. Perry's problem was posed for believers who represent the world in part through descriptions and must connect those descriptions to themselves. A system that never represents the world in world-anchored terms never faces the connection problem, and its readings are indexical without being about itself. The capacity whose breakdown the messy shopper illustrates, linking a description to oneself, belongs to systems that hold both kinds of variable, and it can fail in them in exactly the way Perry's story describes.

37.7 What the Indexical Does Not Settle

The essential indexical is a result about the logic of attitudes and the explanation of action. It shows that some contents are irreducibly centered and that action needs them. It does not show that occupying a center is like anything.

Thomas Nagel framed the larger problem as that of combining the perspective of a particular person inside the world with an objective view of the same world, the person and the person's viewpoint included (Nagel 1986). The centered-worlds formalism accomplishes part of that combination. It adds to the objective description a pointer to a center, and it shows that the pointer carries information no objective description contains. A pointer is still a functional item. A controller with a self-locator, and a fish whose readings are relative to its own field, have centered variables in exactly the sense of this chapter, and the question whether there is anything it is like to be at the center of either is left where Chapter 43 takes it up. Lewis's gods lack a piece of information; nothing in the example says that lacking or having it is an experience.

37.8 Audit

Canonical phenomenon

Perry stops chasing the messy shopper when he comes to believe "I am making a mess," a belief that adds nothing to what he already believed about the world (Perry 1979).

Minimal model

Kaplan's character and content (Definition 37.1); centered worlds and de

se content (Definitions 37.2 and 37.3); self-locating uncertainty (Proposition 37.4) and the dependence of action on the center (Proposition 37.5); the self-locator (Definition 37.6).

What is relational here?
The relation between a world-anchored state of affairs and the one who acts: the mess and my cart, the meeting and my present, the neighbor's frequency and mine.

What physically reads it?
In people, whatever mechanism links world-anchored representations to egocentric and temporally centered ones, the coordinate transformations of Chapter 34 among them. In the fish, the electrosensory pathway, whose input is the beat of the fish's own field with its neighbor's.

What existing field explains it?
The philosophy of indexicals and self-locating attitudes (Perry 1979; Castañeda 1966; Kaplan 1989; Lewis 1979); decision theory with centered credences; the neuroscience of coordinate transformation (Chapter 34); robot localization in engineering practice.

What remains unexplained?
None in the logic. The residue is the step from a centered variable to a subject at the center, which the literature on indexicals does not address and does not claim to.

A discriminating experiment
Separate de dicto from de se information in a behaving agent. Give an animal or a person complete information about a situation in world-anchored form (a map with all positions) and withhold only self-location; then supply self-location alone. Proposition 37.5 predicts that action becomes possible only at the second step. In the fish, present a jamming stimulus whose relation to the fish's own discharge is clamped (Chapter 59), so that the neighbor's frequency relative to the fish's own stays fixed while both change in absolute terms; the response should follow the relation.

37.9 Notes and References

The essential indexical is Perry (1979); the quasi-indicator is Castañeda (1966); the logic of demonstratives, with the distinction between character and content, is Kaplan (1989); belief as self-ascription and the example of the two gods are Lewis (1979). Nagel (1986) places the problem of the first-person perspective at the center of a general account of objectivity. The coordinate transformations that implement spatial self-location in the brain are discussed in Chapter 34; the fish's rule is stated by Metzner (1993b) and developed in Heiligenberg (1991a).

37.10 Exercises

Exercise 37.1. Using Definition 37.1, show that "I am here now" is true in every context of use while its content in each context is a contingent proposition. Which feature of Kaplan's contexts guarantees the first property?

Exercise 37.2. Formalize Lewis's two gods with centered worlds. Describe an observation that would resolve their ignorance, and show that it is irreducibly self-locating in the sense of Definition 37.3.

Exercise 37.3. In the proof of Proposition 37.5, let the credence assign probability q to the hungry center. Find the value of q above which eating maximizes expected utility. How does the threshold change if eating when sated costs $-c$ instead of -1?

Exercise 37.4. Represent Perry's credal state before and after the sugar trail "dawned on him" as distributions on centered worlds. Which changed, the world marginal or the distribution over centers? Could an outside observer who knew everything Perry knew about the world have predicted his stopping?

Exercise 37.5. Write a de dicto description of a group of four wave-type fish with given discharge frequencies. Specify exactly what self-locating information one of them would need to compute the direction of its jamming avoidance response from the description, and explain why the real fish does not need it.

Exercise 37.6. A robot navigates with a perfect map and a localization module. Using Definition 37.6, describe what the robot can and cannot do when localization fails, and state the analogue of Perry's shopper for a robot that localizes itself correctly but misidentifies which of two robots on the map it is.

38

Projective Perspective

> In the PCM, consciousness functions as an algorithm for the maximization of resilience, using projective perspective taking and imagination in order to escape local minima of free energy.
>
> David Rudrauf and colleagues (2017)

In its oldest and most literal sense, a perspective is a projection. A viewpoint maps a three-dimensional scene onto a surface by drawing a ray from each point of the scene through a center, and the image that results is a relation between the scene and the viewpoint. Projective geometry is the mathematics of what such maps keep and what they lose.

The thesis of this chapter is that projective structure gives a precise form to one kind of point of view, the kind that distal, ray-based senses produce, and that recent theories of consciousness have made it central. The Projective Consciousness Model proposes that the spatial field of consciousness is structured by projective geometry and governed by active inference (Rudrauf et al. 2017; Williford et al. 2018). Merker proposes that consciousness arose as a centered arena for action in mobile animals (Merker 2005; Merker 2007). Both proposals describe a functional structure that engineered systems also implement, and neither shows why that structure should be experienced. The electric fish has an agent-centered frame with no projective structure at all, which shows that a functional point of view does not require one.

38.1 The Central Projection

Place a center of projection at the origin and an image plane at distance f in front of it. A scene point $\boldsymbol{X} = (X, Y, Z)$ with $Z > 0$ is joined to the center by a ray, and the ray meets the plane at

$$x = f\,\frac{X}{Z}, \qquad y = f\,\frac{Y}{Z}. \tag{38.1}$$

This is the pinhole model of a camera, and to a first approximation of an eye.

Every point on a ray through the center maps to the same image point, so the projection discards depth along rays. What it records is direction from the center. In the language of Chapter 34, every image point is an agent-indexed variable: the direction of a scene point relative to the viewpoint.

Homogeneous coordinates make the structure linear. Represent an image point by any nonzero multiple of the vector $(x, y, 1)$, and a scene point by any nonzero multiple of $(X, Y, Z, 1)$. A general camera with center $\boldsymbol{c}$, orientation $\boldsymbol{R}$, and internal calibration $\boldsymbol{K}$ acts as

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \sim \boldsymbol{P} \begin{pmatrix} \boldsymbol{X} \\ 1 \end{pmatrix}, \qquad \boldsymbol{P} = \boldsymbol{K}\,[\,\boldsymbol{R} \mid -\boldsymbol{R}\boldsymbol{c}\,], \tag{38.2}$$

where $\sim$ means equality up to a nonzero scale factor. The viewpoint enters $\boldsymbol{P}$ as a parameter. Change $\boldsymbol{c}$ or $\boldsymbol{R}$ and the same scene produces a different image; move to a different scene and a suitably chosen viewpoint can produce the same image.

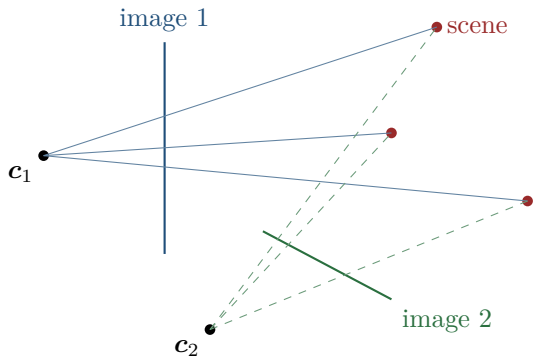


Figure 38.1: One scene, two viewpoints. Rays from each center through the scene points meet each image plane at different places. Each image is a relation between the scene and a viewpoint; the viewpoint is one of its relata.

38.2 What Projection Keeps

A central projection does not preserve distances, angles, ratios of lengths, or parallelism. Parallel lines in the scene with direction $\boldsymbol{d}$ map to image lines that meet at the *vanishing point*, the image $\boldsymbol{K}\boldsymbol{R}\boldsymbol{d}$ of the point at infinity in that direction, which is why railway lines converge toward the horizon. Collinearity and incidence survive: straight lines map to straight lines, and a point on a line maps to a point on the image of the line. One numerical quantity survives as well (Coxeter 2003).

Theorem 38.1 (Invariance of the cross-ratio)**.** *Let four distinct collinear points have coordinates a, b, c, d along their line, and define their* cross-ratio

$$(a, b; c, d) = \frac{(c-a)(d-b)}{(c-b)(d-a)}. \tag{38.3}$$

Any central projection of the line onto another line preserves the cross-ratio of the four points.

Proof sketch. In coordinates along the two lines, a central projection acts as a fractional linear map $u \mapsto (\alpha u + \beta)/(\gamma u + \delta)$ with $\alpha\delta - \beta\gamma \neq 0$. For such a map, $u_i - u_j$ becomes $(\alpha\delta - \beta\gamma)(u_i - u_j)/((\gamma u_i + \delta)(\gamma u_j + \delta))$. In (38.3) each point's denominator factor appears once in the numerator and once in the denominator, and the factors $\alpha\delta - \beta\gamma$ cancel as well, so the cross-ratio is unchanged. □

The cross-ratio is a relation among four image points that does not depend on the viewpoint. It is the projective counterpart of the frame-independent observables of Chapter 1: every individual image coordinate depends on where the viewer stands, and the cross-ratio does not. A projection therefore carries two kinds of relational information at once: viewpoint-indexed variables (the direction of each point from the center) and viewpoint-invariant ones (incidences and cross-ratios among points). A system that must act needs the first; a system that must recognize an object from any angle can use the second.

Example 38.2 (Rotation and translation of the viewpoint). If the camera rotates about its own center, the new image is related to the old one by an invertible projective map of the image plane, whatever the depths of the scene points. Pure rotation therefore reveals nothing about depth. If the center translates, points at different depths shift by different amounts in the image, and the pattern of shifts, motion parallax, reveals relative depth. An animal that moves its head is using its own viewpoint as a probe, in the sense of the active sensing of Chapter 34: the variable it reads is a relation between the scene and a viewpoint it controls.

38.3 The Projective Consciousness Model

Rudrauf et al. (2017) introduced a mathematical model of embodied consciousness, the Projective Consciousness Model (PCM), built on the hypothesis that the spatial field of consciousness is structured by a projective geometry and controlled by a process of active inference. In the model, the field combines multisensory evidence with prior beliefs held in memory and frames them by selecting points of view and perspectives according to preferences. The choice of projective frame governs how expectations are transformed; violations of expectation are encoded as free energy, and the minimization of free energy

drives perspective taking and controls the switch between perception, imagination, and action. The authors apply the model to the spatial character of subjective experience, to the relations between perception, imagination, and action, and to phenomena such as bistable figures and body-swap illusions in virtual reality (Rudrauf et al. 2017).

Williford et al. (2018) summarize the model and argue that it accounts for several aspects of subjective character that are often ignored or conflated: pre-reflective self-consciousness, the first-person point of view, the sense of mineness, and social self-consciousness. They present the model, which they describe as still in development, as the most complete theory to date of what Metzinger called phenomenal selfhood (see Chapter 41).

Model 38.3 (Projective frame selection, schematic)**.** Let an agent hold a world model $\mathcal{M}$ of scene points $\boldsymbol{X}_j$ and a family of candidate viewpoints $\boldsymbol{c}$ with orientations $\boldsymbol{R}$. Each viewpoint defines a projection $\boldsymbol{P}_{\boldsymbol{c},\boldsymbol{R}}$ and hence a predicted image of the scene. The agent evaluates each candidate by a cost $G(\boldsymbol{c}, \boldsymbol{R})$ that scores the predicted consequences of adopting that frame against its preferences, and adopts, imagines from, or acts toward the frame with the lowest cost. In the PCM the cost is an expected free energy (Rudrauf et al. 2017; Williford et al. 2018); the schematic form here, with a generic cost, is ours.

The schematic makes the functional content of the proposal visible. It adds to the agent-centered variables of Chapter 34 two things: a representation in which the whole scene is organized around a single center, with the projective structure that a center imposes, and a choice among centers, including centers the agent does not occupy. Imagining how a room looks from the door, or how a situation looks to another person, becomes a change of frame inside the model. The general framework of active inference that supplies the cost is treated in Chapter 33 (K. Friston 2010; Parr et al. 2022).

38.4 Moving the Center

The center of a visual frame can be moved by an experimenter. Petkova and Ehrsson (2008) gave participants head-mounted displays fed by cameras placed at the eyes of another body, an artificial body or another person, so that the participants saw that body from its own point of view, and they touched the participant's body and the seen body in synchrony. The combination of the relocated visual perspective with correlated multisensory information was sufficient to produce the illusion that the other body was one's own. The effect was strong enough that participants could experience being in another person's body while facing their own body and shaking hands with it (Petkova and Ehrsson 2008).

The experiment separates two things that ordinary experience keeps together, the origin of the visual projection and the location of the owned body, and then moves them together by manipulating relations. The camera sets the center of projection; the synchrony of seen and felt touch, the relation read by the ownership system of Chapter 36, attaches the body to that center. In the terms of Model 38.3, the experimenter has chosen the frame, and the participant's own processes have accepted it. The PCM lists body-swap illusions among the phenomena it accounts for (Rudrauf et al. 2017), and the experiment shows why the claim has content: the center of a point of view is a variable that physical interventions can set.

38.5 A Centered Arena for Action

Merker approaches the same territory from the logistics of behavior. He observes that consciousness presents a stable arena for action, the world, while excluding from awareness the many sensory and sensorimotor transformations by which that stable world is extracted from input corrupted by the animal's own motion (Merker 2005). He proposes that consciousness arose as a solution to problems of decision making in mobile animals with centralized brains, and so has ancient roots. In a later target article, Merker (2007) argues that an upper brainstem system, extending from the roof of the midbrain to the basal diencephalon, integrates target selection, action selection, and motivation, and serves as a penultimate stage of action control. He uses the purposive behavior of mammals after experimental removal of the cortex, and the behavior of children born without a cortex, to argue that this system can support consciousness without cortical input; the argument was published with extensive open peer commentary.

For this chapter the relevant point is structural. Merker's arena is centered on the animal and integrates world, body, and motivation into a single frame for choosing what to do. It need not be projective. A body-centered map in polar coordinates, with distances and directions measured from the animal, would serve the logistics he describes. The PCM makes a stronger claim, that the organization is specifically projective, and the two can be tested against each other.

38.6 A Point of View Without Projection

The electric fish supplies the contrast. Its sense is agent-centered at every step: the electric image of a nearby object is a modulation of the fish's own field on its own skin (Chapter 34), and the variable of the jamming avoidance response is its neighbor's frequency relative to its own. None of it is a central projection. Electrolocation works at short range, in a field the fish itself generates, and the image an object casts on the skin is shaped by the fish's own body and

field as much as by the object (Emde 1999; Caputi and Budelli 2006). There are no rays, no center of projection, and no vanishing points.

The fish therefore shows that a functional point of view, in the sense of the minimal functional perspective defined in Chapter 34, does not require projective geometry. Projective structure is what a point of view looks like when the sense is distal and carried by rays: vision above all. Whether the human field of consciousness is projective is an empirical question about human experience, and the PCM gives it a precise form. It is not a condition on having a perspective at all.

38.7 Is Geometry Enough?

A video game engine holds a model of a three-dimensional world and renders it through a virtual camera with exactly the matrix of (38.2). It can switch cameras, render the scene from a character's eyes or from above, and choose among viewpoints to keep a target in frame. A robot that plans by rendering predicted views from candidate poses and scoring them against a goal implements the schematic of Model 38.3 with an ordinary cost function. Nobody supposes that either system is conscious.

The PCM is not refuted by this, because it claims more than projective rendering. It claims that consciousness *is* a projective field of a particular kind, under a particular kind of inference, in a particular kind of embodied agent. The engine and the robot show only that the ingredients the model names can be assembled without anyone supposing experience results, so the model's power to explain experience must come from something beyond those ingredients. What that something is, the model states as a hypothesis: that the spatial field of consciousness has this structure (Rudrauf et al. 2017). That hypothesis concerns the organization of experience in beings already granted to have it. It is a strong and testable claim about the form of consciousness, and it does not address why a projective frame used for control is accompanied by experience in the first place. Chapter 44 places the PCM among the other theories by asking each the same question.

38.8 Audit

Canonical phenomenon

Parallel rails converge toward a point on the horizon in a photograph and in the visual field; the image of a scene depends on where it is viewed from, and some relations among image points do not.

Minimal model

The central projection (38.1) and the camera matrix (38.2); invariance of the cross-ratio (Theorem 38.1); projective frame selection (Model 38.3).

What is relational here?

Each image coordinate is a relation between a scene point and the viewpoint; the cross-ratio is a relation among image points that does not depend on the viewpoint; motion parallax is a relation between the scene and a moving viewpoint.

What physically reads it?

Optics performs the projection; the retina and the visual system read the projected image, including perspective cues. In the PCM, a hypothesized internal projective frame read by active inference.

What existing field explains it?

Projective geometry (Coxeter 2003); vision science and the geometry of cameras; the Projective Consciousness Model (Rudrauf et al. 2017; Williford et al. 2018); Merker's account of a centered arena for action (Merker 2005; Merker 2007); active inference (K. Friston 2010; Parr et al. 2022); the experimental relocation of the visual point of view (Petkova and Ehrsson 2008).

What remains unexplained?

Whether any projective frame is experienced. The PCM asserts that the field of consciousness is projective; the chapter found no argument that projective structure, frame selection, or free-energy minimization entails experience, and engineered systems possess all three.

A discriminating experiment

A projective internal model and a metric or body-centered one predict different errors when people judge imagined views. Under a projective model, errors in judging an imagined view from a new vantage point should preserve incidences and cross-ratios while distorting distances and angles; under a metric model, the reverse. The same comparison can be made between the PCM and Merker's arena for tasks that require perspective taking.

38.9 Notes and References

Coxeter (2003) is a classical introduction to projective geometry, including the cross-ratio and its invariance. The Projective Consciousness Model is introduced by Rudrauf et al. (2017) and related to phenomenal selfhood by Williford et al. (2018); its active-inference component draws on K. Friston (2010) and is developed at length in Parr et al. (2022). Merker's proposals are Merker (2005) and Merker (2007). The body-swap illusion is reported by Petkova and Ehrsson (2008). Nagel (1986) discusses the philosophical problem of the particular point of view in relation to objectivity. Electric images and

their dependence on the fish's own body and field are analyzed by Emde (1999) and Caputi and Budelli (2006).

38.10 Exercises

Exercise 38.1. Derive (38.1) by similar triangles. Show that the image of a straight line in the scene not passing through the center is a straight line, and find the image of a line through the center.

Exercise 38.2. Complete the proof of Theorem 38.1 by computing the cross-ratio of the images of a, b, c, d under $u \mapsto (\alpha u + \beta)/(\gamma u + \delta)$. Then show by example that the ratio $(c - a)/(c - b)$ of two lengths is not preserved.

Exercise 38.3. For the camera (38.2), show that all scene lines with direction $\boldsymbol{d}$ have images passing through the point $\boldsymbol{KRd}$, and that the vanishing points of all horizontal directions lie on one image line, the horizon.

Exercise 38.4. Show that a pure rotation of the camera about its center induces an invertible linear map of homogeneous image coordinates that does not depend on the scene. Then show that a translation of the center by $\boldsymbol{\tau}$ shifts the image of a point at depth Z by an amount proportional to $1/Z$, and explain why a moving animal can read relative depth from the pattern of shifts.

Exercise 38.5. Explain why an electric image on a fish's skin has no vanishing points and no well-defined center of projection. Which properties of the physics of electrolocation (Chapter 19) are responsible, and which agent-indexed variables does the fish have instead?

Exercise 38.6. List the components of a video game engine that correspond to the elements of Model 38.3. For each additional claim that the PCM makes about conscious agents, state whether a game engine or a planning robot could be given the corresponding feature, and at what cost.

Exercise 38.7. Specify a body-centered map in polar coordinates suitable for Merker's arena, and a projective frame of the PCM type, for an animal choosing between two prey items. Find a judgment the animal must make for which the two representations predict different errors.

Here, now, and I are coordinates with an origin, and the origin is the system. Active sensing builds such coordinates before any self-model exists: a variable of the form "world relative to my action" is agent-centered by construction. Philosophy has studied the logic of these coordinates for fifty years, from Perry's essential indexical to Lewis's centered worlds. Part IX asks what happens when a system reads its own readings.

Part IX

Self-Reading Systems

What changes when the reader becomes one of the relata?

39

The Reader Reads the Reader

> Many complex systems maintain a self-referential check and balance.
>
> Stephen M. Fleming, Raymond J. Dolan, and Christopher D. Frith (2012)

Every reader so far has read something outside itself: a beat, a field, a posture, a time. A second-order reader takes the state of another reader as its input. It can estimate how reliable a reading is, whether a decision is likely to be correct, or whether a representation is present at all. In the notation of the book, a first-order reader forms an estimate $\hat{R}$ of a world variable R; a second-order reader forms an estimate of some property of $\hat{R}$.

The thesis of this chapter is that the simplest second-order reading, confidence, has a clean normative definition, a well-studied neural basis, and a presence in rats, monkeys, and people, and that it is also cheap. Engineered systems routinely read the reliability of their own estimates. Metacognition is therefore a genuine capacity with a genuine mechanism, and a system that has it has acquired a new kind of agent-indexed variable, one about its own readings. It has not thereby acquired a point of view in the sense that Chapter 43 will require.

39.1 From Reading the World to Reading a Reading

Definition 39.1 (Second-order reader)**.** Let D_1 be a reader whose output $z_1 = \mathsf{D}_1(y)$ serves as an estimate $\hat{R}$ of a world variable R. A *second-order reader* is a reader D_2 whose input includes the state of D_1, and whose output z_2 covaries with a property of that state: its reliability, the probability that a decision based on it is correct, its source, or its presence.

The definition moves the reader from the outside of a relation to the inside. The first-order reader stands in a relation to the world; the second-order reader

takes that relation, or the state that realizes it, as one of its relata. Its output is agent-indexed in a new sense. In Chapters 34 to 36 the agent's pose, present, or body was a relatum; here a relatum is the agent's own reading.

Flavell (1979) introduced the term *metacognition* for knowledge and cognition about cognitive phenomena, in the context of children's developing ability to monitor their own memory and comprehension. T. O. Nelson and Narens (1990) gave the idea its standard architecture, in their account of metamemory: an *object level* that does the cognitive work and a *meta level* that holds a model of it, linked by two flows. In *monitoring*, information flows from the object level to the meta level; in *control*, the meta level changes what the object level does, by stopping, continuing, or redirecting it. Monitoring is second-order reading. Control closes a second loop around the first.

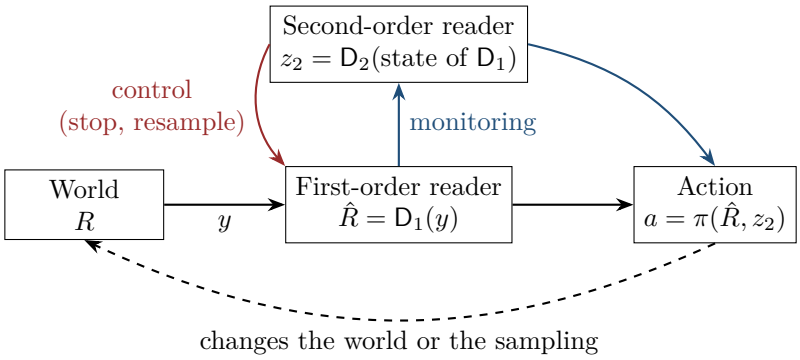


Figure 39.1: Nested reading loops. The second-order reader monitors the state of the first and, through control, changes how the first reads or whether its output is acted on.

39.2 A Signal-Detection Model of Confidence

The simplest case has a precise normative answer.

Model 39.2 (Confidence in a two-choice decision)**.** A stimulus $s \in \{-1, +1\}$ occurs with equal probability. The first-order reader receives evidence $x \mid s \sim \mathcal{N}(\mu s, \sigma^2)$ and decides $d = \operatorname{sgn}(x)$. The second-order reader outputs the probability that the decision is correct given the evidence,

$$c(x) = \mathbb{P}(s = d \mid x) = \frac{1}{1 + \mathrm{e}^{-2\mu|x|/\sigma^2}}. \tag{39.1}$$

The formula follows from Bayes's rule: the posterior odds of $s = +1$ against $s = -1$ given x are $\mathrm{e}^{2\mu x/\sigma^2}$, and the probability of the chosen alternative is the larger of the two posteriors. Confidence depends on the evidence only through its distance from the decision criterion, $|x|$. The first-order reader uses the sign of x; the second-order reader uses its magnitude. One physical quantity is read in two ways, and the two readings serve two purposes.

Proposition 39.3 (Calibration)**.** *If the observer's model is correct, confidence computed by* (39.1) *is calibrated: among decisions made with confidence c, the fraction that are correct is c.*

Proof. Confidence is a function of x, and $c(x) = \mathbb{P}(\text{correct} \mid x)$. Averaging over the values of x that yield a given confidence c gives $\mathbb{P}(\text{correct} \mid c) = c$. □

Calibration fails when the second-order reader's model of the first is wrong: if it assumes a signal strength $\hat{\mu} \neq \mu$, its confidence is systematically too high or too low. A second-order reader can misrepresent the first-order one, and the misrepresentation has the functional consequences that Chapter 30 requires of a genuine error.

A further property does not depend on the observer knowing μ, and so applies when difficulty varies from trial to trial.

Proposition 39.4 (Divergent signatures)**.** *Let confidence be any fixed increasing function of $|x|$, and let the signal strength $\mu > 0$ vary across trials without the observer's knowledge. Then, as μ increases, mean confidence on correct trials increases and mean confidence on error trials decreases.*

Proof sketch. Take $s = +1$. On correct trials x has the distribution $\mathcal{N}(\mu, \sigma^2)$ truncated to $x > 0$, a family that increases stochastically with μ; on error trials $|x| = -x$ has the distribution $\mathcal{N}(-\mu, \sigma^2)$ truncated to positive values, a family that decreases stochastically with μ. Any increasing function of $|x|$ inherits these orderings in its mean. □

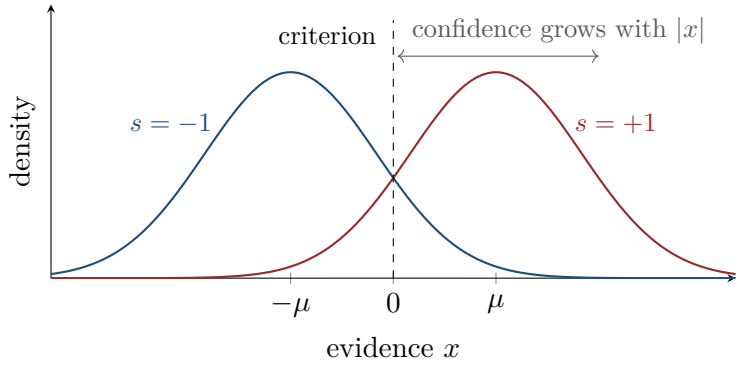


Figure 39.2: Signal-detection model of confidence (Model 39.2). The decision reads the sign of the evidence; confidence reads its distance from the criterion.

39.3 Confidence in Animals

Three lines of work show second-order reading in animals, each with a behavioral readout that does not rely on language.

Kepecs et al. (2008) trained rats on an odor categorization task in which the distance of a mixture from the category boundary controlled the difficulty of each decision. They defined a simple measure of the quality of the evidence behind each decision, found that the firing rates of many single neurons in the orbitofrontal cortex matched the predictions of confidence models and could not readily be explained by alternatives such as learned stimulus–outcome associations, and showed, in a version of the task with delayed reward, that the rats' willingness to wait for a reward increased with confidence, as the model predicted. The authors conclude that confidence estimates, once thought to require metacognition and conscious awareness, are available in the rodent brain and can be computed by relatively simple operations (Kepecs et al. 2008).

Kiani and Shadlen (2009) trained rhesus monkeys to decide the direction of moving random dots at a range of difficulties. On some trials, after viewing the stimulus, the monkeys could opt out of the direction decision in favor of a small but certain reward. They chose the sure option in a way that revealed their degree of certainty, and neurons in parietal cortex that represented the formation of the direction decision also represented the certainty underlying the decision to opt out. The same neurons carried the first-order reading and the information the second-order reading needed.

Smith et al. (2003) review uncertainty monitoring in humans, monkeys, and a dolphin, in perceptual and memory tasks that offer an "uncertain" response alongside the primary ones. The animals' performance closely resembled the humans', and the authors argue that low-level, stimulus-based accounts cannot explain it and that it requires a decisional process that sets and uses criteria. They also mark the limit of what the data show: the animals have functional features of, or parallels to, human conscious cognition, and whether they share its phenomenal features is left open.

Example 39.5 (The opt-out rule)*.* In Model 39.2, let a correct decision earn reward 1, an error 0, and the sure option $r < 1$. Choosing the decision is worth $c(x)$ in expectation, so the optimal policy opts out exactly when $c(x) < r$, that is, when $|x|$ falls below a threshold set by r, μ, and σ. Opting out is a control signal from the meta level, triggered by a second-order reading, and it closes the second loop of Figure 39.1.

39.4 The Value of Reading One's Own Reading

The opt-out rule shows why a second-order reader earns its keep.

Proposition 39.6 (Value of monitoring)**.** *In Example 39.5, a system that always decides earns* $\mathbb{E}[c(x)]$*, a system that always takes the sure option earns* r*, and a system that uses its confidence earns* $\mathbb{E}[\max(c(x), r)]$*. The third is at least as large as either of the others, and strictly larger whenever* $c(x) < r$ *and* $c(x) > r$ *both occur with positive probability.*

Proof. Pointwise, $\max(c(x), r) \geq c(x)$ and $\max(c(x), r) \geq r$; take expectations. Each inequality is strict on a set of positive probability under the stated condition. □

The gain comes entirely from the second-order variable. The first-order reader already contains all the information about the stimulus; what the second reader adds is information about the first reader's reliability on this occasion, and the gain measures what that information is worth for control. The proposition also says when monitoring is worthless: if every decision is easy, or every decision hopeless, confidence never changes the choice.

How well a system's confidence separates its own correct decisions from its errors can be measured, and it varies between individuals and tasks independently of how well the first-order decisions themselves are made (Fleming et al. 2012). The two readings can therefore come apart, and a system can be good at deciding and poor at knowing when it has decided well, or the reverse. The most dramatic dissociation of this kind, above-chance performance accompanied by denial of any awareness of the stimulus, is blindsight, which Chapter 40 uses as a test case for higher-order theories.

39.5 Recursion Is Cheap

Second-order reading is common in engineered systems. A single parity bit appended to a block of data reads whether the block has suffered an odd number of bit errors; it is a reader of the reliability of another reading, and it has a characteristic blind spot, since an even number of errors passes unnoticed. A linear estimator that propagates, along with its estimate of a moving target, the variance of that estimate's error, maintains a calibrated second-order variable at every step and uses it to decide how much weight to give new measurements. A search engine that returns an answer with a confidence score, and suppresses answers below a threshold, implements the opt-out rule of Example 39.5.

None of these systems is a candidate for consciousness on any theory in Chapter 44. The existence of second-order reading in them shows that metacognitive structure, however valuable, is not by itself evidence of experience. Fleming et al. (2012) make a related point in more careful terms: they propose that the level of a representation, the order of the behavior it supports, and access consciousness are orthogonal dimensions, so that first-order and higher-order processes can each occur with or without awareness. Higher-order theorists agree on the corresponding point about their own view: most proponents of higher-order theories do not identify consciousness with metacognition or with confidence (R. Brown et al. 2019). Chapter 40 takes up what they identify it with instead.

39.6 Does the Fish Read Its Own Reading?

Nothing we know of in the jamming avoidance response reads the reliability of the fish's reading of Δf. The response is computed by a distributed population of imprecisely tuned neurons (Chapter 21), and non-jamming combinations of amplitude and phase modulation evoke frequency shifts whose tuning, sensitivity, and dynamics form a smooth continuum with the response to genuine jamming (Carlson and Kawasaki 2007). The pattern fits a system that acts on its first-order reading without a gate that asks how good the reading is. That is an inference from the absence of evidence, and the question is open.

It is also testable. A fish with a confidence reader should respond differently to modulations that are equally strong but differently reliable: for instance, to amplitude and differential-phase modulations whose sign relation (Chapter 21) is consistent across the body surface, and to ones in which different patches of skin report conflicting signs. If the response to the conflicting stimulus is weaker, slower, or more variable than to a consistent stimulus of the same modulation depth, something in the fish is weighing the reliability of its own reading. If the response depends only on the population's summed vote, the fish reads the world and not its reading of it.

39.7 What the Second Reader Adds

A second-order reader adds a variable about the system's own readings, and with it a new kind of self-relation. The variables of Part VIII were centered on the agent's pose, its present, and its body. Confidence is centered on the agent's own epistemic state: it answers "how likely am I to be right?" and it changes behavior in ways that a first-order system cannot match, by waiting, gathering more evidence, or declining to act.

What it does not add is anything that the functional description leaves out. Confidence is a number computed from the state of a reader, calibrated or not, used for control or not. Rats have it, parity bits have a crude version of it, and the question of whether reading one's own reading is accompanied by awareness of that reading is a further question.

The comparative ladder of Part XI makes the point sharper. A radio that hops adaptively among frequencies marks the channels it cannot use reliably, because of interference, and removes them from its hopping set (*Specification of the Bluetooth System, Version 1.2* 2003): it reads the reliability of its own channel readings and acts on the result. By that measure the radio of Chapter 48 has a second-order reader and the fish of Part V, as far as anyone knows, does not. Whatever orders these systems by how plausibly they are conscious, it is not the presence of second-order reading. Higher-order theories of consciousness answer yes, for a particular kind of second-order state; Chapter 40 examines the answer.

39.8 Audit

Canonical phenomenon

Monkeys offered a small certain reward opt out of difficult perceptual decisions more often than easy ones, and parietal neurons that represent the decision also represent the certainty behind opting out (Kiani and Shadlen 2009).

Minimal model

The signal-detection model of confidence (Model 39.2), with calibration (Proposition 39.3), the divergent signatures of correct and error trials (Proposition 39.4), and the opt-out rule (Example 39.5).

What is relational here?

The relation between the first-order reader's state and the truth of what it reads: the distance of the evidence from the criterion, as a stand-in for the probability of being right.

What physically reads it?

Neurons in rat orbitofrontal cortex and monkey parietal cortex whose activity tracks confidence (Kepecs et al. 2008; Kiani and Shadlen 2009); in engineered systems, error-detecting codes and estimators that propagate their own error variance.

What existing field explains it?

Metacognition research in developmental and cognitive psychology (Flavell 1979; T. O. Nelson and Narens 1990), signal detection theory, comparative studies of uncertainty monitoring (Smith et al. 2003), the neuroscience of decision confidence (Kepecs et al. 2008; Kiani and Shadlen 2009), and the conceptual framework of Fleming et al. (2012).

What remains unexplained?

None in the mechanism. The residue belongs to the next chapter: whether some second-order state makes its object conscious, which the confidence literature neither claims nor tests.

A discriminating experiment

Measure whether a system's behavior depends on the reliability of its own reading when the strength of the input is held fixed. In the electric fish, compare responses to modulations with consistent and with conflicting sign relations across the body surface at matched modulation depth (Section 39.6). In rats and monkeys, dissociate confidence from stimulus strength by manipulating noise at fixed mean signal, and ask whether neural and behavioral readouts follow confidence or strength.

39.9 Notes and References

The term metacognition was introduced by Flavell (1979); the object-level and meta-level architecture with its monitoring and control flows is from Nelson and Narens's account of metamemory (T. O. Nelson and Narens 1990). Fleming et al. (2012) introduce a theme issue that surveys the computation, neurobiology, and function of metacognition and relates it to consciousness. The neural and behavioral evidence for confidence in rats is Kepecs et al. (2008) and in monkeys Kiani and Shadlen (2009); Smith et al. (2003) review uncertainty monitoring across species. R. Brown et al. (2019) state that most higher-order theorists do not equate consciousness with metacognition or confidence. The electric-fish material follows Chapters 21 and 23, with the phantom-jamming data of Carlson and Kawasaki (2007).

39.10 Exercises

Exercise 39.1. Derive (39.1) from Bayes's rule. Then suppose the second-order reader assumes signal strength $\hat{\mu}$ while the true strength is μ. Compute the fraction of correct decisions among those made with nominal confidence 0.8, and state when the reader is overconfident.

Exercise 39.2. Simulate Model 39.2 with $\sigma = 1$ and μ drawn from $\{0.25, 0.5, 1, 2\}$ on each trial, with confidence equal to $|x|$. Plot mean confidence against μ separately for correct and error trials, and compare the result with Proposition 39.4.

Exercise 39.3. In Example 39.5, find the threshold on $|x|$ below which the optimal policy opts out, as a function of r, μ, and σ. Plot the probability of opting out against μ for $r = 0.6$ and $r = 0.8$.

Exercise 39.4. Describe a parity bit in the language of Definition 39.1: what is the first-order reading, what property of it does the parity bit read, and what is its blind spot? Propose a second-order reader with a smaller blind spot and state what it costs.

Exercise 39.5. Design the fish experiment of Section 39.6 in detail: specify the stimulus geometry that makes the sign relation consistent or conflicting across the body surface, the matched modulation depth, and the measured response variables. State the outcome that would count as evidence for a second-order reader and the outcome that would count against one.

Exercise 39.6. Place each of the following in the three dimensions of Fleming et al. (2012): a thermostat with a sensor-fault detector, a rat waiting for a reward in proportion to its confidence, and a person reporting that they are unsure whether they saw a faint light. Which of them, if any, does the confidence literature show to be aware of its own readings?

40

Higher-Order Representation

> Higher-order theories of consciousness argue that conscious awareness crucially depends on higher-order mental representations that represent oneself as being in particular mental states.
>
> Hakwan Lau and David Rosenthal (2011)

Chapter 39 showed that systems read their own readings and that the simplest such reading, confidence, is common, useful, and cheap. Higher-order theories of consciousness make a stronger claim about a particular kind of second-order state. A mental state is conscious, they hold, when the subject is suitably aware of being in it, and that awareness consists in a higher-order representation of the state.

The thesis of this chapter is that higher-order theories supply a precise and testable criterion for which states are conscious, with real empirical support and real difficulties, and that the criterion is functional. It says which second-order organization accompanies consciousness. It leaves the question of why that organization is accompanied by experience where it found it, and a system built to satisfy the criterion would inherit the question rather than answer it.

40.1 The Transitivity Intuition

Suppose you are in a mental state, a perception of red, say, but in no way aware of being in it. Higher-order theorists find it natural to say that the perception is then not a conscious state. D. M. Rosenthal (2005) makes this the starting point of his theory: a conscious state is one whose subject is aware of being in it, and the question is what that awareness consists in. His answer is that it consists in a suitable thought, a *higher-order thought* (HOT), to the effect that one is in the state.

The appeal of the idea for this book is that it names a relation. A state is conscious, on the higher-order view, in virtue of standing in a relation to another state that represents it. Consciousness becomes a relational property

of mental states, of the kind that Chapter 1 taught us to look for, and the relatum that makes the difference is itself a reading.

40.2 Varieties of Higher-Order Theory

The theories differ in what the higher-order state is and in whether it must be present or only available.

Higher-order perception (HOP), or inner sense, models the awareness on perception. D. M. Armstrong (1968) proposed that consciousness is a kind of perception of our own mental states, and Lycan (1996) developed the view with internal monitoring mechanisms that scan first-order states and represent them in a quasi-perceptual way.

Actualist higher-order thought requires an occurrent thought. On Rosenthal's theory (D. M. Rosenthal 2005), a state is conscious when it is accompanied by a thought that one is in that state, a thought that is assertoric and that does not arise from any inference or observation of which the subject is aware. The higher-order thought is usually not itself conscious; it becomes conscious only if a further thought is about it, which is what happens in introspection.

Dispositional higher-order thought requires only availability. On Carruthers's theory (Carruthers 2000), a state is phenomenally conscious when it is available to higher-order thought, held in a system whose contents can become the objects of such thoughts, whether or not any is actually formed.

Definition 40.1 (Higher-order conditions)**.** Let s_1 be a first-order state of a subject with content concerning a world variable R.

1. *HOP*: s_1 is conscious if and only if a monitoring state represents s_1 in a quasi-perceptual way.
2. *Actualist HOT*: s_1 is conscious if and only if there is an occurrent state s_2 whose content is that the subject is in s_1, and s_2 is not based on inference or observation of which the subject is aware.
3. *Dispositional HOT*: s_1 is conscious if and only if s_1 is available to states of the kind s_2 describes.

The figure marks a distinction that critics and proponents both insist on. Confidence and higher-order representation are both second-order, and their contents differ. Confidence concerns the accuracy of a state; a higher-order thought concerns its existence and character. R. Brown et al. (2019) state that most higher-order theorists do not equate consciousness with metacognition or confidence, and the evidence of Chapter 39 that rats compute confidence therefore bears only indirectly on whether rats have higher-order states.

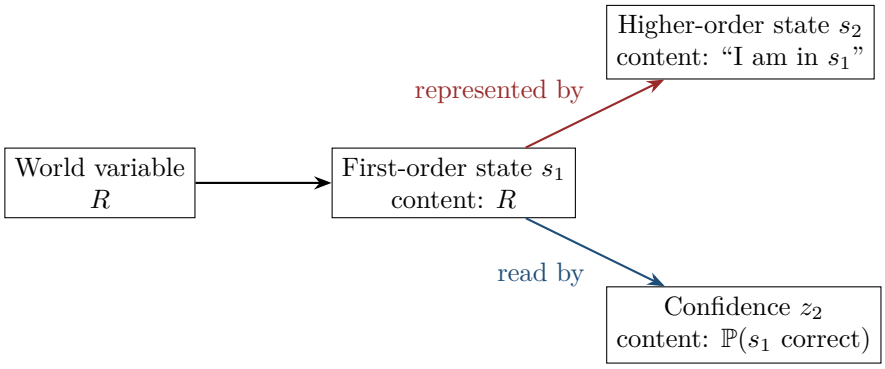


Figure 40.1: Two second-order readings of one first-order state. A higher-order state represents that the subject is in s_1; a confidence reading (Chapter 39) represents how likely s_1 is to be correct. Higher-order theories make consciousness depend on the first kind.

40.3 Unconscious First-Order States

Higher-order theories require that first-order states can occur without being conscious, and the strongest evidence that they can comes from neurology. Weiskrantz (1986) studied a patient with damage to primary visual cortex who reported seeing nothing in part of his visual field, yet, when required to guess about stimuli presented there, located and discriminated them well above chance. Weiskrantz called the phenomenon *blindsight*.

The higher-order explanation is direct. The patient's first-order visual states carry enough information about the stimuli to support discrimination, so they exist; he is not aware of being in them, so, on the theory, they are not conscious. The first-order information reaches the mechanisms that produce forced-choice responses and fails to reach, or fails to produce, the higher-order state that would make it conscious.

Other theories explain blindsight differently. A first-order theorist can say that the residual visual states are too degraded, or too weakly integrated, to be conscious, without mentioning any higher-order state. Blindsight therefore does not decide between the views. It establishes the premise they share: discrimination and awareness can come apart, and whatever makes the difference is not the presence of stimulus information alone. In the terms of Chapter 39, blindsight is an extreme dissociation between a first-order reading, which is good, and the second-order readings that report on it, which say nothing is there.

40.4 Evidence and Clarification

Lau and D. Rosenthal (2011) review empirical evidence for the higher-order view, concentrating on findings that distinguish it from first-order theories,

global workspace theories, and recurrent processing theories (compared in Chapter 44). They defend the view against two major criticisms: that the prefrontal activity associated with awareness reflects attention rather than awareness, and that prefrontal lesions do not abolish awareness. They conclude that although the approach began in philosophical discussion, it is testable and has received substantial empirical support.

R. Brown et al. (2019) address misunderstandings of the theory. They argue that higher-order states, though cognitively assembled, require considerably less than critics often presume, so that the theory is not an overintellectualization of consciousness, and that it occupies an intermediate position between global workspace theories and early sensory views. They also argue that the higher-order approach accounts better than its rivals for complex everyday experiences such as emotions and episodic memories, which makes it useful for thinking about pathological mental states.

The empirical program has a recognizable design. Match first-order performance while varying awareness, then ask where the brain differs. If performance on a discrimination is the same under two conditions and reported awareness differs, first-order theories predict that the difference lies in the first-order sensory representation, and higher-order theories predict that it lies in the higher-order machinery. The strength of the design is that it holds the first-order reading fixed, which is the reader-intervention logic of Chapter 58 applied to consciousness.

40.5 Objections

Four objections recur, and each clarifies what the theory claims.

The rock. Representing a rock does not make the rock conscious, so why should representing a mental state make the state conscious? Higher-order theorists restrict the claim to mental states, but then the question returns in a different form: what is it about being a mental state that makes being represented matter?

Targetless higher-order states. A higher-order thought can misrepresent. It can represent the subject as being in a state it is not in, or as being in a state that does not exist at all. The theory must say what, if anything, is then experienced. Rosenthal's answer is that what it is like for the subject is determined by how the higher-order thought represents the subject as being (D. M. Rosenthal 2005). The answer is coherent, and it has a striking consequence: the first-order state, which looked like the thing that was conscious, drops out of the account of what the experience is like.

Animals and infants. If consciousness requires thoughts about one's own mental states, creatures without the concepts needed to have such thoughts lack consciousness. The dispositional version makes the requirement a capacity rather than an occurrent thought, and on it a creature without the capacity

for higher-order thought would lack phenomenal consciousness; critics have pressed this consequence for animals and infants. R. Brown et al. (2019) argue that the conceptual requirements are much weaker than this objection assumes.

The explanatory target. Grant that a state is conscious exactly when it is suitably represented by a higher-order state. The criterion then says which states are conscious. It does not say why a state's being represented in that way is accompanied by there being something it is like to be in it, rather than by nothing. A program that, whenever it enters an internal state, also writes a record that it is in that state satisfies the bare structure of the condition. Higher-order theorists require more (thoughts, with the right concepts, about oneself, formed noninferentially), and a system meeting those further requirements could be engineered with ordinary means. A consistent functionalist reading of the theory should say that such a system is conscious. The theory would then have located consciousness in an organization, which is progress, and left open why that organization has experience attached, which is the question of Chapter 43.

Example 40.2 (One reading, three second-order states). A subject judges whether a faint grating is tilted left or right and is correct. A confidence reading assigns the judgment a probability of being correct, say 0.7. A higher-order thought represents the subject as seeing a leftward tilt. A report, "I saw it tilt left," expresses the higher-order thought. In blindsight, the first-order judgment can be correct while the confidence reading is low and the higher-order thought and the report are absent. In a targetless higher-order thought, the thought and the report can be present while the first-order state is absent. The three second-order items can be dissociated from one another as well as from the first-order state, and a theory must say which of them consciousness follows.

40.6 The Higher-Order State Is Indexical

The content of a higher-order thought, on Rosenthal's formulation, is that *one* is in the state: the thought represents the subject, as the subject, as being in it. A thought that some creature is in a state of seeing red, or that the organism with such-and-such a brain is, would not make the state conscious, even if the creature or organism happened to be oneself. The higher-order condition therefore contains the essential indexical of Chapter 37. What makes a state conscious, on this view, is a de se representation of it.

The observation connects the higher-order theory to the machinery of Part VIII in both directions. In one direction, it shows that the theory presupposes a self-locator: a system can form the required thought only if it can represent a state as its own, which requires the kind of self-specifying readings described in Chapter 36 and the kind of self-location formalized in Chapter 37. In the other direction, it shows how little that presupposition adds

on its own. A controller with a self-locator already has de se variables, and a system that represents its own internal states as its own, with the self-locator attached, satisfies the indexical part of the condition without further ingenuity.

The reading-loop notation puts the point compactly. A first-order reader D_1 produces a state s_1 from the mixture y. A higher-order reader D_{HO} takes s_1 as input and produces a state whose content is that the reader's own system is in s_1. It is a reader whose relatum is another of the system's readers and whose output is indexed to the system that contains both. Nothing in that description distinguishes a brain from a sufficiently elaborate program, and higher-order theorists who are functionalists accept the consequence. The question that remains is the one this part of the book has carried from chapter to chapter: why a reading of this form would be accompanied by experience.

40.7 The Fish and the Higher-Order Question

Nothing in the jamming avoidance response suggests a higher-order state. The fish reads a relation, acts on it, and, as far as is known, does not read the reliability of its reading (Chapter 39), let alone represent itself as being in a reading state. If a higher-order theory is right, the fish's electrosensory states are unconscious however intricate the computation that produces them, and the question of fish consciousness moves to other parts of the fish's brain, if anywhere. If a first-order theory is right, the answer depends on features of the fish's first-order processing that higher-order theories treat as irrelevant. The fish does not decide between the theories. It shows how much of the reading machinery of this book can be present without any state that a higher-order theory would count as conscious, and Chapter 50 returns to the question with the evidence from fish neurology.

40.8 Audit

Canonical phenomenon

A patient with damage to primary visual cortex denies seeing stimuli in part of the visual field yet discriminates them well above chance when required to guess (Weiskrantz 1986).

Minimal model

First-order state s_1 and higher-order state s_2 with content that the subject is in s_1; the HOP, actualist HOT, and dispositional HOT conditions (Definition 40.1).

What is relational here?

The relation between a mental state and a representation of it; in blindsight, the relation between a first-order discrimination and the second-order readings that report on it.

What physically reads it?
On higher-order theories, mechanisms that represent first-order states, which proponents associate with prefrontal cortex (Lau and D. Rosenthal 2011); in blindsight, whatever supports report fails to read what supports discrimination.

What existing field explains it?
Higher-order theories in philosophy of mind (D. M. Armstrong 1968; Lycan 1996; D. M. Rosenthal 2005; Carruthers 2000); the empirical case and its clarifications (Lau and D. Rosenthal 2011; R. Brown et al. 2019); the neuropsychology of blindsight (Weiskrantz 1986); the study of metacognition (Fleming et al. 2012).

What remains unexplained?
Why being represented by a higher-order state is accompanied by experience. The theories give a criterion for which states are conscious; the criterion is functional, can be satisfied by an engineered system, and leaves the explanatory question of Chapter 43 unanswered.

A discriminating experiment
Hold first-order performance fixed while varying reported awareness, by adjusting stimulus visibility until discrimination accuracy is matched across conditions that differ in reported awareness. Higher-order theories predict that the conditions differ in higher-order machinery and need not differ in first-order sensory representation; first-order theories predict the reverse. Chapter 61 discusses how to preregister such predictions so that each theory can lose.

40.9 Notes and References

Inner-sense or higher-order perception theories go back to D. M. Armstrong (1968) and are developed by Lycan (1996). The actualist higher-order thought theory is set out in D. M. Rosenthal (2005); the dispositional version in Carruthers (2000). Lau and D. Rosenthal (2011) review the empirical case for the higher-order view and reply to criticisms; R. Brown et al. (2019) clarify the theory's commitments and distinguish it from global workspace and early sensory views. Blindsight is described in Weiskrantz (1986). The relation between metacognition and consciousness is discussed by Fleming et al. (2012), and the comparison with other theories of consciousness continues in Chapter 44.

40.10 Exercises

Exercise 40.1. Write the three conditions of Definition 40.1 as predicates on a set of states with a "represents" relation and an "available to" relation.

Construct a case in which a state is conscious on the dispositional condition and not on the actualist one, and a case in which the reverse holds.

Exercise 40.2. Using the signal-detection model of Chapter 39, describe a subject whose first-order discrimination has sensitivity well above zero while its confidence is constant at the lowest level. What must the second-order reader be doing to produce this pattern, and which feature of blindsight does the model fail to capture?

Exercise 40.3. Describe a targetless higher-order thought in the formalism of Definition 40.1. State what each of the three conditions implies about whether anything is conscious, and what is experienced, in that case.

Exercise 40.4. Write a short program in which every internal state change is accompanied by a record that the program is in the new state. List the further requirements that higher-order theorists place on higher-order thoughts, and for each say whether your program could be extended to meet it and how.

Exercise 40.5. For a perceptual decision about the direction of moving dots, write the content of a first-order state, of a confidence reading, of a higher-order thought, and of a verbal report. Identify a manipulation that should change each while leaving the others fixed.

Exercise 40.6. Apply Definition 40.1 to the jamming avoidance response. Which of the conditions could be tested in the electric fish, what would count as evidence of a higher-order state, and what would a negative result show?

41

Self-Models

> For my part, when I enter most intimately into what I call myself, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch myself at any time without a perception, and never can observe any thing but the perception.
>
> David Hume, *A Treatise of Human Nature* (1739), 1.4.6

Every reader in this book so far has read something other than itself: a beat, a neighbor's field, the gradient of an attractant, the attention of another agent. A self-model is a reader whose object includes the system that does the reading. It estimates the system's own body, boundaries, location, and states, and the system uses the estimate to act. Self-models of this kind are among the best-understood structures in motor control and multisensory perception, and they can be built into robots. Thomas Metzinger's self-model theory adds a stronger claim: when a system cannot recognize its self-model as a model, the system is, for itself, a self.

The thesis of this chapter is that the first half of this story is established and the second half is where the residue lies. Self-models explain the structure of self-experience (ownership, self-location, perspective, and agency) and they explain why that structure can be manipulated in the laboratory. They do not by themselves explain why modeling oneself should be experienced, and the ingredient most often proposed to close the gap, recursion, is too cheap to do the work.

41.1 A Reader Whose Object Includes Itself

In the notation of the reading loop, a system reads $z = \mathsf{D}(y)$ from a mixture $y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})$ in which its own state s_{A} is one of the terms. Most of the readers in Parts IV to VIII extract a relation between the system and something else. A self-model extracts, or estimates, the system's own contribution.

Definition 41.1 (Self-model)**.** A *self-model* is a component of a system's internal model whose variables are estimates $\hat{s}_{\mathrm{A}}$ of the state of the system that contains the model (body configuration, location, boundaries, internal states, or the states of its own processes), and which the system uses in the control of its own behavior.

The definition is functional and says nothing about experience. It has two clauses on purpose. A stored description of a system that the system never uses is a record, and a control loop that never estimates the controlled system's state is a reflex; a self-model is the estimate put to work.

The standard example comes from motor control. Sensory input caused by one's own movement must be separated from input caused by the world, and the brain is thought to do this with a forward model that predicts the sensory consequences of its own motor commands (Wolpert et al. 1995; Kawato 1999).

Model 41.2 (Self-model as a forward model and state estimator)**.** Let a_t be the motor command at time t, $\hat{s}_t$ the estimate of the system's own state, and y_{t+1} the next sensory input. The self-model predicts, compares, and corrects:

$$\hat{s}^{-}_{t+1} = \hat{f}(\hat{s}_t, a_t), \tag{41.1}$$

$$\hat{y}_{t+1} = \hat{g}(\hat{s}^{-}_{t+1}), \tag{41.2}$$

$$\hat{s}_{t+1} = \hat{s}^{-}_{t+1} + K\,(y_{t+1} - \hat{y}_{t+1}). \tag{41.3}$$

Here $\hat{f}$ is the model's estimate of its own dynamics, $\hat{g}$ its estimate of how its state appears to its own sensors, and K a gain. The discrepancy $e_{t+1} = y_{t+1} - \hat{y}_{t+1}$ is attributed to the world when it is large and to the self when it is small.

The discrepancy e_{t+1} is a relational variable in the sense of Chapter 1: a difference between an actual and a predicted signal, belonging to neither alone. The comparator account of the sense of agency uses exactly this variable. Self-produced touch is predicted and attenuated, which is why one cannot tickle oneself (Blakemore et al. 1998), and several disturbances in the awareness of action have been modeled as failures of the comparison (Frith et al. 2000; Haggard 2017).

Remark 41.3. Conant and Ashby proved that every good regulator of a system must be a model of that system (Conant and Ross Ashby 1970), and the internal model principle of control theory makes a related claim for tracking and disturbance rejection (Francis and Wonham 1976). Applied to a system that regulates its own body, these results give a formal reason to expect self-models wherever self-regulation is good. The application is an interpretation of the theorems, and the theorems concern regulation, not experience.

Example 41.4 (A minimal self-model in an electric fish). The pulse-type mormyrid fishes of Africa receive, at every electric organ discharge, a large sensory signal caused by their own discharge. A corollary discharge from the

electromotor command drives an adaptive negative image in cerebellum-like circuits of the electrosensory lateral line lobe, which cancels the predictable part of this self-caused input; when the reafference changes, the negative image changes with it (C. C. Bell 1981; C. Bell et al. 1997; Requarth and Sawtell 2011). The negative image is a self-model in the sense of Definition 41.1: a learned estimate of the system's own contribution to its input, used to separate self from world. The wave-type fishes of Part V show the contrast. The jamming avoidance response of *Eigenmannia* and *Gymnarchus* is computed from the sensory mixture alone, with no reference to the fish's own pacemaker command (Heiligenberg 1991a; Kawasaki 1993). An animal can read a relation between itself and its neighbor without modeling itself at all (Chapter 22).

The example fixes a point that the rest of the chapter depends on. Self-models come in degrees and in kinds. A negative image models one channel of reafference; a body schema models the configuration of the limbs; a global self-model, if there is one, models the organism as a whole and places it in a model of the world. The philosophical claims concern only the last kind.

41.2 Metzinger's Self-Model Theory

Metzinger's *Being No One* develops the strongest version of the thesis that selves are self-models (Metzinger 2003). Its central claim is that there are no selves in the world, only systems that operate under a particular kind of self-model; the experience of being someone is the content of that model (Metzinger 2003; Metzinger 2009). Three constructs carry the theory.

The *phenomenal self-model* (PSM) is the conscious part of a system's representation of itself as a whole: its body, its states, its location, and its relations to the world. The *phenomenal model of the intentionality relation* represents the system as standing in relations to objects (seeing this, attending to that, wanting the other) and thereby supplies the structure of a first-person perspective (Metzinger 2003). The third construct, transparency, does the most work.

Definition 41.5 (Transparency)**.** A representation is *transparent* for a system when the system has access to its content but not to its character as a representation: the earlier processing stages, and the fact that the content is an estimate, are not available to the system's own introspection. A representation is *opaque* when its representational character is available, as when one experiences an image as an image.

The claim is that a system operating under a transparent self-model does not experience itself as having a model of a self. It experiences itself as a self, because the model's vehicle is invisible to it and only the content remains (Metzinger 2003). The world-model is transparent in the same way: we experience the world, not a model of the world. On this view the first-person

perspective is the centering of a transparent world-model on a transparent self-model, together with a transparent model of the relations between them.

In the terms of this book, Metzinger's theory takes the agent-centered variables of Part VIII (here, now, me) and adds one functional condition: the system cannot represent those variables as variables. Transparency is a limit on what one reader, introspection, can read of another reader, the self-model. It is therefore a relational property, a fact about how two components of the same system stand to each other, and it is specifiable in functional terms.

41.3 Building and Breaking the Model

If self-experience is the content of a model, it should be possible to change the experience by feeding the model misleading evidence. It is.

In the rubber hand illusion, a participant watches a rubber hand being stroked while their own hidden hand is stroked in synchrony, and comes to feel the touch on the rubber hand and to feel the rubber hand as their own (Botvinick and J. Cohen 1998). Full-body versions of the procedure use virtual reality. When participants in virtual reality saw a virtual body in front of them receive touches synchronized with touches they felt, they came to feel the virtual body as their own and mislocalized themselves toward it, outside their bodily borders (Lenggenhager et al. 2007). Ehrsson induced the experience of being located outside one's physical body and looking at it from another place by combining the visual perspective of a camera behind the participant with correlated touch (Ehrsson 2007). Blanke and Metzinger proposed that such experiments isolate the components of *minimal phenomenal selfhood*: self-identification with a body, self-location, and a first-person perspective (Blanke and Metzinger 2009). Tsakiris's neurocognitive model of body ownership assigns the relevant comparisons to a network of multisensory areas that includes parietal and premotor cortex (Tsakiris 2010).

The variable that drives these illusions is relational. Stroking the rubber hand without stroking the real hand does nothing, and so does stroking the real hand alone; what matters is the temporal relation between the seen and the felt touch.

Model 41.6 (Ownership as inference from a temporal relation (illustrative))**.** Let Δt be the delay between a seen touch and a felt touch. Suppose the system entertains two hypotheses: the two signals have a common cause (the seen hand is its hand), with prior probability p_c and delays distributed as $\mathcal{N}(0, \sigma^2)$; or they have independent causes, with delays spread uniformly over a window of width W. The posterior probability of a common cause is

$$P(\text{common} \mid \Delta t) = \frac{p_c \mathcal{N}(\Delta t; 0, \sigma^2)}{p_c \mathcal{N}(\Delta t; 0, \sigma^2) + (1 - p_c)/W}. \tag{41.4}$$

The parameters are illustrative; the model's point is structural.

Proposition 41.7 (Ownership reads a relation)**.** *In Model 41.6, the attribution of ownership depends on the two sensory streams only through their temporal relation Δt. Changing the timing of both streams together, or the intensity of either alone within the model's assumptions, leaves the attribution unchanged.*

Proof. The right-hand side of (41.4) is a function of Δt alone, and Δt is invariant under a common shift of both event times. □

The proposition is a statement about the model, and the model is a hypothesis about the brain. Its value is that it names the discriminating manipulation: vary Δt while holding each stream's content fixed, and ownership should follow Δt.

Self-models can also be built. Bongard, Zykov, and Lipson gave a four-legged robot no fixed description of its own body. The robot inferred its structure from the relations between its actuation and its sensation, used the inferred self-model to generate locomotion, and, when part of a leg was removed, revised the model and generated a new gait (Bongard, Zykov, et al. 2006). The robot satisfies Definition 41.1 without difficulty. Nobody supposes that it is a self, and the question of what it lacks is the question of this chapter.

41.4 Model and Model-Owner

If the self is a model, whose model is it? The question threatens a regress: a model seems to need a user, and if the user is a self, the self has not been explained. Metzinger's answer is that the model belongs to the organism in the way that any of its states belongs to it, as a part used by other parts, and that the *experience* of ownership is one more content of the self-model (Metzinger 2003). There is no additional owner to be found, and the search for one repeats the error Hume described: introspection finds perceptions, never a perceiver standing apart from them.

Other traditions reach similar conclusions from different premises. Gallagher distinguishes a *minimal self*, the pre-reflective sense of being the owner and agent of one's experiences, from a *narrative self* built over time from stories and memories (Gallagher 2000). Dennett treats the self as a center of narrative gravity, an abstraction useful for organizing a life's story, real in the way a center of mass is real (Dennett 1991). Both accounts, and Metzinger's, locate the self in a structure of representations and deny that it is an extra entity.

The dispute that remains concerns the word "experience." A minimal self-model can be specified completely in the vocabulary of Definitions 41.1 and 41.5. Metzinger's theory says that operating under such a model, with the right degree of transparency and integration, is what having a phenomenal self consists in. That is an identity claim. It explains why selves seem to exist and

why the seeming has the structure it has; whether it explains why the seeming is felt is the question Part X takes up.

41.5 Why Recursion Is Not Enough

Self-reference has often been proposed as the missing ingredient of subjectivity. Hofstadter argues that an "I" is a strange loop: a pattern in a symbol-manipulating system that comes to represent itself, and whose self-representation feeds back on the system's behavior (Hofstadter 2007). The idea is attractive and captures something right about how selves are organized. As a criterion for subjectivity it faces a difficulty of cost.

Proposition 41.8 (Self-reference is cheap)**.** *Let a system have any state variable s and a register m updated by $m \leftarrow h(s)$ for some function h, and let the system's behavior depend on m. Then the system contains a representation of its own state that it uses in control, and so satisfies Definition 41.1 in its minimal form. Iterating the construction ($m_2 \leftarrow h_2(m)$, and so on) yields self-models of self-models to any finite depth.*

Proof. By construction m is a function of the system's state and enters its dynamics; the same holds for each m_k with respect to m_{k-1}. □

A thermostat that records its own setpoint and adjusts its hysteresis accordingly satisfies the proposition, as does an operating system whose process table lists the process that manages the table, and so does the robot of Section 41.3. If self-reference alone sufficed for subjectivity, subjectivity would be everywhere in engineering, and the claim would lose the discriminating force that made it interesting.

What distinguishes a candidate self-model for subjectivity therefore has to lie in additional conditions, and Metzinger supplies them: the self-model must be transparent, it must be integrated into a global model of the world, and it must be embedded in a window of presence, a represented now (Chapter 35) (Metzinger 2003). Each condition is functional. Each can, in principle, be built into an artificial system. The residue survives the addition of all of them: it is the claim that a system satisfying them is someone, as opposed to a system that behaves, reports, and attributes as if it were. The theories of Chapter 44 differ precisely in what they add to close that gap, and Chapter 43 states the gap itself.

41.6 Audit

Canonical phenomenon

The rubber hand and full-body illusions: synchronous visuotactile stimula-

tion shifts felt ownership and self-location to a body that is not one's own (Botvinick and J. Cohen 1998; Lenggenhager et al. 2007; Ehrsson 2007).

Minimal model

A forward model and state estimator of the system's own state (Model 41.2); ownership as inference of a common cause from the temporal relation between two sensory streams (Model 41.6); transparency as the inaccessibility of a representation's vehicle to introspection (Definition 41.5).

What is relational here?

The discrepancy between predicted and actual reafference; the delay Δt between seen and felt touch; and, in Metzinger's theory, the centering of the world-model on the self-model together with the relation of transparency between introspection and the self-model.

What physically reads it?

Multisensory circuits in premotor and parietal cortex compare the streams and update the body model (Tsakiris 2010; Blanke and Metzinger 2009); in the mormyrid, cerebellum-like circuits of the electrosensory lateral line lobe hold the negative image (C. Bell et al. 1997).

What existing field explains it?

Motor control and internal models (Wolpert et al. 1995; Kawato 1999), comparator accounts of agency (Frith et al. 2000; Haggard 2017), multisensory integration and body ownership (Botvinick and J. Cohen 1998; Tsakiris 2010), control theory (Conant and Ross Ashby 1970; Francis and Wonham 1976), robotics (Bongard, Zykov, et al. 2006), and the philosophy of self from Hume to Metzinger (Metzinger 2003; Gallagher 2000; Dennett 1991). Together they explain the construction, the structure, and the manipulability of self-experience.

What remains unexplained?

Why operating under a transparent, integrated self-model should amount to being someone for whom things are some way. Metzinger's identity claim answers the question by stipulation; recursion does not answer it at all.

A discriminating experiment

Dissociate transparency from content. In the rubber hand paradigm, vary Δt parametrically while telling participants, truthfully, which hand is theirs and asking them to rate both ownership and their belief about ownership. The self-model account predicts that felt ownership tracks Δt even when belief does not, because the transparent model is not updated by knowledge of its own falsity; an account on which ownership is a judgment predicts that the two ratings move together.

41.7 Notes and References

Metzinger (2003) is the full statement of the self-model theory of subjectivity, and Metzinger (2009) is its accessible version. Blanke and Metzinger (2009) connect the theory to the experimental study of bodily self-consciousness; the founding experiments are Botvinick and J. Cohen (1998) for the rubber hand and Lenggenhager et al. (2007) and Ehrsson (2007) for the full body. Tsakiris (2010) gives a neurocognitive model of body ownership. Forward models in motor control are introduced by Wolpert et al. (1995) and reviewed by Kawato (1999); the comparator account of self-produced sensation is tested by Blakemore et al. (1998) and extended to disorders of agency by Frith et al. (2000); Haggard (2017) reviews the sense of agency. For the mormyrid negative image, see C. Bell et al. (1997) and Requarth and Sawtell (2011); the contrast with the wave-type jamming avoidance response is developed in Chapter 22. Gallagher (2000) separates minimal from narrative selves; Dennett (1991) gives the center-of-narrative-gravity view; Hofstadter (2007) gives the strange-loop view. Bongard, Zykov, et al. (2006) is the robot that models itself.

41.8 Exercises

Exercise 41.1. Implement Model 41.2 for a one-dimensional limb with dynamics $s_{t+1} = s_t + a_t + \eta_t$ and observation $y_t = s_t + \nu_t$, with Gaussian noise. Add a delay of d steps to the observation and show how the discrepancy e_t grows with d when the model ignores the delay. At what delay would a threshold on $|e_t|$ start attributing self-produced movement to the world?

Exercise 41.2. Plot (41.4) against Δt for $\sigma = 100\,\mathrm{ms}$, $W = 1\,\mathrm{s}$, and $p_c \in \{0.2, 0.5, 0.8\}$. Find the delay at which the posterior falls to one half, as a function of p_c. State which parameter a prior belief such as "that is a rubber hand" would change, and predict the effect on the illusion.

Exercise 41.3. A negative image $n(\tau)$ is subtracted from the self-caused input $u(\tau)$ at latency τ after each discharge, and adapts by $n \leftarrow n + \eta\,(u - n)$ after each discharge. Show that n converges to u for $0 < \eta < 2$ when u is constant, and compute the residual when u switches to a new constant. Which variable of the fish is the system modeling, and which is it failing to model?

Exercise 41.4. Give a functional criterion for transparency (Definition 41.5) that could be applied to an artificial agent. Using your criterion, decide whether the robot of Bongard, Zykov, et al. (2006) has a transparent self-model, and whether the answer could be changed by adding a component to the robot.

Exercise 41.5. Construct three physical systems that satisfy Proposition 41.8 with depth at least two. For each, list which of Metzinger's further conditions (transparency, integration into a world-model, a window of presence) it lacks.

Exercise 41.6. Place *Eigenmannia*, a mormyrid, the robot of Bongard, Zykov, et al. (2006), and a human on a scale of self-modeling, using Definition 41.1 and the kinds of self-model distinguished at the end of Section 41.1. Which step on your scale, if any, would a theory of consciousness have to point to, and why?

42

Attention Models

> Every one knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought.
>
> William James, *The Principles of Psychology* (1890), ch. 11

Attention is a selection process: many signals compete for a limited capacity to influence behavior, and some win. A system that controls its own attention faces a control problem, and control problems are solved better with a model of the thing controlled. The attention schema theory of Michael Graziano proposes that the brain builds such a model, a simplified description of its own attention, and that the content of this model is what we call awareness (Graziano 2013; Graziano and Webb 2015).

The thesis of this chapter is that the attention schema is a reader whose input is another reader, that the theory built on it makes predictions of an unusual kind for a theory of consciousness, predictions about the *control* of attention, and that several have been tested and partly confirmed (Webb et al. 2016; Wilterson and Graziano 2021). The theory explains why systems claim to be aware, and why they attribute awareness to others. It declines to explain an additional phenomenal property, because it denies that there is one, and that refusal is where it meets the illusionism of Chapter 43.

42.1 Attention as Selection

Two findings anchor the modern study of attention. Posner showed that a cue indicating where a target is likely to appear speeds its detection at the cued location and slows it elsewhere, even when the eyes do not move (Posner 1980). Attention could be moved covertly, and its movement had measurable consequences. Desimone and Duncan then organized a large body of single-unit and behavioral evidence under the *biased competition* account: objects in the visual field compete for representation in neural populations with limited

capacity, and the competition is biased by bottom-up salience and by top-down signals that favor task-relevant objects (Desimone and Duncan 1995).

A schematic model makes the structure explicit.

Model 42.1 (Biased competition, schematic)**.** Let items $i = 1, \dots, n$ present inputs of strength $x_i > 0$, and let an attentional bias $b_i > 0$ weight each item. The share of processing capacity that item i obtains is

$$\alpha_i = \frac{b_i\, x_i}{\sum_{j=1}^{n} b_j\, x_j}, \qquad \sum_i \alpha_i = 1. \tag{42.1}$$

Attentional control is the adjustment of the biases $\boldsymbol{b}$ toward items relevant to the current task.

The model is illustrative and deliberately minimal; the literature contains far richer versions. It shows two things this chapter needs. The shares α_i are relational: each depends on every item's input and bias, and an item's share can fall because another item grew stronger. And the controller, the process that sets $\boldsymbol{b}$, needs to know something about $\boldsymbol{\alpha}$ to do its job: where attention currently is, how strongly it is held, and what is about to capture it. That knowledge is a reading of a reading.

Example 42.2 (Attention in the electric fish)*.* The jamming avoidance response of Part V involves no attention in this sense. The fish's midbrain pools sign information across the body surface, and the behavior emerges from a large population of imprecisely tuned neurons with no bottleneck at which one input wins (Heiligenberg, Baker, et al. 1978; Heiligenberg and G. Rose 1985). Pooling is the opposite of selection. The contrast marks the step this chapter adds: a system that must choose among relations, and that therefore benefits from representing its own choosing.

42.2 Controlling a Process Requires a Model of It

Engineering has a precise version of the claim that control needs a model. Conant and Ashby showed that the simplest optimal regulator of a system is, in a defined sense, a model of that system (Conant and Ross Ashby 1970), and the internal model principle states that a controller that rejects a class of disturbances must contain a model of the process that generates them (Francis and Wonham 1976). Graziano's argument has the same shape. The brain controls the body better with a body schema, a simplified model of the body's configuration used for movement planning; by analogy, it should control attention better with an attention schema, a simplified model of attention (Graziano 2013; Graziano and Webb 2015).

Definition 42.3 (Attention schema)**.** An *attention schema* is a simplified internal model of a system's own attentional state (what is being selected, how

strongly, and with what likely consequences) that the system uses to control its attention.

Why a model helps can be made exact in the simplest control problem with a delay, which is the situation of any controller that observes its own process through slow channels.

Proposition 42.4 (A model of the process restores the delay-free stability range)**.** *Let a controlled quantity evolve as $q_{t+1} = q_t + u_t$ and be observed with a delay of d steps, $o_t = q_{t-d}$. Under proportional control on the observation, $u_t = -k\,o_t$, the loop is stable if and only if $0 < k < 2\sin\big(\pi/(2(2d+1))\big)$. If instead the controller acts on the prediction $\hat{q}_t = o_t + \sum_{j=1}^{d} u_{t-j}$ supplied by an exact model of the process, $u_t = -k\,\hat{q}_t$, the loop is stable if and only if $0 < k < 2$, whatever the delay.*

Proof. Without the model, the characteristic equation is $z^{d+1} - z^d + k = 0$. At $k = 0$ its roots are $z = 1$ and $z = 0$; stability is lost when a root crosses the unit circle at $z = \mathrm{e}^{\mathrm{i}\omega}$, which requires $|\mathrm{e}^{\mathrm{i}\omega} - 1| = k$ and $\omega d + \arg(\mathrm{e}^{\mathrm{i}\omega} - 1) = \pi$. Since $\arg(\mathrm{e}^{\mathrm{i}\omega} - 1) = \pi/2 + \omega/2$ for $0 < \omega < 2\pi$, the second condition gives $\omega = \pi/(2d+1)$, and the first gives $k = 2\sin(\omega/2)$. With the model, $\hat{q}_t = q_t$ exactly, the loop is $q_{t+1} = (1-k)\,q_t$, and it is stable if and only if $|1-k| < 1$. □

With a delay of one step, the stable gains shrink from $(0, 2)$ to $(0, 1)$; with a delay of five steps, to about $(0, 0.28)$. A controller that must act fast on a slow observation of its own process gains a large margin from a model of that process. The proposition is about controllers in general, and it gives the attention schema theory's central premise a precise form: if attention is controlled through delayed signals about attention, a model of attention pays for itself.

In the notation of the reading loop, the first-order reader is the competition of Model 42.1, whose output is the allocation $\boldsymbol{\alpha}$. The attention schema is a second reader, $\hat{\boldsymbol{\alpha}} = \mathsf{D}_{\mathrm{AS}}(\boldsymbol{\alpha}, \ldots)$, whose output feeds the policy that sets the biases, $\boldsymbol{b} = \pi(\hat{\boldsymbol{\alpha}}, \text{goals})$. Its object is a relation between the system and whatever it is selecting.

Whether a controller benefits from such a model is an empirical question about controllers, and it has been asked of artificial ones. Wilterson and Graziano trained a deep reinforcement-learning agent to move an attention spotlight to track a stimulus in a simple catch task. With an attention schema available, the agent learned to control its spotlight and to solve the task. Disabling the schema after learning greatly reduced performance, and removing it before learning impaired learning, even though the agent received enough information in both conditions that it could in principle have learned without it (Wilterson and Graziano 2021). The result shows that a descriptive model of attention can help a controller of attention. It shows nothing yet about brains or about awareness.

42.3 Awareness as the Content of the Schema

The attention schema theory makes a further claim. The schema, like the body schema, is simplified: it does not describe competing neural signals and normalization. It describes something closer to a mental possession of an object by a self: an inner grasp that can be directed from one object to another. When the system consults its schema and reports on it, what it reports is that it *is aware* of the object. On the theory, awareness is the content of the attention schema: the brain's description of its own attention, and the reason the brain concludes that it has a subjective experience (Graziano 2013; Graziano and Webb 2015).

The theory offers three advantages, which its authors state as predictions (Graziano and Webb 2015). Awareness and attention should usually go together, because one is a model of the other, and should sometimes come apart, because models are imperfect. Awareness should extend to whatever attention can select, internal states included. And attention without awareness should still be possible, but its control should suffer.

The first prediction had evidence before the theory existed. In a patient with blindsight after damage to primary visual cortex, cues presented at locations the patient could not see still directed attention, speeding responses to stimuli that also did not enter awareness (Kentridge et al. 1999). Koch and Tsuchiya reviewed psychophysical evidence that top-down attention and consciousness can be manipulated separately and can even have opposite effects (Koch and Tsuchiya 2007). Such dissociations are consistent with the attention schema theory, and also with several of its rivals, so they do not discriminate among theories.

The third prediction is the distinctive one, because it concerns the quality of control. Webb, Kean, and Graziano used metacontrast masking to make participants aware or unaware of a visual stimulus that served as a spatial cue. Attention was drawn to the stimulus whether or not participants were aware of it. Without awareness, attention changed in three ways: attention to a task-relevant stimulus was less stable over time, inhibition of return was reduced, and attention was driven more by the stimulus's luminance contrast (Webb et al. 2016). The authors interpret these as the control deficits the theory predicts.

Hypothesis 42.5 (Control deficit without awareness)**.** If awareness is the content of a model used to control attention, then manipulations that remove awareness of an attended stimulus while leaving its capture of attention intact will degrade the endogenous control of that attention (its stability, its disengagement, its resistance to salience) in ways predictable from control theory, and manipulations that improve the model will improve control.

The hypothesis is stated so that it can fail. If awareness could be removed with no change in the dynamics of attentional control, the theory would lose

its main empirical handle.

42.4 A Reader of Readers, and What It Reads

The attention schema is a relational reader in two senses. Its object is a relation: attention is a relation between a system and what it selects, and the schema represents that relation. And its output is agent-centered in the sense of Part VIII: "I am attending to that" is a variable indexed to the system itself. The schema is also used outward. The theory proposes that the same machinery models the attention of other agents, which is why we attribute awareness to people, animals, and sometimes machines, and why the attribution follows cues to where others are attending (Graziano 2013).

The theory's account of why awareness seems nonphysical follows from the simplification. A model that described attention as competing signals under normalization would be accurate and useless for fast control. A model that describes attention as a movable inner grasp is inaccurate and useful. Its content is what introspection reports, and that content has no physical mechanism in it, so the system reports an experience that seems to float free of mechanism (Graziano 2013).

This account makes the attention schema theory a mechanistic partner of illusionism. Frankish argues that phenomenal consciousness is an introspective illusion: we represent our experiences as having phenomenal properties that they do not have, and the task is to explain the representation rather than the property (Frankish 2016). The attention schema theory proposes a specific representation, a specific reason for its simplification, and a specific control function that it serves. Chapter 43 states what an illusionist position gains and what it gives up.

42.5 Limits

Three limits bound what the theory can claim.

First, the dissociations between attention and awareness that support it are predicted by other theories too. A global workspace account ties conscious access to amplification and broadcast, which attention can influence without being identical to it; a higher-order account ties awareness to a representation of one's own states, of which a model of attention is one example (Chapters 40 and 44). Only the control predictions separate the attention schema theory from these rivals, and those predictions are new enough that independent replication is still the main requirement.

Second, the theory explains reports, attributions, and control, all of which are functional. A reader who holds that phenomenal consciousness is real and irreducible will say that the theory has changed the subject. Its proponents accept the charge and answer that the subject was confused. The dispute

cannot be settled by the experiments above, because both sides predict their outcomes.

Third, a model of attention is a reader of a reader, and Chapter 41 showed that such recursion is cheap. An operating system's scheduler that monitors its own queue models its own allocation of resources. What distinguishes the attention schema, if anything does, is its simplification and its double use, for controlling one's own attention and for modeling others', and its integration into a global model of self and world. Recursion alone does not distinguish it.

42.6 Audit

Canonical phenomenon

Attention without awareness: masked or blindsight cues still capture attention, and the control of that attention degrades (Kentridge et al. 1999; Webb et al. 2016).

Minimal model

Biased competition for capacity (Model 42.1), with a controller that sets the biases from a simplified model of the allocation (Definition 42.3).

What is relational here?

The attentional relation between a system and what it selects, represented by the schema as the variable "I am attending to X", agent-centered and simplified.

What physically reads it?

In the theory, a control network that monitors the allocation of attention and, in social contexts, the attention of others (Graziano 2013); in the artificial agent of Wilterson and Graziano (2021), an explicit schema module whose removal impairs control.

What existing field explains it?

Attention research (Posner 1980; Desimone and Duncan 1995), control theory (Conant and Ross Ashby 1970; Francis and Wonham 1976), social cognition, and illusionism (Frankish 2016). The theory assembles these into one account.

What remains unexplained?

Whether a simplified internal model of attention exhausts what "awareness" names. The theory says yes; phenomenal realists say no; the residue is the disagreement itself, which the chapter's experiments cannot resolve.

A discriminating experiment

Change the schema without changing attention. Train participants, with feedback on their awareness reports only, to report awareness of masked cues at a different threshold, while holding cue capture constant as measured by

reaction times. The theory predicts that the stability and disengagement of attention will shift with the trained reports; a theory in which reports are downstream readouts with no control function predicts no change in attentional dynamics.

42.7 Notes and References

Graziano (2013) is the book-length statement of the attention schema theory, and Graziano and Webb (2015) a compact review with predictions. Webb et al. (2016) tests the control predictions in human observers, and Wilterson and Graziano (2021) tests the benefit of a schema in an artificial agent. The attention literature the theory builds on begins with Posner (1980) and is organized by Desimone and Duncan (1995); dissociations of attention and awareness are reviewed by Koch and Tsuchiya (2007), and attention without awareness in blindsight is demonstrated by Kentridge et al. (1999). The control-theoretic results are Conant and Ross Ashby (1970) and Francis and Wonham (1976). The relation to illusionism is developed in Frankish (2016) and discussed further in Chapter 43.

42.8 Exercises

Exercise 42.1. In Model 42.1, let $n = 2$, $x_1 = 1$, and $b_1 = b_2 = 1$. Compute α_1 as a function of x_2 and show that α_1 falls as x_2 rises although nothing about item 1 has changed. What bias ratio b_1/b_2 restores $\alpha_1 = 0.8$ when $x_2 = 3$?

Exercise 42.2. A controller adjusts b_1 to hold α_1 at a target, but it observes α_1 only with a delay d. Simulate proportional control with and without a predictive model of the allocation (a model that forecasts α_1 over the delay). Show that the model-free controller oscillates for large gains and delays, and relate the result to the claim that an attention schema improves control.

Exercise 42.3. In a Posner cueing experiment with 80% valid cues, mean reaction times are 300 ms on valid trials and 340 ms on invalid trials (illustrative values). Define the validity effect and compute it. Design a version in which the cue is masked, and state which outcome would support Hypothesis 42.5 and which would count against it.

Exercise 42.4. Describe an operating-system scheduler that maintains a model of its own allocation of processor time. Does it satisfy Definition 42.3? List what it lacks relative to the attention schema proposed for brains, and decide whether any item on your list is non-functional.

Exercise 42.5. Take one attention–awareness dissociation (for example, attention without awareness in blindsight) and write the prediction that the attention schema theory, a global workspace theory, and a higher-order theory each make for it. Which of the three predictions, if any, differ?

Exercise 42.6. The jamming avoidance response pools information across the body instead of selecting it (Example 42.2). Invent a variant of the task that would require the fish to select among neighbors. What would an attention schema for this fish represent, and what behavioral evidence would show that it has one?

Confidence, metacognition, higher-order representation, self-models, and attention schemas are readers whose input includes other readers. Each has been proposed, by someone, as the ingredient that turns processing into experience. Recursion is cheap in engineering and ubiquitous in biology, so the proposal needs more than recursion. Part X states the gap that all of this machinery must cross, and surveys the theories that claim to cross it.

Part X

Consciousness

Why should any of these relations be experienced?

43

Functional Perspective Versus Phenomenal Perspective

> [A]n organism has conscious mental states if and only if there is something that it is like to be that organism
>
> Thomas Nagel, "What Is It Like to Be a Bat?" (1974)

Parts IV to IX built up, one capability at a time, systems that read relations, act on what they read, maintain themselves, misrepresent, predict, center their variables on themselves, and model their own reading. Each capability was specified in functional terms and anchored in systems that have been studied. This chapter states the distinction that all of that machinery leaves standing.

A system has a *functional perspective* when it reads and uses variables centered on itself. A system has a *phenomenal perspective* when there is something it is like for the system to be in its states. The thesis of the chapter is that system-relative information is not yet experience: every property on the functional list can be specified, measured, and built, and none of them, as specified, settles whether the phenomenal fact obtains. The gap between the two is the explanatory gap of the philosophy of mind, reached here from below, and the chapter closes by stating what would count as closing it.

43.1 Two Perspectives

The book's vocabulary makes the first notion precise.

Definition 43.1 (Functional perspective)**.** A system S has a *functional perspective* to the extent that

1. some of the variables $z = \mathsf{D}(y)$ it reads are agent-centered: indexed to its own location, time, body, or action (Part VIII);
2. it uses those variables in control, through a policy $a = \pi(z)$ whose effects feed back on the next mixture; and

3. the variables are integrated into a model in which S figures as the origin of the coordinates.

The definition admits degrees. A thermostat satisfies none of the clauses in any interesting sense; the jamming avoidance response of Part V satisfies the first two clearly and the third in a minimal form; a mammal with a hippocampal map, a body model, and an attention schema satisfies all three richly.

Definition 43.2 (Phenomenal perspective)**.** A system S has a *phenomenal perspective* if there is something it is like for S to be in its states.

The second definition borrows Nagel's formula, and the words matter. Nagel's full sentence says that an organism has conscious mental states if and only if there is something it is like to be that organism, "something it is like for the organism" (Nagel 1974, p. 436). The last clause is perspectival. It says that the facts in question obtain *for* someone, and Section 43.6 returns to it.

The gap can be stated as a list. A system can possess agent-relative sensing, self–world discrimination, prediction, representation, error correction, and self-maintaining control, and it can remain open whether there is something it is like to be that system. Each item on the list has appeared in an earlier chapter with a concrete carrier.

Example 43.3 (The fish, item by item). The weakly electric fish generates the field it senses, so its sensory variables are relative to its own action (Chapter 19). Mormyrid fishes cancel the predictable part of their own reafference with a learned negative image, which separates self-caused from world-caused input and predicts the former (C. Bell et al. 1997). Midbrain neurons of *Eigenmannia* represent the sign of the frequency difference between the fish and its neighbor (G. Rose and Heiligenberg 1986b). Non-jamming stimuli can evoke the jamming avoidance response, so the fish's classification can be wrong (Carlson and Kawasaki 2007). The response itself is closed-loop control that restores a relation the fish needs for electrolocation (Heiligenberg 1991a; Madhav et al. 2013). And the fish, like every organism, maintains the conditions of its own existence (Chapter 27). Every item on the list has a carrier in this animal. Whether there is something it is like to be the fish is the question Chapter 50 asks and cannot answer.

43.2 Access and Phenomenal Consciousness

Block drew the distinction that consciousness science has used ever since. *Phenomenal consciousness* is experience: what it is like to see red, to feel pain, to hear a tone. *Access consciousness* is a functional notion: a state is access-conscious when its content is poised for use in reasoning and in the rational control of action and speech (Block 1995). Block's title named the

confusion he diagnosed: arguments that attribute a function to phenomenal consciousness often establish only a function of access, because the two usually go together and the evidence cannot separate them.

Blindsight is the standard case at the boundary. After damage to primary visual cortex, some patients can discriminate stimuli presented in the blind part of the visual field, when forced to guess, well above chance, while denying that they see anything there (Weiskrantz 1986). Information about the stimulus is available to some consumers (the guessing response) and not to others (report, spontaneous use). Block described a hypothetical patient, a "superblindsighter," trained to prompt themselves to guess and to use the guesses in reasoning, who would have access without phenomenality, and argued that the case is conceivable even if no actual patient is like that (Block 1995).

In this book's terms, access consciousness is a property of a functional perspective: it concerns which readers inside the system can use a variable, and how widely. The global workspace theory of Chapter 44 develops it into a neural model. Phenomenal consciousness is the property that Definition 43.2 names, and Block's point is that the two definitions pick out different things even if, in every actual case, they are found together.

43.3 The Hard Problem and the Explanatory Gap

Chalmers divided the problems of consciousness into two kinds (Chalmers 1995; Chalmers 1996b). The *easy problems* are problems of explaining functions: how a system discriminates and categorizes stimuli and reacts to them, integrates information, reports its mental states, accesses its own internal states, focuses attention, controls its behavior deliberately, and differs between waking and sleep. They are easy in principle, not in practice; each may take decades, but each asks for a mechanism that performs a function, and the methods of cognitive science and neuroscience are built for that. The *hard problem* is to explain why the performance of these functions is accompanied by experience at all.

The chapters of this book have been working through easy problems. Reading a relation, computing its sign, acting to change it, maintaining oneself, misrepresenting, predicting, centering variables on oneself, and modeling one's own attention are all functions. The hard problem asks why any of them, or all of them together, should be experienced.

Three arguments make the problem vivid.

Levine observed that some scientific identities leave nothing to explain. Once heat is identified with the motion of molecules, and the motion is shown to produce the effects by which we recognized heat, the question "but why is molecular motion heat?" has no remaining content. An identity such as "pain is the firing of certain fibers" leaves the question "but why does that firing feel like this?" with its content intact. Levine called the difference an *explanatory*

gap, and argued that it concerns our explanations whether or not it concerns the world (Levine 1983).

Jackson imagined Mary, a scientist who knows every physical fact about color vision but has lived all her life in a black-and-white room. When she first sees a red object, it seems that she learns something new, namely what seeing red is like; if so, the physical facts she already knew were not all the facts (Jackson 1982).

Chalmers argued that a being physically and functionally identical to a conscious human but without experience, a "zombie," is conceivable, that what is conceivable in this strong sense is possible, and that experience is therefore not entailed by the physical and functional facts (Chalmers 1996b).

Each argument has been answered repeatedly, and no answer is accepted by everyone. Their common premise can be stated as a thesis about explanation.

Principle 43.4 (The explanatory gap, as a thesis)**.** For a complete functional description $\mathcal{F}$ of a system S, the question "does S have a phenomenal perspective?" is not answered by $\mathcal{F}$ in the way that "is this sample water?" is answered by a complete chemical description of the sample.

The thesis is contested, and the environment marks it as a position under discussion, not a result. Those who accept it disagree about what follows: an epistemic gap in our concepts (Levine 1983), or a gap in the world (Chalmers 1996b). Those who reject it hold either that the question is answered by $\mathcal{F}$ once $\mathcal{F}$ is complete enough, or that the question rests on a mistaken picture of what experience is.

43.4 Deflating the Gap

The most direct rejection is Dennett's. In Dennett (1991) he argued that consciousness is a set of functions, including the functions that produce our convictions about consciousness, and that once those functions are explained nothing is left over. His method, *heterophenomenology*, treats subjects' reports as data about what they believe their experience is like, to be explained like any other data, without assuming either that the beliefs are true or that they are false (Dennett 2003).

Frankish's illusionism sharpens the view. Our introspective representations of our experiences represent them as having phenomenal properties, a qualitative and ineffable character, and on the illusionist account nothing has such properties. The hard problem is replaced by the *illusion problem*: explain why we represent our states this way (Frankish 2016). Dennett called illusionism the obvious default theory (Dennett 2016). The attention schema theory of Chapter 42 supplies one mechanism for the illusion: a simplified model of attention whose content is a nonphysical-seeming inner grasp.

Illusionism turns the gap into an easy problem, at a price. It must hold that the phenomenal perspective of Definition 43.2 is empty, or rather that the phrase refers to nothing beyond what Definition 43.1 and a suitable self-misrepresentation describe. Its critics reply that an illusion is itself something experienced, so that the view presupposes what it denies; its defenders reply that "seeming" can be analyzed as representing, and that a misrepresentation need not be experienced in any further sense. The exchange has not been settled, and nothing in this book settles it.

Between denial and addition lies the program of the theories in Chapter 44. Each names a property (global broadcast, integrated information, recurrence, higher-order representation, a transparent self-model) and proposes that systems with that property have a phenomenal perspective. Whether any of them explains why, or only where and when, is the question that chapter asks of each.

43.5 What Would Count as Closing the Gap?

Four routes have been proposed, and they can be distinguished by what they would require.

Derivation. A theory would close the gap if the phenomenal facts could be derived from the physical and functional facts, as the properties of water are derived from its chemistry. The most promising version targets the *structure* of experience: if the similarity relations among experiences (among colors, say) could be derived from relations among neural states, the structure of experience would be explained. Even a complete correspondence of structure would not yet derive that the structure is experienced.

Prediction. A theory would earn confidence if it predicted, in advance and against rivals, which systems are conscious and what their experience is like, including cases no rival predicts, and if independent evidence then confirmed the predictions. This is the route of Chapters 44 and 61.

Intervention. A theory would earn confidence if manipulating the property it names changed experience in the predicted way while other functions were held fixed. The difficulty is the qualification: manipulations of the brain that change experience usually change function too.

Dissolution. The illusionist route would close the gap by showing that the question it poses is confused.

The second and third routes are the empirical ones, and both meet the same obstacle. Every measure of consciousness is validated against report somewhere.

Model 43.5 (The calibration problem)**.** Let $C(S) \in \{0, 1\}$ indicate whether system S has a phenomenal perspective, $K(S) \in \{0, 1\}$ whether it has the capacity to report, and $M(S)$ a candidate neural measure. Validation data consist of systems in which report is available and taken as evidence of C.

Consider two hypotheses about the measure: H_C, that M depends on C, and H_K, that M depends on the capacity for report K.

Proposition 43.6 (Report-bound validation)**.** *In any validation population in which $C(S) = K(S)$ for every system, the data assign equal likelihood to H_C and H_K. Distinguishing them requires systems in which C and K come apart, and the presence of C in such systems can be established only by delayed report, by an independent measure already validated against report, or by theory.*

Proof. If $C(S) = K(S)$ for every S in the data, any likelihood $\mathbb{P}(M \mid C)$ under H_C equals $\mathbb{P}(M \mid K)$ under H_K with the same functional form, so no observation of M in this population favors one hypothesis. A population with $C(S) \neq K(S)$ for some S is needed, and the value of C for such S must come from a source other than present report. □

The proposition is elementary and its consequence is not. The perturbational measures of Chapter 45 are among the most carefully validated in the field, and their validation uses exactly the loophole the proposition allows: volunteers under ketamine are unresponsive, yet on emergence they report long, vivid dreams, and the complexity of their cortical responses during unresponsiveness resembles wakefulness (Sarasso et al. 2015). The report is delayed, but it is still a report. Lamme proposed that neural measures be put on an equal footing with behavioral ones, so that a neural criterion could in principle overrule a report (Lamme 2006). That proposal is a decision to let a theory adjudicate where report cannot, and it makes the theory's credentials the whole question.

43.6 The Word "For"

The book can now say precisely where its own residue lies.

A functional perspective is a set of relations: variables indexed to the system ("world relative to my action," "before and after now," "this body is mine"), read by components of the system and used by it. Every item in Definition 43.1 is a relation that some reader inside the system reads.

Nagel's phrase contains a further relation: something it is like *for* the organism. The phenomenal perspective seems to be a relation between the organism and its own states, the relation of those states being present to it. Two readings of that relation are possible, and they divide the field.

On the first reading, "for" names one more agent-centered relation of the kind the book has studied, perhaps the relation between a first-order state and a representation of it (Chapter 40), or between the world-model and a transparent self-model (Chapter 41), or between attention and an attention schema (Chapter 42). Then some reader inside the system reads it, and the question becomes which reader, and why its reading is experienced when the

readings of the fish's torus are, perhaps, not. On the second reading, "for" names something that no reader supplies: an intrinsic property of the system, as in integrated information theory (Chapter 44), which denies that experience requires a reader at all.

This is the book's residue in its sharpest form. Everything below the word "for" has been subtracted by existing sciences, chapter by chapter. The word itself has not. Chapters 63 and 64 return to it with the whole ladder in view.

43.7 Audit

Canonical phenomenon

Blindsight: discrimination of stimuli in a cortically blind field, well above chance, with denial of seeing them (Weiskrantz 1986).

Minimal model

Two predicates of a system, a functional perspective defined by agent-centered variables that are read, used, and integrated (Definition 43.1), and a phenomenal perspective defined by Nagel's formula (Definition 43.2); the calibration model relating experience, report capacity, and a neural measure (Model 43.5).

What is relational here?

Agent-centered variables, which relate the world to the system's own position, time, body, and action; and the "for" of Nagel's formula, which relates the system to its own states.

What physically reads it?

For access: the consumers of a variable inside the system, which in blindsight include a guessing response and exclude report. For phenomenality: unknown, and whether any reader is involved is itself the disputed question.

What existing field explains it?

The philosophy of mind from Nagel (1974) and Levine (1983) through Jackson (1982), Block (1995), Chalmers (1995) and Chalmers (1996b), Dennett (1991), and Frankish (2016); consciousness science's methods for relating measures to reports.

What remains unexplained?

The hard problem in its entirety. The book subtracts the easy problems one by one and does not subtract this one.

A discriminating experiment

No experiment tests the hard problem directly. The discriminating experiment targets the calibration step that every empirical route depends on: validate a candidate measure in reporting subjects, then apply it in

conditions designed to separate the capacity for report from delayed report of experience (anesthetics that produce dreaming versus dreamless unresponsiveness, sleep with and without dream reports) and ask whether the measure tracks the later reports or the present capacity to respond (Proposition 43.6).

43.8 Notes and References

Nagel (1974) introduced the "something it is like" formula and argued that objective description leaves out the subjective character of experience; Nagel (1986) develops the theme. Levine (1983) named the explanatory gap. Jackson (1982) gave the knowledge argument. Block (1995) distinguished access from phenomenal consciousness and diagnosed the confusion between them; blindsight is reviewed in Weiskrantz (1986). Chalmers (1995) separated the easy problems from the hard problem, and Chalmers (1996b) gives the conceivability argument in full. The deflationary side is Dennett (1991), with the method of Dennett (2003); illusionism is stated by Frankish (2016) and endorsed by Dennett (2016). On putting neural measures on an equal footing with report, see Lamme (2006). The perturbational evidence that exploits delayed report is in Sarasso et al. (2015), discussed in Chapter 45.

43.9 Exercises

Exercise 43.1. Classify each of the following as an easy problem or as the hard problem in Chalmers's sense, and justify each classification in one sentence: explaining the jamming avoidance response; explaining why a patient with blindsight denies seeing; explaining why seeing red differs from hearing a tone; explaining the rubber hand illusion; explaining why there is anything it is like to undergo the rubber hand illusion.

Exercise 43.2. Score a thermostat, *Eigenmannia*, a mormyrid, a laboratory rat, and yourself on the three clauses of Definition 43.1, using a scale of your own design. Which clause does most of the work in separating the cases, and what evidence from earlier chapters supports each score?

Exercise 43.3. In Model 43.5, suppose that in a validation population $C = K$ for 95% of subjects and that the remaining 5% have $C = 1$, $K = 0$, established by delayed report. If a measure M is a noisy indicator of C under H_C and of K under H_K, with the same noise, compute the expected log-likelihood ratio between the hypotheses per subject. How many subjects are needed to reach a likelihood ratio of 100?

Exercise 43.4. Reconstruct Jackson's knowledge argument as a numbered sequence of premises and a conclusion. Identify the premise that an illusionist

denies and the premise that a defender of an a posteriori identity denies, and state each denial in one sentence.

Exercise 43.5. Specify Block's superblindsighter as a reading loop: what is the mixture, what reader extracts the stimulus variable, which consumers receive it, and what is missing. Could the missing component be supplied by adding a reader? What would a proponent of Principle 43.4 say about your answer?

Exercise 43.6. State the "for" of Nagel's formula in the notation of this book in two ways: once as a relation read by some reader inside the system, and once as an intrinsic property that requires no reader. For each version, name one theory from Chapter 44 that adopts it and one observation that would count against it.

44

Major Theories of Consciousness

Chapter 43 separated a functional perspective, which the book knows how to build, from a phenomenal perspective, which it does not. Theories of consciousness are proposals for crossing that gap. Each names a property of a physical or functional system and claims that systems with the property are conscious, or that the property is what consciousness is. This chapter compares the major theories by asking each the same five questions.

The thesis is threefold. The theories name different properties, and several of the properties are relational in the sense of this book: broadcast, integration, recurrence, and higher-order representation are all relations among parts of a system. Most of the evidence offered for each theory is evidence that several theories predict, and so confirms none of them over the others. And where two of them, the global neuronal workspace and integrated information theories, were made to predict divergent outcomes in advance and tested together, in the adversarial collaboration reported in 2025, each failed key predictions (Cogitate Consortium et al. 2025). No theory has shown why its property entails experience rather than accompanying it.

44.1 Five Questions for Every Theory

A comparison needs a fixed format. We use five questions, applied to each theory in turn.

Definition 44.1 (Theory profile)**.** The *profile* of a theory of consciousness T is its answer to five questions:

1. *Property*: what physical or functional property P_T matters?
2. *Link*: why is P_T supposed to entail, constitute, or produce consciousness?
3. *Prediction*: what does T predict that its rivals do not?
4. *Falsifier*: what observation would count against T?
5. *Perspective*: does T explain why experience is *for* a subject, in the sense of the closing section of Chapter 43?

The third and fourth questions carry the empirical weight, for a reason that can be stated exactly.

Proposition 44.2 (Only divergent predictions confirm)**.** *Let T_1 and T_2 be rival theories and D an observation. The posterior odds after observing D are the prior odds multiplied by the Bayes factor:*

$$\frac{\mathbb{P}(T_1 \mid D)}{\mathbb{P}(T_2 \mid D)} = \frac{\mathbb{P}(T_1)}{\mathbb{P}(T_2)} \cdot \frac{\mathbb{P}(D \mid T_1)}{\mathbb{P}(D \mid T_2)}. \tag{44.1}$$

If both theories predict D with equal probability, observing D leaves their relative credibility unchanged.

Proof. Apply Bayes's rule to each theory and divide; $\mathbb{P}(D)$ cancels. When $\mathbb{P}(D \mid T_1) = \mathbb{P}(D \mid T_2)$ the Bayes factor is one. □

The proposition is elementary, and it disposes of a large part of the evidence in the field. That some neural quantity changes when consciousness is lost under anesthesia is predicted by every theory on the list below. A theory earns credit only where it sticks its neck out.

Example 44.3 (An observation that confirms nothing). Under propofol anesthesia the complexity of the cortical response to magnetic stimulation falls, and it recovers with consciousness (Casali et al. 2013; Sarasso et al. 2015). A proponent of integrated information theory reads the fall as a loss of integration and differentiation. A proponent of the global workspace reads it as a loss of the long-range recurrent interactions that ignition requires. A proponent of a higher-order theory reads it as the loss of the prefrontal states that represent first-order states. All three predicted the fall. The observation is important for clinical practice (Chapter 45) and leaves the relative credibility of the three theories where it was.

44.2 Global Neuronal Workspace

Baars proposed that consciousness is a limited-capacity workspace whose contents are broadcast to many specialized, unconscious processors (Baars 1988). Dehaene, Kerszberg, and Changeux gave the idea a neuronal form: a population of workspace neurons with long-range axons, densely distributed in prefrontal, parieto-temporal, and cingulate cortices, which can link otherwise separate processors (Dehaene, Kerszberg, et al. 1998; Dehaene and Changeux 2011). In the mature version, conscious access occurs when a representation triggers a nonlinear *ignition* of the workspace network, which, with recurrent processing, amplifies and sustains the representation and makes it globally available (Dehaene and Changeux 2011; Mashour et al. 2020).

Property. Global availability of a content through ignition and broadcast.
Link. On the theory, being conscious of a content just is that content's

being globally available to the system's consumers: reasoning, memory, report, and the control of action. *Prediction.* Conscious access should show late amplification of sensory activity, long-distance synchronization in the beta and gamma bands, and ignition of a prefronto-parietal network (Dehaene and Changeux 2011); conscious contents should be represented in prefrontal cortex. *Falsifier.* Conscious experience of a content with no workspace involvement, in particular no prefrontal representation of the content. *Perspective.* The workspace supplies a single broadcast shared by many consumers, which is a functional unity; the theory does not claim to explain why the broadcast is experienced for anyone.

In the vocabulary of this book, the global workspace is a theory of *readers*. A content becomes conscious when many readers inside the system can read it.

44.3 Integrated Information Theory

Integrated information theory starts at the other end. Tononi proposed that consciousness corresponds to a system's capacity to integrate information, measured by a quantity Φ (Tononi 2004). Later versions begin from axioms taken to be self-evident properties of experience (it exists intrinsically, is structured, specific, unified, and definite) and derive postulates that a physical substrate must satisfy (Oizumi et al. 2014; Albantakis et al. 2023). The substrate of consciousness is proposed to be a maximum of intrinsic cause-effect power: the set of elements, in a state, whose causal structure is maximally irreducible to the structures of its parts (Tononi, Boly, et al. 2016). The quantity of experience is given by the irreducibility, and its quality by the full structure of causal distinctions and relations that the substrate specifies (Albantakis et al. 2023).

Property. A maximally irreducible cause-effect structure. *Link.* An identity: the experience is the cause-effect structure, as it exists for the system itself. *Prediction.* The substrate of perceptual experience lies in a posterior cortical "hot zone" (Koch, Massimini, et al. 2016; Tononi, Boly, et al. 2016); purely feedforward systems have $\Phi = 0$ and are not conscious whatever they do (Oizumi et al. 2014); a system's potential states, including units that are inactive but could be active, contribute to what it experiences (Tononi, Boly, et al. 2016). *Falsifier.* Consciousness in a system whose cause-effect structure is reducible, or its absence where the structure is maximally irreducible; in practice Φ cannot be computed for brains, so tests rely on predictions about where and when activity should be sustained and connected. *Perspective.* Integrated information theory is the one theory whose axioms include the intrinsic, for-the-system character of experience; it claims to explain perspective by building it in.

In the vocabulary of this book, integrated information theory is the one major theory that requires no reader. The cause-effect structure exists intrinsi-

cally, whether or not any process inside or outside the system reads it. The contrast with the global workspace could hardly be sharper, and Section 44.11 shows what happened when the two were tested against each other.

44.4 Recurrent Processing

Lamme and Roelfsema distinguished two modes of cortical processing. A fast feedforward sweep routes activity through many visual areas and groups features that are hardwired, but is probably incapable of yielding awareness; recurrent processing, through horizontal connections and feedback from higher areas, is needed before the features of an object are grouped and the stimulus can enter consciousness (Lamme and Roelfsema 2000). Lamme later argued that local recurrent processing within sensory cortex suffices for phenomenal experience, while widespread recurrence that engages frontoparietal regions underlies access and report, and that neural criteria should be given the same standing as behavioral ones in defining consciousness (Lamme 2006).

Property. Recurrent interaction between feedforward and feedback activity in sensory cortex. *Link.* Recurrence binds features into an integrated representation, and that representation is the experience. *Prediction.* Phenomenal experience can occur without report or access, wherever local recurrence occurs. *Falsifier.* Local recurrence reliably present without experience; the difficulty is that the theory's key cases are, by design, cases without report. *Perspective.* No specific account.

44.5 Higher-Order Theories

Higher-order theories hold that a mental state is conscious when the subject is, in a suitable way, aware of being in it. Armstrong and Lycan developed versions on which the awareness is a kind of inner perception (D. M. Armstrong 1968; Lycan 1996); Rosenthal's higher-order thought theory holds that a state is conscious when accompanied by a suitable thought that one is in that state (D. M. Rosenthal 2005); Carruthers developed a dispositional variant (Carruthers 2000). Lau and Rosenthal reviewed evidence that separates higher-order from first-order views, including dissociations between task performance and awareness, and defended a role for prefrontal cortex (Lau and D. Rosenthal 2011). Brown, Lau, and LeDoux argue that the theory has been misread: higher-order states are assembled cognitively but require less than critics suppose, and the theory does not equate consciousness with metacognition or confidence (R. Brown et al. 2019).

Property. A higher-order representation of a first-order state as one's own. *Link.* To be conscious of a state is to be aware of being in it; a state of which one is in no way aware is not conscious. *Prediction.* Awareness can be lost while first-order discrimination is preserved (blindsight is the standard

case (Weiskrantz 1986)); manipulations of prefrontal function should change awareness without changing performance. *Falsifier.* Preserved awareness when higher-order representation is abolished; whether prefrontal damage has this effect is contested, with reviews reaching opposite conclusions from overlapping evidence (Boly et al. 2017; Odegaard et al. 2017). *Perspective.* Yes, in a specific sense: the higher-order state represents the subject as being in the state, which gives experience its for-me character (Chapter 40).

44.6 Predictive Processing Accounts

Predictive processing treats the brain as a hierarchical generative model that predicts its inputs and updates itself on precision-weighted prediction errors; active inference extends the account to action (K. Friston 2010; Clark 2013; Hohwy 2013). It is a framework for perception and action more than a theory of consciousness, and Seth and Bayne group its consciousness-related versions with re-entry theories (Seth and Bayne 2022). A prominent version ties the experience of emotion and of being a self to predictions about the body's internal state, interoceptive inference (Seth 2013).

Property. The contents of a hierarchical generative model's best current hypotheses, weighted by precision. *Link.* Conscious contents are the brain's best guesses about the causes of its inputs; selfhood is the best guess about the body that does the guessing. *Prediction.* Expectations should shape conscious perception, and interoceptive predictions should shape emotion and self-experience. *Falsifier.* As a framework it can accommodate many outcomes; specific consciousness theories built on it must add commitments that can fail (Seth and Bayne 2022). *Perspective.* The generative model is centered on the organism that holds it, so the account supplies the structure of a perspective (Chapter 33).

44.7 Attention Schema, Self-Model, and Projective Theories

Three theories locate consciousness in a model that a system has of itself.

The attention schema theory of Chapter 42 holds that awareness is the content of a simplified model of attention, used to control attention and to attribute awareness to others (Graziano 2013; Graziano and Webb 2015). Its distinctive prediction concerns control: without awareness, attention should still be possible but its control should degrade (Webb et al. 2016).

The self-model theory of Chapter 41 holds that the experience of being someone is the content of a transparent, globally integrated self-model (Metzinger 2003). Its evidence comes from experiments that shift self-location and body ownership by manipulating multisensory relations (Lenggenhager et al. 2007; Blanke and Metzinger 2009).

The Projective Consciousness Model proposes that the spatial field of consciousness is structured by a projective geometry and controlled by active inference, which selects points of view according to preferences; its authors apply it to the spatial character of experience, the dynamics of bistable figures, and body-swap illusions in virtual reality (Rudrauf et al. 2017), and to pre-reflective self-consciousness and the first-person point of view (Williford et al. 2018).

All three explain the *structure* of perspective in detail: a point of view with an origin, a body, a location, and an attentional focus. None claims to derive from that structure that it is experienced; the attention schema theory denies that there is anything further to derive.

44.8 Merker's Egocentric Integration

Merker locates the origin of consciousness in a problem that mobile animals face: to separate the effects of their own movement from changes in the world, and to decide what to do next in a stable arena (Merker 2005). Consciousness, on this view, presents a stable world from which the multiple sensorimotor transformations that construct it have been removed, and it arose early in the evolution of mobile animals with centralized brains. Merker proposed that an upper brainstem system, extending from the roof of the midbrain to the basal diencephalon, integrates target selection, action selection, and motivation into a single egocentric simulation, and cited the purposive behavior of decorticated mammals and the evidence that children born without a cortex are conscious (Merker 2007). Both lines of evidence are contested, and the proposal was published together with open peer commentary (Merker 2007).

Property. Integration of target selection, action selection, and motivation in a centralized egocentric simulation. *Link.* The simulation is the arena of experience, a reality space centered on the animal. *Prediction.* Consciousness should survive extensive loss of cortex and should be widespread among vertebrates. *Falsifier.* Evidence that cortical integrity is necessary for any consciousness at all. *Perspective.* Yes: an egocentric reality space is a perspective by construction. Merker's account is the one most directly continuous with this book's fish, whose midbrain computes the relations it acts on.

44.9 Illusionism

Illusionism, described in Chapter 43, holds that introspection represents experiences as having phenomenal properties that nothing has (Frankish 2016; Dennett 2016). It is a theory of consciousness only in a qualified sense: it proposes that the target of the other theories is an introspective representation, and that the task is to explain the representation.

Property. A systematic introspective misrepresentation of one's own states. *Link.* There is no further property to link; the conviction that there is one is what needs explaining. *Prediction.* Reports of phenomenal character should be explainable by the mechanisms that produce them, including their errors. *Falsifier.* None that experiments currently reach; the dispute is conceptual. *Perspective.* The represented perspective is real as a representation; a phenomenal perspective beyond it is denied.

44.10 The Profiles Side by Side

Table 44.1 compresses the profiles. It is a summary of the positions as their proponents state them, with our assessment of what each is most directly challenged by; the sections above give the sources.

Table 44.1: Profiles of the major theories of consciousness (Definition 44.1).

Theory	**Property**	**Distinctive prediction**	**Most directly challenged by**	**Perspective**
Global workspace	Ignition and global broadcast	Prefrontal representation of contents; ignition at onset and offset	Conscious content without prefrontal representation or offset ignition	Unity of broadcast; not a subject
Integrated information	Maximally irreducible cause-effect structure	Posterior substrate; sustained connected activity for the duration of experience; feedforward systems unconscious	Absence of sustained posterior synchronization during experience	Intrinsic; built into the axioms
Recurrent processing	Local recurrence in sensory cortex	Experience without report wherever recurrence occurs	Recurrence reliably present without experience	None specific
Higher-order	Representation of a state as one's own	Awareness lost with discrimination preserved; prefrontal role	Preserved awareness without higher-order representation (disputed)	For-me character via self-attribution
Predictive processing	Precision-weighted best hypotheses	Expectation shapes experience; interoception shapes self	Few; as a framework it accommodates many outcomes	Organism-centered generative model
Attention schema	Simplified model of attention	Degraded attentional control without awareness	Awareness removed with control unchanged	Self-attributed attention
Self-model	Transparent, integrated self-model	Self-experience shifts with multisensory relations	Self-experience dissociating from the self-model	World-model centered on self-model

Theory	Property	Distinctive prediction	Most directly challenged by	Perspective
Projective model	Projective geometry under active inference	Perspectival phenomena as projective transformations	Failure to predict the targeted phenomena	Geometry of a point of view
Egocentric integration	Brainstem integration of selection and motivation	Consciousness without cortex; widespread in vertebrates	Cortex necessary for any consciousness	Egocentric reality space
Illusionism	Introspective misrepresentation	Reports explainable with their errors	No experiment reaches the dispute	Represented, not phenomenal

Two patterns stand out. First, most of the properties are relational. Broadcast is a relation between a content and its consumers; irreducibility is a relation among the causal powers of parts and wholes; recurrence is a relation between feedforward and feedback activity; higher-order representation is a relation between a state and a representation of it. The theories differ on *which* relation matters and on whether it must be read. Second, the theories that explain perspective in detail (self-model, projective, egocentric integration, higher-order) explain it as a structure that some process in the system constructs and uses, which places them on the functional side of the line drawn in Chapter 43; the one theory that claims to capture the intrinsic, for-the-system character of experience directly, integrated information theory, does so by making it an axiom.

44.11 Adversarial Testing

The theories developed largely in parallel, each accumulating its own supporting evidence. Melloni and colleagues organized an adversarial collaboration between proponents of the global neuronal workspace and integrated information theories. The proponents agreed on contrasting predictions about the location and timing of the neural correlates of visual consciousness, and on how the outcomes would be interpreted; the predictions and analyses were preregistered; and the experiments were run by theory-impartial laboratories with functional magnetic resonance imaging, magnetoencephalography and electroencephalography, and intracranial recordings, with built-in replications (Melloni et al. 2023).

The results were reported by the Cogitate Consortium in 2025 (Cogitate Consortium et al. 2025). Two hundred and fifty-six participants viewed suprathreshold stimuli for variable durations. Information about the conscious content was found in visual, ventrotemporal, and inferior frontal cortex; sustained responses in occipital and lateral temporal cortex reflected stimulus duration; and content-specific synchronization was found between frontal and

early visual areas. The consortium reported that the results align with some predictions of each theory while substantially challenging key tenets of both. For integrated information theory, the lack of sustained synchronization within posterior cortex contradicted the claim that network connectivity specifies consciousness. For the global workspace theory, the general lack of ignition at stimulus offset, and the limited representation of certain conscious dimensions in prefrontal cortex, were the main challenges. The consortium noted that these challenges extend to other theories that share the tested predictions (Cogitate Consortium et al. 2025).

Three lessons follow, and none of them depends on which theory one favors. The first is methodological: the collaboration shows that theories of consciousness can be made to commit to divergent predictions in advance, which is the condition of Proposition 44.2. The second is that each theory needed auxiliary assumptions to generate predictions at all (for integrated information theory, that irreducible causal structure should show up as sustained synchronization measurable with current methods; for the global workspace, that offset ignition should be detectable), so a failed prediction can be absorbed by revising an auxiliary, as Lakatos described for research programmes generally (Lakatos 1970). The third is that the experiment tested where and when content is represented and sustained, which is a question about correlates; it could not test, and was not designed to test, whether either property is what experience is. Chapter 61 develops the method; Chapter 60 proposes experiments in the same spirit for relational measures.

44.12 What Each Theory Says About the Fish

The book's hydrogen atom provides a small test of the theories' reach. *Eigenmannia* reads a relation, computes its sign in a midbrain circuit that pools over the body surface, and changes its own frequency in a closed loop (Part V). The global workspace theory, whose workspace is a network of long-range cortical neurons, gives no clear verdict for an animal without a neocortex, and the verdict depends on whether some other structure can play the workspace's role. Integrated information theory gives a verdict in principle, since the fish's circuits have some cause-effect structure, but no one can compute it. Higher-order theories would ask whether the fish represents its own states as its own; nothing in the jamming avoidance response requires that it does. Merker's account, which locates the core of consciousness in a conserved upper brainstem system, is the most hospitable to attributing some consciousness to vertebrates generally (Merker 2007). The theories disagree about the book's best-understood reader, and that disagreement is itself informative: a complete functional account of a system does not, on any current theory, settle whether the system is conscious without further, theory-specific premises. Chapter 50 takes up the question in comparative context.

44.13 Audit

Canonical phenomenon

Conscious perception of suprathreshold visual stimuli of variable duration, as tested in the adversarial collaboration (Melloni et al. 2023; Cogitate Consortium et al. 2025).

Minimal model

The theory profile (Definition 44.1) and the Bayes-factor rule that only divergent predictions confirm (Proposition 44.2).

What is relational here?

Most candidate properties: broadcast (content and consumers), irreducible cause-effect structure (parts and whole), recurrence (feedforward and feedback), higher-order representation (a state and its representation), and synchronization (phase relations between areas).

What physically reads it?

Workspace consumers for the global workspace; higher-order states for higher-order theories; an attention schema or self-model for the model theories; an introspective mechanism for illusionism. Integrated information theory posits no reader.

What existing field explains it?

The theories themselves, reviewed and compared by Seth and Bayne (2022); the method of adversarial collaboration (Melloni et al. 2023).

What remains unexplained?

For every theory, the link from its property to experience. The evidence locates correlates; no theory derives experience from its property, and the theories that explain the structure of perspective best explain it functionally.

A discriminating experiment

Extend the adversarial method to the relational claims of this book: have proponents of two theories that disagree about whether consciousness requires a reader (for example, integrated information theory and a higher-order theory) preregister predictions for a case in which a relational structure is present but no plausible reader reads it, such as a cortical region that remains causally connected while its outputs are pharmacologically disconnected from downstream areas.

44.14 Notes and References

Seth and Bayne (2022) is the standard recent comparison of theories. For the global workspace: Baars (1988), Dehaene, Kerszberg, et al. (1998), Dehaene and Changeux (2011), and Mashour et al. (2020). For integrated information

theory: Tononi (2004), Oizumi et al. (2014), Tononi, Boly, et al. (2016), and Albantakis et al. (2023); the posterior hot zone is argued in Koch, Massimini, et al. (2016). For recurrent processing: Lamme and Roelfsema (2000) and Lamme (2006). For higher-order theories: D. M. Armstrong (1968), Lycan (1996), Carruthers (2000), D. M. Rosenthal (2005), Lau and D. Rosenthal (2011), and R. Brown et al. (2019); the front-versus-back debate is set out in the paired reviews of Boly et al. (2017) and Odegaard et al. (2017). For predictive processing: K. Friston (2010), Clark (2013), Hohwy (2013), and Seth (2013). For the model theories: Graziano (2013), Metzinger (2003), Rudrauf et al. (2017), and Williford et al. (2018). For egocentric integration: Merker (2005) and Merker (2007). For illusionism: Frankish (2016) and Dennett (2016). The adversarial collaboration's protocol is Melloni et al. (2023) and its results Cogitate Consortium et al. (2025); on how research programmes absorb failed predictions, see Lakatos (1970). An early and influential neurobiological proposal, which the exercises examine, is Crick and Koch (1990).

44.15 Exercises

Exercise 44.1. Crick and Koch (1990) proposed that synchronized oscillations in the gamma range bind the features of an attended object and are a correlate of visual awareness. Write the profile of this proposal (Definition 44.1). Which of the other theories in Table 44.1 predict synchronization, and what observation would separate the proposal from them?

Exercise 44.2. Suppose that theories T_1 and T_2 assign probabilities 0.9 and 0.6 to an observation D and 0.1 and 0.4 to its negation. Starting from even prior odds, compute the posterior odds after observing D three times independently, and after observing D twice and its negation once. How many observations of D would raise the odds to 10 in favor of T_1?

Exercise 44.3. Integrated information theory assigns $\Phi = 0$ to purely feedforward systems. Construct a feedforward network and a recurrent network with identical input–output behavior on a finite set of inputs. Which theories in Table 44.1 would treat the two networks differently, and on what grounds?

Exercise 44.4. For one of the two tested theories, identify an auxiliary assumption that connects its property to the measurements of Cogitate Consortium et al. (2025). Propose a revision of the auxiliary that would absorb the reported challenge, and state a new prediction that the revised theory makes and the original did not.

Exercise 44.5. Write a one-paragraph prediction, for each of four theories of your choice, about whether *Eigenmannia* is conscious, stating which empirical facts about the fish each prediction depends on. Which facts would you most want to know, and how could they be measured?

Exercise 44.6. Classify each theory in Table 44.1 by whether it requires a reader of the property it names, and if so whether the reader is inside the system. Design an observation that would favor reader-based theories over integrated information theory, or the reverse.

45

Neural Integration and Differentiation

When consciousness fades in deep sleep, under general anesthesia, in a generalized seizure, or after severe brain injury, the brain does not fall silent. Its activity changes in structure. Across these conditions, the level of consciousness covaries with the brain's capacity to support activity that is at once *integrated*, in the sense that its parts depend on one another, and *differentiated*, in the sense that the whole can take many distinct states.

The thesis of this chapter is that this covariation is among the best-established findings in consciousness science, that its best-validated measure, the perturbational complexity index, can classify states and patients with a reliability no earlier measure achieved, and that none of this establishes that integration and differentiation constitute experience. The measures are relational; they are read by experimenters, not by the brains that carry them; and they are validated against report, which limits what they can show (Chapter 43).

45.1 Segregation and Integration

Cortical areas are functionally specialized, and during perception and action they work together. Tononi, Sporns, and Edelman proposed a measure of the interplay between the two properties, *neural complexity*, built from information theory (Tononi, Sporns, and Edelman 1994).

Definition 45.1 (Integration and neural complexity)**.** For a system $X = (x_1, \ldots, x_n)$ of random variables, the *integration* is the total deviation from statistical independence,

$$I(X) = \sum_{i=1}^{n} H(x_i) - H(X). \tag{45.1}$$

For a subset X_j^k of k variables, $I(X_j^k)$ is defined in the same way. The *neural complexity* of X is

$$C_N(X) = \sum_{k=1}^{n} \Big[\tfrac{k}{n}\, I(X) - \big\langle I(X_j^k) \big\rangle_j \Big], \tag{45.2}$$

where $\langle \cdot \rangle_j$ averages over all subsets of size k.

Each term compares the integration of subsets of size k with the share of the whole system's integration that a subset of that size would carry if integration were spread evenly. Tononi and colleagues showed that C_N is high when segregation of small subsets coexists with integration of large ones, low when the components are either completely independent or completely dependent, and high for connection patterns resembling those of cerebral cortex: dense, locally clustered, patchy, and reciprocal (Tononi, Sporns, and Edelman 1994). They later proposed the *dynamic core* hypothesis, that the substrate of consciousness is a cluster of neural activity that is both highly integrated and highly differentiated from moment to moment (Tononi 1998). The idea is the ancestor of integrated information theory (Chapter 44).

Integration is a relational quantity in exactly the sense of Chapter 1: it is defined on the joint distribution and cannot be computed from the variables one at a time. Appendix H shows that it is one of a family of measures of multivariate dependence, and that the family's members disagree in instructive ways.

Integration and segregation also meet in dynamics. Coordination dynamics describes a brain whose regions express their own tendencies while coupling transiently into coalitions, a regime called metastability (Tognoli and J. A. S. Kelso 2014), and Chapter 11 developed its mathematics. Metastability is the dynamical counterpart of high neural complexity: a system that never fully synchronizes and never fully fragments.

45.2 Criticality

A second family of proposals holds that the waking brain operates near a critical point, between a regime in which activity dies out and one in which it runs away.

Beggs and Plenz recorded spontaneous local field potentials from cultured and acute slices of rat cortex on a 60-electrode array and found cascades of activity, *neuronal avalanches*, whose sizes followed a power law with exponent $-3/2$, with a branching parameter close to one (Beggs and Plenz 2003). The branching parameter is a ratio.

Definition 45.2 (Branching ratio)**.** In a cascade of activity, the *branching ratio* σ is the mean number of units activated in the next time step per unit active in the current one.

Proposition 45.3 (Critical branching)**.** *In a Galton–Watson branching process with mean offspring number σ and finite offspring variance, the expected total size of a cascade started by one unit is $1/(1-\sigma)$ for $\sigma < 1$ and is infinite for $\sigma \geq 1$. At $\sigma = 1$, with finite and nonzero offspring variance, the probability of a cascade of total size s falls as $s^{-3/2}$ for large s.*

Sketch. The expected number of units in generation g is σ^g, and $\sum_{g \geq 0} \sigma^g = 1/(1-\sigma)$ for $\sigma < 1$. The $s^{-3/2}$ tail at criticality is a classical result for the total progeny of a critical branching process, obtained from the Lagrange inversion of the offspring generating function. □

The proposition is a result about an idealized process, and its relevance to cortex is a hypothesis. It shows why $\sigma = 1$ is special: it is the magnitude balance of Chapter 5, a dimensionless ratio at which two tendencies, growth and decay, exactly offset. Beggs and Plenz argued from simulations that a branching parameter at this value optimizes information transmission while preventing runaway excitation (Beggs and Plenz 2003), and Chialvo reviewed the case for criticality as an organizing principle of brain dynamics (Chialvo 2010).

The evidence has a known weakness. Touboul and Destexhe showed that event-size distributions in local field potentials can look close to power laws on logarithmic axes without any underlying criticality, that similar apparent scaling appears in surrogate signals, and that the appearance does not survive more rigorous statistical tests; thresholded stochastic processes produce it generically (Touboul and Destexhe 2010). A power law on a log–log plot is therefore not by itself evidence of criticality.

With that caution in place, the relevance to consciousness is a specific finding. Tagliazucchi and colleagues studied functional magnetic resonance imaging during wakefulness, propofol sedation, loss of consciousness, and recovery. During unconsciousness, activity in frontothalamic regions showed reduced long-range temporal correlations, and functional connectivity departed from anatomical constraints; a model of a system passing through a phase transition reproduced these findings and the reduced sensitivity of cortex to external perturbation (Tagliazucchi et al. 2016). The authors interpret loss of consciousness as a departure from critical dynamics.

45.3 Perturbational Complexity

Spontaneous activity can be complex for reasons that have nothing to do with integration: independent noise sources produce high entropy. Massimini and colleagues asked a sharper question: when one cortical area is perturbed, how does the perturbation spread? They stimulated premotor cortex with transcranial magnetic stimulation and recorded the response with high-density electroencephalography. During quiet wakefulness, an initial response at the

stimulation site was followed by a sequence of waves that moved to connected cortical areas several centimeters away. During non-rapid-eye-movement sleep, the initial response was stronger but was rapidly extinguished and did not propagate beyond the stimulation site (Massimini et al. 2005). They proposed that the fading of consciousness in this stage of sleep is related to a breakdown in cortical effective connectivity.

Casali and colleagues turned the observation into a measure.

Model 45.4 (Perturbational complexity, schematic)**.** Perturb the cortex with a magnetic pulse and estimate, at each cortical source i and time t after the pulse, whether the evoked response is significantly different from baseline, giving a binary matrix $\boldsymbol{S} = (S_{it})$. The *perturbational complexity index* is the algorithmic complexity of $\boldsymbol{S}$, estimated by Lempel–Ziv compression, normalized so that it can be compared across recordings with different amounts of significant activity:

$$\mathrm{PCI} = \frac{c_{\mathrm{LZ}}(\boldsymbol{S})}{c_{\max}(\boldsymbol{S})}, \tag{45.3}$$

where c_{LZ} is the Lempel–Ziv complexity and $c_{\max}$ the value expected for a random matrix of the same size and source entropy. The exact normalization is given in Casali et al. (2013); the schematic form suffices here.

The measure has two ingredients that correspond to the two properties. The perturbation engages distributed interactions, so a response that stays local, as in sleep, yields a small matrix of significant activity; that is the integration side. The compression rewards responses that are spatially and temporally diverse, so a response that spreads as a single stereotyped wave compresses well and scores low; that is the differentiation side (Casali et al. 2013).

Casali and colleagues tested the index on a large set of recordings from healthy subjects in wakefulness, dreaming, non-rapid-eye-movement sleep, and sedation with midazolam, xenon, and propofol, and from patients who had emerged from coma into a vegetative state, a minimally conscious state, or locked-in syndrome. The index discriminated the level of consciousness in single individuals across these conditions (Casali et al. 2013). Casarotto and colleagues then validated a cutoff in a benchmark population of 150 healthy controls and communicative brain-injured patients, who could confirm the presence or absence of consciousness by report, in states of conscious wakefulness, disconnected consciousness, and unconsciousness. The cutoff separated conscious from unconscious conditions with 100% sensitivity and specificity in the benchmark. Applied to patients who could not report, it detected the minimally conscious state with 94.7% sensitivity and identified 9 of 43 patients in a vegetative state whose values overlapped the distribution of the benchmark conscious condition (Casarotto et al. 2016).

The anesthetic comparison is the most informative case.

Example 45.5 (Three anesthetics, two ways to fail). Sarasso and colleagues measured the index in 18 volunteers under propofol, xenon, and ketamine, all of which abolish behavioral responsiveness. Propofol produced a low-amplitude slow wave with a local, low-complexity pattern; xenon produced a high-amplitude slow wave with a global but stereotyped, low-complexity pattern; ketamine produced a complex, wakefulness-like pattern. On emergence, participants reported no conscious experience after propofol and xenon, and long, vivid dreams after ketamine (Sarasso et al. 2015). The two low-complexity cases fail in opposite ways, one by staying local and one by spreading without variety, which is exactly the distinction between the two ingredients of the measure.

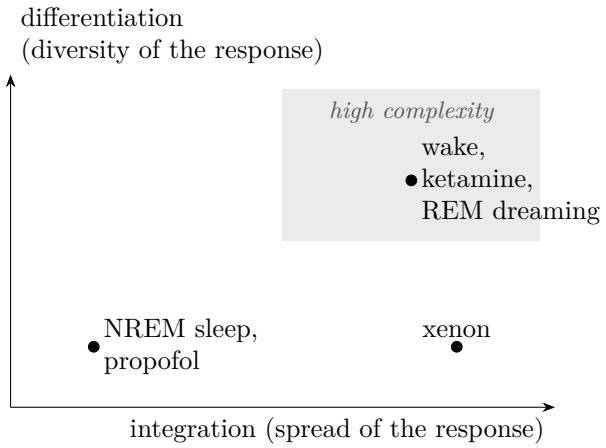


Figure 45.1: Schematic placement of conditions on the two ingredients of perturbational complexity, after Massimini et al. (2005), Casali et al. (2013), and Sarasso et al. (2015). Positions are qualitative.

45.4 Anesthesia, Sleep, Coma, and Seizures

The same picture recurs across the conditions in which consciousness is lost. Brown, Lydic, and Schiff compared general anesthesia with sleep and coma, and described general anesthesia as a reversible, drug-induced coma whose neurophysiology can be tracked and whose mechanisms differ by drug (E. N. Brown et al. 2010). Alkire, Hudetz, and Tononi reviewed the evidence that anesthetics produce unconsciousness when they disrupt the integration of information: either by functionally disconnecting a posterior complex of cortical regions, interrupting cortical communication, or by producing bistable, stereotyped responses that reduce the repertoire of available states (Alkire et al. 2008). They also noted that some anesthetics abolish responsiveness without necessarily abolishing consciousness, which is the lesson of the ketamine data above.

Epileptic seizures impair consciousness through different mechanisms that converge on a common set of cortical and subcortical structures. Blumenfeld reviewed absence, generalized tonic–clonic, and complex partial seizures, and argued that they disrupt a specialized system that normally maintains alertness, attention, and awareness (Blumenfeld 2012).

Disorders of consciousness after brain injury add a clinical dimension. The minimally conscious state was defined by consensus as a condition with inconsistent but clearly discernible behavioral evidence of consciousness, distinguishable from coma and the vegetative state (Giacino et al. 2002). Owen and colleagues showed that a patient who met the behavioral criteria for the vegetative state, when asked to imagine playing tennis or moving through her home, activated the predicted cortical areas in a manner indistinguishable from healthy volunteers (Owen et al. 2006). Behavior underestimated consciousness in that patient. Perturbational complexity, which requires neither sensory processing nor behavior, was built for exactly this problem, and the nine vegetative-state patients with high values in Casarotto et al. (2016) are its clinical payoff and its open question.

45.5 Why None of This Establishes Phenomenality

The covariation is robust. Four reasons prevent it from settling what experience is.

Calibration. Every measure in this chapter was validated against report, immediate or delayed. The ketamine result depends on reports given after emergence; the benchmark of Casarotto et al. (2016) depends on subjects who could confirm consciousness by report. The calibration argument of Chapter 43 shows why this limits what the data can distinguish: a measure that tracks the capacity for experience and one that tracks the capacity for later report agree wherever the two coincide.

Level and content. The measures track the *level* of consciousness, whether a subject is conscious at all. They say nothing about *what* is experienced, which is where the theories of Chapter 44 make their most distinctive claims. The debate over whether the contents of experience depend on posterior or on prefrontal cortex is conducted with different evidence, and it remains open (Koch, Massimini, et al. 2016; Boly et al. 2017; Odegaard et al. 2017).

Arousal. A measure can track arousal rather than experience. Lendner and colleagues showed that the slope of the aperiodic component of the electrophysiological power spectrum, a single non-oscillatory quantity, separates wakefulness from propofol anesthesia, non-rapid-eye-movement sleep, and rapid-eye-movement sleep (Lendner et al. 2020). Rapid-eye-movement sleep is the stage most often associated with dreaming (Lendner et al. 2020). A measure that places it with the unconscious states is tracking something other than the

presence of experience, and every measure in this chapter must be checked against the same confound.

Underdetermination. The same observations support different theories. Integrated information theory reads falling perturbational complexity as falling integrated information; the global workspace reads it as the loss of the long-range recurrent interactions that ignition needs (Chapter 44). The measure was designed from the first theory and is compatible with the second.

A fifth point concerns the book's own vocabulary. Integration, differentiation, the branching ratio, and perturbational complexity are relational variables. In every case discussed here, the reader that extracts them is the experimenter: a compression algorithm, a statistical estimator, a model fit. Whether any process inside the brain reads its own complexity, and uses it, is a question the measures do not address. Integrated information theory answers that no reader is needed; theories built on readers must say which reader, if any, uses the integration that the experimenter measures.

45.6 Audit

Canonical phenomenon

The cortical response to a magnetic pulse spreads and diversifies in wakefulness, dreaming, and ketamine anesthesia, stays local in non-rapid-eye-movement sleep and propofol anesthesia, and spreads without diversity under xenon (Massimini et al. 2005; Casali et al. 2013; Sarasso et al. 2015).

Minimal model

Integration and neural complexity (Definition 45.1); the branching ratio and critical branching (Proposition 45.3); perturbational complexity (Model 45.4).

What is relational here?

Statistical dependence among parts (integration), diversity of joint states (differentiation), the ratio of descendants to ancestors in a cascade (branching), and the relation between a perturbation and the spatiotemporal pattern it evokes.

What physically reads it?

The experimenter: significance tests, compression algorithms, and model fits. No internal reader of these variables has been identified.

What existing field explains it?

Information theory (Tononi, Sporns, and Edelman 1994), the physics of branching processes and criticality (Beggs and Plenz 2003; Chialvo 2010), clinical neurophysiology of anesthesia, sleep, coma, and epilepsy (E. N. Brown et al. 2010; Alkire et al. 2008; Blumenfeld 2012; Giacino et al. 2002), and the theories of Chapter 44.

What remains unexplained?

Whether integration and differentiation constitute experience or enable it, and which cortical regions carry the contents of experience.

A discriminating experiment

Compare perturbational complexity with matched non-relational measures on the same recordings: the amplitude of the evoked response, the aperiodic spectral slope, and band power. For each, compute its accuracy at separating conscious from unconscious conditions as confirmed by report, including rapid-eye-movement sleep with dream reports and ketamine anesthesia. A relational measure earns its keep only if it succeeds where the matched constituent measures fail (Chapters 60 and 62).

45.7 Notes and References

Neural complexity is introduced by Tononi, Sporns, and Edelman (1994) and the dynamic core hypothesis by Tononi (1998); metastability is reviewed by Tognoli and J. A. S. Kelso (2014). Neuronal avalanches are reported by Beggs and Plenz (2003), the criticality program is reviewed by Chialvo (2010), and the statistical caution is Touboul and Destexhe (2010); the application to anesthesia is Tagliazucchi et al. (2016). The perturbational approach begins with Massimini et al. (2005); the index is defined and tested in Casali et al. (2013), validated against report in Casarotto et al. (2016), and applied to anesthetics with different phenomenology in Sarasso et al. (2015). General anesthesia, sleep, and coma are compared by E. N. Brown et al. (2010); anesthesia and integration by Alkire et al. (2008); impaired consciousness in epilepsy by Blumenfeld (2012). The minimally conscious state is defined in Giacino et al. (2002), and covert awareness in the vegetative state is shown by Owen et al. (2006). The aperiodic slope as a marker of arousal is Lendner et al. (2020). The posterior hot zone is argued by Koch, Massimini, et al. (2016), and the front-versus-back debate by Boly et al. (2017) and Odegaard et al. (2017).

45.8 Exercises

Exercise 45.1. Compute the integration (45.1) of three binary variables in three cases: independent fair bits; three copies of one fair bit; and X, Y independent fair bits with $Z = X \oplus Y$. Which case has the largest integration? Compute the average integration of pairs in each case and comment on what the comparison says about pairwise versus higher-order dependence (Chapter 1).

Exercise 45.2. Simulate a branching process with Poisson offspring of mean $\sigma \in \{0.8, 0.95, 1.0\}$, starting each cascade from one unit and capping cascades at 10^4 units. Plot the distribution of cascade sizes on log–log axes and estimate

the exponent at $\sigma = 1$. Then threshold a sequence of independent Gaussian noise samples at a high level and plot the distribution of run lengths above threshold; explain why a straight segment on log–log axes is not enough to establish criticality (Touboul and Destexhe 2010).

Exercise 45.3. Write a Lempel–Ziv complexity function for binary matrices read in column order. Compute it for a matrix with activity only in one row, for a matrix in which every row carries the same pulse, and for a matrix with independent random entries of the same density. Relate the three cases to the three anesthetics in Figure 45.1.

Exercise 45.4. In the benchmark of Casarotto et al. (2016), conscious and unconscious conditions were confirmed by report. Describe a population in which a perfect classifier of the benchmark could still fail, and say what additional data would reveal the failure.

Exercise 45.5. Using Lendner et al. (2020), explain why a measure that separates wakefulness from rapid-eye-movement sleep cannot be a pure measure of the presence of experience. Propose a design that would separate a measure of arousal from a measure of experience using dream reports collected on awakening.

Exercise 45.6. For each measure in this chapter, identify the reader that extracts it. Choose one and describe what a biological reader of that variable inside the brain would have to do, which structures could in principle perform it, and what evidence would show that it does.

46

Neural Frequency Relations

The book began with beats: two nearby frequencies, an envelope that belongs to neither, and a difference frequency that becomes physical only when something nonlinear receives the sum (Chapters 3 and 4). The brain is full of rhythms, and the rhythms stand in relations of exactly these kinds. The phase of a slow rhythm modulates the amplitude of a fast one; rhythms lock to periodic stimulation inside tongue-shaped regions of a parameter plane; two flickering inputs produce responses at sums and differences of their frequencies; spectral components show phase coupling that the power spectrum cannot see. All of these relations change with the state of consciousness.

The thesis of this chapter is that a relational neural measure earns a place in the science of consciousness only when it outperforms the simpler measures it is built from, and that the history of these measures supplies both successes and instructive failures. The most celebrated clinical relational measure, the bispectral component of a commercial anesthesia monitor, turned out to add little to ordinary spectral power. The most convincing relational measures are the ones designed with a control that destroys the relation while sparing its relata.

46.1 Oscillations and Their Background

Cortical networks oscillate over several orders of magnitude of frequency, and the oscillations are conserved across mammals, which suggests that they are functionally relevant (Buzsáki and Draguhn 2004; Buzsáki 2006). The classical bands (delta, theta, alpha, beta, gamma) are defined by frequency ranges, and much of the literature reports power in those ranges.

Two cautions precede any relational analysis. First, electrophysiological power spectra contain an aperiodic, roughly $1/f$-like component in addition to oscillatory peaks, and the aperiodic component changes with age, task, and state. Standard band-power analyses can conflate a change in the aperiodic component with a change in oscillatory power, and algorithms now exist to separate them (Donoghue et al. 2020). Second, neural oscillations are often not sinusoidal. Their waveforms can be sharp, asymmetric, or arch-shaped,

and a nonsinusoidal waveform has harmonics that spectral methods report as separate components (Cole and Voytek 2017).

Both cautions matter here because relational measures inherit them. A relation computed between two bands is only as meaningful as the claim that each band contains an oscillation.

Example 46.1 (An arousal marker without a relation). Lendner and colleagues analyzed intracranial and scalp recordings from four independent data sets and showed that the slope of the aperiodic component alone separates wakefulness from propofol anesthesia, non-rapid-eye-movement sleep, and rapid-eye-movement sleep (Lendner et al. 2020). The slope is a property of one signal's spectrum, not a relation between rhythms. Any relational measure proposed as a marker of conscious state must beat this simple competitor, and must also explain why it would do better than a quantity that, by grouping dreaming sleep with the unconscious states, evidently tracks arousal rather than experience.

46.2 Cross-Frequency Coupling

Canolty and colleagues recorded electrocorticograms from human patients and found that the phase of the theta rhythm (4 to 8 Hz) modulates power in the high gamma band (80 to 150 Hz), with stronger modulation at higher theta amplitudes and different patterns across the cortex for different tasks (Canolty, Edwards, et al. 2006). Phase–amplitude coupling has since been reported widely and proposed as a mechanism that links large-scale slow rhythms with fast local processing (Canolty and Knight 2010).

Definition 46.2 (Modulation index)**.** Let $\phi_L(t)$ be the phase of a low-frequency band and $A_H(t)$ the amplitude envelope of a high-frequency band. Divide the phase circle into N bins, let $\bar{A}_j$ be the mean of A_H when ϕ_L falls in bin j, and form the distribution $P(j) = \bar{A}_j / \sum_k \bar{A}_k$. The *modulation index* is the normalized divergence of P from the uniform distribution,

$$\mathrm{MI} = \frac{\log N - H(P)}{\log N}, \qquad H(P) = -\sum_{j=1}^{N} P(j) \log P(j), \tag{46.1}$$

which is zero when the high-frequency amplitude is independent of the low-frequency phase and one when all amplitude falls in a single bin (Tort et al. 2010).

The index is a relational observable in the sense of Chapter 1: it is defined on the joint behavior of two bands and on neither alone. Tort and colleagues reviewed seven other coupling measures and stressed the use of surrogate controls (Tort et al. 2010).

The measure also has a reader problem, and it is the one Part I predicts.

Example 46.3 (Coupling created by the measurement). Consider a rhythm with a sharp, nonsinusoidal waveform, such as a train of brief peaks at 8 Hz. Its Fourier series contains the fundamental and many harmonics, all phase-locked to the fundamental by construction. Band-pass filtering around a high harmonic yields an "amplitude" that rises at every peak, and so at a fixed phase of the filtered fundamental. The modulation index reports strong coupling between the two bands. Nothing couples two oscillators here: there is one oscillation, and the "relation" is created by the decomposition that the analysis imposed (Cole and Voytek 2017; Aru et al. 2015).

The example is the brain-sized version of the lesson of Chapter 16: which relational variables appear depends on the reader, and here the reader is the analysis pipeline. Aru and colleagues catalogued further confounds, including spurious coupling produced by common nonstationarities in the absence of any interaction between frequency components, and gave recommendations for avoiding them (Aru et al. 2015). Coupling measures remain useful, but only with controls that show the coupling survives when the waveform and the nonstationarity are accounted for (Appendices J and K).

46.3 Frequency Ratios

If rhythms interact, the ratios of their frequencies matter, for the reasons developed in Chapter 7. Two proposals about the brain's ratios illustrate how easily a relational claim outruns its evidence.

Pletzer, Kerschbaum, and Klimesch argued that the classical frequency bands form a geometric series with a ratio of neighboring center frequencies equal to the golden mean, $1.618\ldots$, and that the excitatory phases of two oscillations whose frequencies stand in that ratio never coincide in the mathematical sense and coincide rarely and irregularly in the physiological sense; the golden mean, as the "most irrational" number, would then provide a maximally desynchronized resting state from which coupling can arise (Pletzer et al. 2010). Klimesch later argued for a different architecture: a binary hierarchy in which neighboring center frequencies stand in a ratio of two, with brain and body oscillations aligned and communicating through phase-to-envelope and harmonic phase-to-phase coupling (Klimesch 2018).

The two proposals come from overlapping authors and assign different ratios to the same bands, which already shows that band definitions do not fix the ratio. There is a deeper point, which the theory of Chapter 7 makes precise. Whether two coupled oscillators lock depends on their coupling strength and their detuning, and the locking regions (Arnold tongues) surround every rational ratio, widening with coupling (Pikovsky et al. 2001; Glass and Mackey 1988). An irrational ratio protects oscillators from locking only when coupling is weak enough that they sit outside all the tongues that matter, and in that case the protection is supplied by the weak coupling as much as by the number. The

golden mean is special in number theory because every term of its continued fraction is 1, which makes it as badly approximable by rationals as any number can be, so it stays as far as possible from the rationals that anchor the tongues (Appendix F); but real neural rhythms are narrowband noisy processes whose center frequencies drift across individuals and states, and a statement about the exact irrationality of their ratio has no measurable content. The relational variable that matters is the pair (coupling, detuning), not the arithmetic of the ratio.

46.4 Entrainment and Arnold Tongues

Periodic stimulation gives direct access to the tongue picture.

Proposition 46.4 (Locking in the Adler equation)**.** *Let φ be the phase difference between an oscillator with natural frequency ω_0 and a periodic drive of frequency ω, and suppose that to first order in the coupling $\varepsilon > 0$ it obeys*

$$\dot{\varphi} = \Delta\omega - \varepsilon \sin \varphi, \qquad \Delta\omega = \omega_0 - \omega. \tag{46.2}$$

The oscillator locks to the drive (the phase difference has a stable fixed point) if and only if $|\Delta\omega| \le \varepsilon$. In the plane of detuning and coupling, the locking region is the tongue $|\Delta\omega| \le \varepsilon$.

Proof. Fixed points satisfy $\sin \varphi^* = \Delta\omega/\varepsilon$, which has solutions if and only if $|\Delta\omega| \le \varepsilon$. The fixed point with $\cos \varphi^* > 0$ is stable, since the derivative of the right-hand side there is $-\varepsilon \cos \varphi^* < 0$. If $|\Delta\omega| > \varepsilon$, $\dot{\varphi}$ never vanishes and the phase difference drifts. □

The Adler equation is the first-order phase reduction of a weakly driven limit-cycle oscillator (Appendices D and F), and its prediction is qualitative and strong: locking should be easier near the natural frequency and at higher stimulation intensity, and the region of locking should widen with intensity.

Three experiments bear on the prediction. Herrmann presented ten subjects with flickering light at every frequency from 1 to 100 Hz in steps of 1 Hz; steady-state responses appeared at all frequencies up to at least 90 Hz, with resonance-like enhancement around 10, 20, 40, and 80 Hz (Herrmann 2001). Thut and colleagues delivered bursts of transcranial magnetic stimulation at the preferred frequency of a parietal alpha generator identified with magnetoencephalography, and confirmed three predictions of entrainment: alpha oscillations were induced, they grew progressively over the train, and the effect depended on the phase of the ongoing alpha rhythm before stimulation; control burst profiles and a phantom head showed that the effect required synchronization between the train and the neural oscillator (Thut et al. 2011). Notbohm, Kurths, and Herrmann compared rhythmic with jittered visual stimulation across frequencies and intensities and found phase coupling that increased with intensity and was

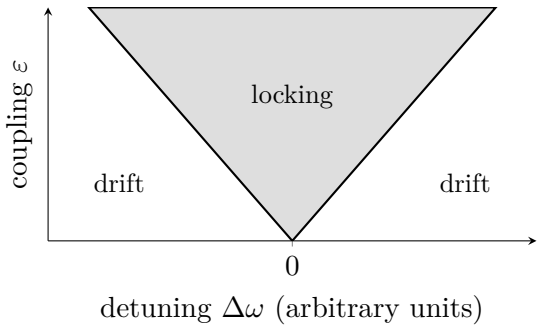


Figure 46.1: The first-order Arnold tongue of the Adler equation (46.2). Entrainment experiments test whether neural rhythms respond to periodic stimulation with this structure: stronger locking near the intrinsic frequency and at higher intensity.

greater at frequencies closer to each participant's intrinsic frequency, significant only inside the Arnold tongue, together with intermittent phase locking that contradicts a linear superposition of evoked responses (Notbohm et al. 2016). Herrmann and colleagues review the theory of how periodic stimulation shapes intrinsic rhythms (Herrmann et al. 2016).

These experiments pass the test this chapter sets. The jittered stimulus delivers the same stimuli with the relation to the ongoing rhythm destroyed, and the dependence on pre-stimulus phase shows that the response depends on the relation between drive and oscillator, not on the drive alone. Whether the width of these tongues tracks the state of consciousness is an open question, taken up as an experiment in Chapter 60.

46.5 Intermodulation

When two inputs at frequencies f_1 and f_2 drive a system, a response at a combination $nf_1 + mf_2$ with both n and m nonzero is an *intermodulation* component. Regan and Regan developed intermodulation analysis as a general technique for characterizing nonlinearities in biological systems (M. Regan and D. Regan 1988). Its diagnostic value rests on a simple fact.

Proposition 46.5 (Intermodulation requires a common nonlinearity)**.** *Let inputs $x_1(t) = \cos(2\pi f_1 t)$ and $x_2(t) = \cos(2\pi f_2 t)$, with f_1/f_2 irrational, pass through separate time-invariant channels g_1 and g_2, possibly nonlinear, whose outputs are added: $y = g_1(x_1) + g_2(x_2)$. Then the spectrum of y contains only integer multiples of f_1 and integer multiples of f_2, and no intermodulation components. If instead the inputs meet in one nonlinearity, $y = g(x_1 + x_2)$ with g not affine, intermodulation components are generically present.*

Proof. $g_1(x_1(t))$ is periodic with period $1/f_1$, so its Fourier series contains only multiples of f_1; the same holds for $g_2(x_2(t))$ with f_2, and addition does

not create new frequencies. For a common nonlinearity, expanding g in a power series, the quadratic term alone contributes $2x_1x_2 = \cos(2\pi(f_1+f_2)t) + \cos(2\pi(f_1-f_2)t)$, and higher terms contribute further combinations; the combinations cancel only for special g. □

The proposition is the neural version of the square-law detector of Chapter 4. An intermodulation component in a brain response shows that the two inputs were combined by a single nonlinear reader somewhere between the stimulus and the recording.

Boremanse, Norcia, and Rossion applied the idea to faces. Participants viewed the left and right halves of a face flickering at different frequencies, 5.88 and 7.14 Hz. Besides responses at the two fundamentals, which reflect processing of the parts, the electroencephalogram showed intermodulation components (for example at $7.14 - 5.88 = 1.26$ Hz) over the right occipito-temporal cortex. The part responses did not depend on whether the halves were aligned, separated, upright, or inverted. The intermodulation components were virtually absent when the halves were misaligned and separated, and inverting the whole face reduced them specifically (Boremanse et al. 2013). Steady-state methods of this kind are reviewed by Norcia et al. (2015).

The design is the model for this chapter. The misaligned condition keeps both relata (the same two half-faces at the same two frequencies) and destroys their relation as a face; the measure that tracks the relation changes while the measures that track the relata do not. That is the intervention criterion of Chapter 1, carried out in a human brain.

46.6 Bispectrum, Bicoherence, and the Bispectral Index

The power spectrum discards phase. Quadratic phase coupling, in which the phase of a component at $f_1 + f_2$ is consistently the sum of the phases at f_1 and f_2, is a signature of a quadratic nonlinearity, and the bispectrum detects it (Nikias and Mendel 1993; Sigl and Chamoun 1994).

Definition 46.6 (Bispectrum and bicoherence)**.** For a signal with Fourier coefficients $X(f)$ estimated over many epochs, the *bispectrum* is

$$B(f_1, f_2) = \mathbb{E}\big[X(f_1)\, X(f_2)\, X^{*}(f_1 + f_2)\big], \tag{46.3}$$

and the *bicoherence* is its normalized magnitude,

$$b^2(f_1, f_2) = \frac{|B(f_1, f_2)|^2}{\mathbb{E}\big[|X(f_1)X(f_2)|^2\big]\; \mathbb{E}\big[|X(f_1+f_2)|^2\big]}. \tag{46.4}$$

If the phases of the three components vary independently across epochs, B averages toward zero; if the third phase is locked to the sum of the other two, B does not.

Sigl and Chamoun introduced bispectral analysis for the electroencephalogram and the concept of a bispectral index, arguing that the power spectrum ignores the phase coupling expected from nonlinear sources such as the central nervous system (Sigl and Chamoun 1994). The commercial bispectral index monitor combines several features of the electroencephalogram, including a bispectral parameter, into a proprietary number used to guide the depth of anesthesia (Rampil 1998).

The relational ingredient was tested directly. Miller and colleagues recorded the electroencephalogram during induction of anesthesia in 39 patients and compared the monitor's bispectral parameter (SynchFastSlow) with an analogous parameter computed from the power spectrum (PowerFastSlow). The two were correlated ($r = 0.84$) and neither was superior at predicting the awake or anesthetized state (areas under the receiver operating characteristic curve of 0.85 and 0.93, respectively). An amplitude-independent version computed from the bicoherence did not change over induction and did not correlate with the bispectral parameter ($r = 0.07$). The authors concluded that they could not show that bispectral analysis gave more information than power spectral analysis, and that most of the change in the bispectral values resulted from the fall in the relative high-frequency content of the electroencephalogram (Miller et al. 2004).

The clinical record is mixed in a separate way. In the B-Aware trial of 2463 adults at high risk of awareness, monitor-guided anesthesia was associated with two reports of awareness against eleven with routine care, a risk reduction of 82% (95% confidence interval 17 to 98%) (Myles et al. 2004). The B-Unaware trial randomized 2000 high-risk patients to a monitor-guided protocol or to one guided by end-tidal anesthetic gas concentration and found two cases of definite awareness in each group; awareness occurred even with values within the target ranges, and the authors concluded that their findings did not support routine monitoring as standard practice (Avidan et al. 2008). Neither trial tested the bispectral component separately, and neither needed to: the lesson for this book is that a relational feature was built into a clinical index on theoretical grounds, and the evidence that it contributed beyond ordinary spectral power did not materialize.

46.7 Signatures of Propofol Unconsciousness

The best-characterized electroencephalographic signatures of loss of consciousness come from propofol. Purdon and colleagues recorded high-density electroencephalograms during gradual induction and emergence while subjects responded to interleaved verbal and click stimuli. Loss of consciousness was marked simultaneously by an increase in power below 1 Hz, the loss of spatially coherent occipital alpha oscillations, and the appearance of spatially coherent frontal alpha oscillations, and these reversed on recovery. The phase

of the slow oscillation modulated alpha amplitude in two distinct patterns: during profound unconsciousness, alpha amplitude was maximal at the peaks of the slow oscillation; during the transitions into and out of unconsciousness, it was maximal at the troughs, and this second pattern predicted recovery (Purdon et al. 2013). Mukamel and colleagues, with electroencephalography and electrocorticography, found a rapid drug-induced transition between two states with opposite phase–amplitude coupling and different cortical source distributions (Mukamel et al. 2014).

Here a relational measure, the phase relation between two rhythms, distinguishes states that power alone would blur, and the result is specific in time and in space. Two questions keep it from settling anything about consciousness in general. The signatures are characterized for propofol, and whether they generalize to other anesthetics, to sleep, and to injury is a separate empirical matter; and the transitions were identified by responsiveness, so the calibration problem of Chapter 43 applies.

46.8 The Rule

The cases in this chapter can be summarized in a single requirement.

Principle 46.7 (Relational measures must beat their constituents)**.** A relational neural measure $R(x_1, x_2)$ earns explanatory credit as a marker or mechanism of conscious state only if

1. it predicts the state better than the best matched non-relational measures computed from the same signals (band power, aperiodic slope, evoked amplitude);
2. its value survives surrogate tests that destroy the relation while preserving the constituents' spectra (Appendix J) (Theiler et al. 1992; Lancaster et al. 2018);
3. it is not produced by waveform shape or nonstationarity of a single signal (Cole and Voytek 2017; Aru et al. 2015);
4. it generalizes across mechanisms of unconsciousness (sleep, different anesthetics, injury); and
5. where possible, it responds to an intervention on the relation that spares the relata, as the misaligned face halves and jittered stimulation do.

Table 46.1: Relational measures against Principle 46.7. The assessments summarize the evidence cited in this chapter.

Measure	Relation	Main confound	Status
Phase–amplitude coupling	Phase of a slow band and amplitude of a fast band	Nonsinusoidal waveforms; nonstationarity	Useful with controls; state-specific patterns under propofol
Frequency ratios	Ratios of band center frequencies	Band definitions; drifting, noisy rhythms	Competing proposals; no discriminating test
Entrainment (Arnold tongues)	Phase of drive and intrinsic rhythm	Evoked-response superposition	Passes jitter and phase-dependence controls; link to consciousness untested
Intermodulation	Joint nonlinear processing of two inputs	Separate channels summed	Passes a configural control that spares the parts
Bispectral component	Quadratic phase coupling	Change in relative high-frequency power	Not shown to beat a matched power measure

The requirement is demanding, and it should be. A relational measure that merely correlates with the state of consciousness has shown that something changes; almost everything does. A relational measure that beats its constituents, survives destruction of the relation, and follows interventions on it has shown that the relation carries information about the state that the relata do not.

46.9 What Relational Measures Have Not Yet Shown

Three things remain undemonstrated. No relational measure has been shown to track the presence of experience in a case where matched non-relational measures and perturbational complexity both fail. No relational measure has been shown to be causally necessary for consciousness, by an intervention that alters the relation alone and changes experience as reported. And no relational measure has been tied to a reader: the intermodulation components of the face experiment show that some neural population combines the two halves nonlinearly, and that population is a reader in the book's sense, but whether its reading contributes to the experience of the face, or only to its processing, is exactly the question Chapter 43 leaves open. Chapter 60 proposes experiments aimed at each gap.

46.10 Audit

Canonical phenomenon

Intermodulation components at $f_1 - f_2$ in the response to two half-faces flickering at different frequencies, present when the halves form a face and nearly absent when they are misaligned and separated (Boremanse et al. 2013).

Minimal model

The modulation index (Definition 46.2); the Adler equation and its tongue (Proposition 46.4); intermodulation as a signature of a common nonlinearity (Proposition 46.5); the bispectrum (Definition 46.6).

What is relational here?

Phase–amplitude relations between bands, ratios of frequencies, the phase relation between a drive and an intrinsic rhythm, joint nonlinear processing of two inputs, and phase coupling among three spectral components.

What physically reads it?

For intermodulation, a neural population that receives both inputs and combines them nonlinearly; for coupling and bispectral measures, often the analysis pipeline itself, which can create the relation it reports.

What existing field explains it?

Nonlinear systems analysis (M. Regan and D. Regan 1988), synchronization theory (Pikovsky et al. 2001), spectral and higher-order spectral analysis (Nikias and Mendel 1993), clinical neurophysiology of anesthesia (Purdon et al. 2013), and the methodological literature on coupling (Tort et al. 2010; Aru et al. 2015; Cole and Voytek 2017).

What remains unexplained?

Whether any relation among neural rhythms is necessary for experience, as opposed to accompanying the states in which experience occurs.

A discriminating experiment

Measure entrainment tongues (Figure 46.1) at several levels of propofol and of ketamine sedation, together with matched non-relational measures, and relate each to reports collected on emergence. If tongue width tracks the later reports across both drugs while power and aperiodic slope do not, the relation carries information the constituents lack; if power or slope does as well, the relational measure has not earned its keep.

46.11 Notes and References

Buzsáki and Draguhn (2004) and Buzsáki (2006) survey cortical rhythms. The aperiodic component is parameterized by Donoghue et al. (2020), waveform

shape is treated by Cole and Voytek (2017), and the aperiodic slope as a marker of arousal is Lendner et al. (2020). Phase–amplitude coupling is reported by Canolty, Edwards, et al. (2006), reviewed by Canolty and Knight (2010), measured by Tort et al. (2010), and criticized by Aru et al. (2015). The golden-mean and binary-hierarchy proposals are Pletzer et al. (2010) and Klimesch (2018); the dynamics of locking is in Pikovsky et al. (2001) and Glass and Mackey (1988). Entrainment evidence: Herrmann (2001), Thut et al. (2011), Notbohm et al. (2016), and the review Herrmann et al. (2016). Intermodulation: M. Regan and D. Regan (1988), Boremanse et al. (2013), and the steady-state review Norcia et al. (2015). Higher-order spectra: Nikias and Mendel (1993) and Sigl and Chamoun (1994); the bispectral index is described by Rampil (1998), its bispectral component tested by Miller et al. (2004), and its clinical use tested in Myles et al. (2004) and Avidan et al. (2008). Propofol signatures: Purdon et al. (2013) and Mukamel et al. (2014). On surrogate data: Theiler et al. (1992) and Lancaster et al. (2018). Communication-through-coherence proposals, which give phase relations between areas a mechanistic role in routing, are Fries (2005) and Fries (2015); F. Varela, Lachaux, et al. (2001) and Lachaux et al. (1999) introduced phase-synchrony measures for large-scale integration.

46.12 Exercises

Exercise 46.1. Generate a signal consisting of a periodic train of narrow Gaussian pulses at 8 Hz with added white noise. Compute the modulation index (46.1) between the 6–10 Hz phase and the 60–100 Hz amplitude. Repeat for a pure 8 Hz sinusoid plus independent 60–100 Hz noise, and for surrogates that randomize the Fourier phases of the pulse train. Explain the three results.

Exercise 46.2. Integrate (46.2) for $\varepsilon = 0.3$ and $\Delta\omega$ from -0.6 to 0.6, and plot the mean frequency of $\dot{\varphi}$ against $\Delta\omega$. Show that the plot has a plateau of width 2ε and that outside it the mean drift is $\sqrt{\Delta\omega^2 - \varepsilon^2}$ in absolute value.

Exercise 46.3. Pass $\cos(2\pi \cdot 5.88\,t) + \cos(2\pi \cdot 7.14\,t)$ through $g(x) = x + 0.3x^2 + 0.1x^3$ and list all output frequencies below 20 Hz with their amplitudes. Then pass each tone through the same g separately and add the outputs. Which frequencies disappear, and how does the comparison relate to the misaligned-face control of Boremanse et al. (2013)?

Exercise 46.4. Simulate epochs of $\cos(2\pi f_1 t + \alpha) + \cos(2\pi f_2 t + \beta) + \cos(2\pi(f_1 + f_2)t + \gamma)$ with α, β uniform random per epoch, and either $\gamma = \alpha + \beta$ or γ independent. Estimate the bicoherence (46.4) at (f_1, f_2) in both cases as a function of the number of epochs. How many epochs are needed to tell the cases apart with noise of the same power as each component?

Exercise 46.5. Choose a published claim that a relational neural measure tracks consciousness. Evaluate it against the five clauses of Principle 46.7, and design the single additional analysis that would most improve the claim.

Exercise 46.6. Two oscillators with frequencies f and ϕf, where ϕ is the golden mean, are coupled with strength K. Using the circle map of Appendix F, argue that for sufficiently large K they lock to a nearby rational ratio. What does this imply for the claim that golden-mean ratios guarantee desynchronization?

47

Neural Moiré Theories

This book began by asking whether consciousness is a moiré effect, and set the question aside as too loose to test. Neuroscience has twice proposed moiré and interference mechanisms for features of the mammalian cortex that are far better defined than consciousness: the hexagonal firing fields of entorhinal grid cells, and the maps of orientation preference in primary visual cortex. Both proposals were mathematically exact, both made specific predictions, and both met specific evidence against them.

The thesis of this chapter is the lesson of that history. A beautiful relational mechanism does not imply correct biology. What decides between a relational mechanism and its rivals is a prediction on which they differ, tested in the tissue; and a failed mechanism can leave the relational variable standing while replacing the physical process proposed to produce it. If moiré models of grid cells and orientation maps needed that discipline, a moiré theory of consciousness needs it more.

47.1 Grid Cells

Hafting, Fyhn, Molden, Moser, and Moser found that neurons in the dorsocaudal medial entorhinal cortex of the rat fire whenever the animal's position coincides with any vertex of a regular grid of equilateral triangles spanning the environment. Neighboring cells share the orientation and spacing of their grids and differ in the grids' offsets, or phases; spacing and field size increase from dorsal to ventral; and the grid is anchored to landmarks but persists in their absence, which suggested that grid cells belong to a path-integration system (Hafting et al. 2005). Later recordings from up to 186 grid cells in individual rats showed that the cells cluster into a small number of modules with distinct scale, orientation, asymmetry, and theta-frequency modulation (Stensola et al. 2012).

A grid is a relational object twice over. Its geometry is a lattice, a structure of relations among firing locations; and it must be computed from relations, since the animal has no sensor for its own position and must integrate its

velocity over time. Two families of mechanism were proposed, and one of them was a beat.

47.2 Oscillatory Interference

The interference idea began with the hippocampus. O'Keefe and Recce found that a place cell starts firing at a particular phase of the theta rhythm as the rat enters its place field and then fires progressively earlier on each theta cycle as the rat crosses the field. The phase precession ranged from about 100 to 355 degrees across cells, was better correlated with position than with time, and appeared to arise because the cells' interburst frequency was higher than the theta frequency (O'Keefe and Recce 1993). Two oscillations at slightly different frequencies slip against each other at the difference frequency: a beat, as in Chapter 3.

O'Keefe and Burgess proposed that the same interference could produce grid cells (O'Keefe and Burgess 2005), and Burgess, Barry, and O'Keefe developed the proposal into a two-dimensional model. Dendritic subunits of entorhinal stellate cells oscillate at a frequency that exceeds the theta frequency by an amount proportional to the animal's speed in a preferred direction; the phase difference between each dendritic oscillation and a somatic theta input then integrates velocity, so that its interference pattern reflects the distance traveled in that direction; and several such patterns, along directions separated by multiples of 60 degrees, multiply to give the grid (Burgess et al. 2007).

Model 47.1 (Path integration by a beat)**.** Let the animal move with velocity $\boldsymbol{v}(t)$, and let a baseline oscillation run at frequency f_θ. A velocity-controlled oscillator with preferred direction $\hat{\boldsymbol{d}}$ runs at

$$f_{\hat{\boldsymbol{d}}}(t) = f_\theta + \beta\, \boldsymbol{v}(t) \cdot \hat{\boldsymbol{d}}, \tag{47.1}$$

where β is a gain in cycles per unit distance. The interference of the two oscillations drives firing when their phase difference is near zero.

Proposition 47.2 (The beat phase is position)**.** *In Model 47.1, the phase difference between the velocity-controlled oscillator and the baseline is*

$$\Delta\phi(t) = 2\pi\beta\, (\boldsymbol{x}(t) - \boldsymbol{x}_0) \cdot \hat{\boldsymbol{d}}, \tag{47.2}$$

where $\boldsymbol{x}(t)$ is the animal's position. Firing locked to $\Delta\phi \equiv 0 \pmod{2\pi}$ therefore occurs on bands perpendicular to $\hat{\boldsymbol{d}}$, spaced $1/\beta$ apart. The product of three band patterns whose directions differ by 60 degrees has maxima on a triangular lattice with spacing $2/(\sqrt{3}\,\beta)$.

Proof. The phase difference accumulates at the frequency difference: $\Delta\phi(t) = 2\pi \int_0^t \beta\, \boldsymbol{v}(\tau) \cdot \hat{\boldsymbol{d}}\, \mathrm{d}\tau = 2\pi\beta(\boldsymbol{x}(t) - \boldsymbol{x}_0) \cdot \hat{\boldsymbol{d}}$. Each band pattern is a plane wave

with wave vector of magnitude $2\pi\beta$; three plane waves of equal magnitude at 60-degree separations have their joint maxima on a triangular lattice whose reciprocal vectors have that magnitude, which for a triangular lattice of spacing a is $4\pi/(\sqrt{3}a)$; setting $4\pi/(\sqrt{3}a) = 2\pi\beta$ gives $a = 2/(\sqrt{3}\beta)$. □

Example 47.3 (An illustrative grid). Take $\beta = 1/(40\,\text{cm})$, an illustrative gain. A rat running at 20 cm/s along the preferred direction raises the oscillator's frequency by $\beta v = 0.5\,\text{Hz}$ above theta, so the phase difference completes one cycle every two seconds, once per 40 cm of travel. The three-direction product then peaks on a triangular lattice with spacing $2/(\sqrt{3}\beta) \approx 46\,\text{cm}$. Running twice as fast doubles the frequency difference and halves the time per cycle, and the spacing in space is unchanged, which is the point: the relation integrates velocity into position.

The proposition is exact within the model, and the model is the cleanest relational mechanism in systems neuroscience. The variable the system needs, its position, is not a property of either oscillator; it is their phase difference, a beat phase, and velocity integration comes for free because a frequency difference integrates into a phase difference. In the book's terms, the mixture M is the convergence of two oscillations on one cell, the reader D is the cell's firing threshold applied to their sum, and the relational variable read is the animal's displacement along $\hat{\boldsymbol{d}}$. The electric fish of Part V uses a beat as a *signal* to be read; the interference model uses a beat as the *computation*.

47.3 Moiré Grids

Blair, Welday, and Zhang took the moiré idea one step further. They showed that grid fields can combine to form moiré interference patterns, "moiré grids," that replicate the hexagonal lattice over an unlimited range of spatial scales, and proposed that the grids recorded in the medial entorhinal cortex are moiré grids formed by interference between much smaller "theta grids," hypothesized to be the source of the movement-related theta rhythm (Blair et al. 2007). The formation of moiré grids from theta grids obeys two scaling rules, one for differences in length and one for differences in rotation, which the authors related to the properties of grid cells in different entorhinal layers, and they argued that memory representations built from moiré grids would inherit scale invariance (Blair et al. 2007).

The two rules are the moiré formulas of Chapter 2, applied to hexagonal lattices. Two lattices of spacing λ rotated by a small angle θ produce a moiré lattice of spacing

$$\Lambda_{\text{rot}} = \frac{\lambda}{2\sin(\theta/2)} \approx \frac{\lambda}{\theta}, \tag{47.3}$$

and two aligned lattices of slightly different spacings λ_1 and λ_2 produce one of spacing

$$\Lambda_{\text{len}} = \frac{\lambda_1 \lambda_2}{|\lambda_1 - \lambda_2|}. \tag{47.4}$$

Small mismatches amplify into large scales, which is why the proposal could produce large grids from small ones (Figure 47.1).

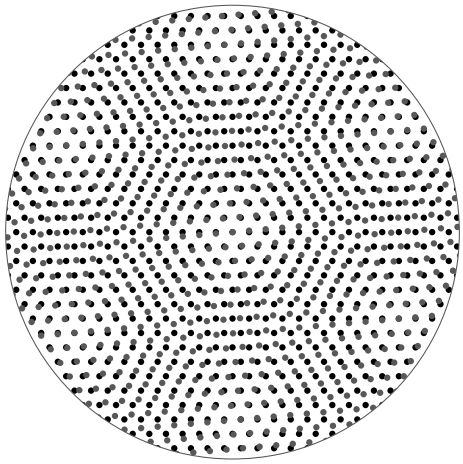


Figure 47.1: Two triangular lattices of equal spacing rotated by five degrees. The large hexagonal pattern of dense and sparse regions is a moiré lattice with spacing given by (47.3), about eleven times the spacing of either lattice. In the model of Blair et al. (2007), recorded grids are patterns of this kind.

47.4 Tests and Failures

Oscillatory interference and continuous attractor networks predicted different things, and between 2011 and 2022 the differences were tested. In attractor models, recurrent connections among grid cells sustain a periodic bump of activity that moves across the network as the animal moves, so that the relevant variable is the bump's position on a two-dimensional periodic space, a torus (Fuhs and Touretzky 2006; Burak and Fiete 2009). Giocomo, Moser, and Moser reviewed the models as they stood when the tests began (Giocomo et al. 2011).

Dependence on theta. Interference models require continuous theta-band oscillations. Yartsev, Witter, and Ulanovsky recorded grid cells in the medial entorhinal cortex of the Egyptian fruit bat and found grids closely resembling those of rodents in the absence of continuous theta oscillations and with almost no theta modulation of grid-cell spiking, which the authors argued refutes the proposed causal link between grids and theta (Yartsev et al. 2011). In rats, the evidence pointed the other way: inactivating the medial septum reduced theta and abolished the spatial periodicity of grid cells while head-direction

cells kept their selectivity (Brandon et al. 2011), and septal infusions that reduced theta made grid-like firing vanish while place cells and other spatially modulated cells kept their fields to a larger extent (Koenig et al. 2011). The two results can be reconciled, but the reconciliation removes theta from the role the interference models gave it: the septal input may be necessary for grids in rodents for reasons other than supplying the carrier of a beat.

Membrane potential. Interference models predict theta-amplitude modulations of the membrane potential that determine where fields lie; attractor models predict slow depolarizing ramps. Domnisoru, Kinkhabwala, and Tank recorded grid cells intracellularly in mice running along virtual linear tracks and found large, reproducible ramps of depolarization tightly correlated with firing fields; the cells also showed intracellular theta oscillations that influenced spike timing, but the properties of the theta amplitude modulations were not consistent with their determining field locations (Domnisoru et al. 2013). Schmidt-Hieber and Häusser, recording stellate cells in navigating mice, found that the membrane potential during field crossings was a slow depolarization best predicted by networks that supply sustained depolarizing input, as attractor models do, while the phase precession of intracellular theta and spiking was best captured by an interference model (Schmidt-Hieber and Häusser 2013).

Connectivity and population structure. Attractor models require specific recurrent connectivity. Couey and colleagues recorded from more than 600 neuron pairs in entorhinal slices and found that stellate cells, the principal cells of the layer II grid network, are interconnected mainly through inhibitory interneurons, and showed that a simple recurrent inhibitory network can generate stable grid firing (Couey et al. 2013). Gardner and colleagues recorded many hundreds of grid cells simultaneously and showed with topological data analysis that the joint activity of a module lies on a toroidal manifold, with positions on the torus corresponding to the animal's position and maintained across environments and from wakefulness to sleep, as continuous attractor models predict and the feedforward alternatives do not (Gardner et al. 2022).

Table 47.1: Discriminating predictions for grid-cell mechanisms and what was found.

Question	**Oscillatory interference**	**Continuous attractor**	**Finding**
Is continuous theta needed?	Yes	No	Grids without continuous theta in bats (Yartsev et al. 2011); grid periodicity lost with septal inactivation in rats (Brandon et al. 2011; Koenig et al. 2011)

Question	Oscillatory interference	Continuous attractor	Finding
What does the membrane do in a field?	Theta-amplitude modulation sets the field	Slow depolarizing ramp	Ramps; theta amplitude does not set fields (Domnisoru et al. 2013; Schmidt-Hieber and Häusser 2013)
Where does phase precession come from?	Interference	Needs an added mechanism	Intracellular precession best fit by interference (Schmidt-Hieber and Häusser 2013)
Is recurrent connectivity required?	No	Yes	Recurrent inhibitory circuitry among stellate cells (Couey et al. 2013)
How is population activity organized?	No constraint	Low-dimensional torus, stable across states	Torus, preserved across environments and sleep (Gardner et al. 2022)

The verdict of the tests is not that relations do not matter. In the attractor account the variable the system reads is still relational: a phase on a torus, the position of a bump relative to the network that carries it. What failed was the proposed physical reader and generator of that relation, the interference of velocity-controlled oscillators, as the source of the grid. The beat survives in a smaller role, as a plausible account of spike timing and phase precession.

47.5 Retinal Moiré and Orientation Maps

The second neural moiré theory concerns vision. In primary visual cortex of many mammals, preferred orientation varies smoothly across the cortical surface and forms iso-orientation domains arranged around pinwheel centers. Soodak used a computer simulation to argue that the tangential organization of orientation columns follows from the orderly projection of the mosaic of retinal ganglion cells onto the cortex, consistent with the presence of orientation columns in visually inexperienced animals (Soodak 1987). Ringach proposed that the blueprint for receptive fields, feature maps, and their interrelations may reside in the layout of retinal ganglion cell mosaics together with a simple statistical scheme of connectivity from thalamus to cortex (Ringach 2007). Paik and Ringach then proposed that quasi-periodic orientation maps are established by moiré interference between the regularly spaced mosaics of ON-center and OFF-center retinal ganglion cells. Their theory predicted that the centers of iso-orientation domains should lie on a hexagonal lattice on the cortical surface, and they reported such a pattern in individuals of four species (monkeys, cats, tree shrews, and ferrets); the mechanism would explain how

maps develop without precise patterns of spontaneous activity or molecular guidance, and why rodents have orientation-tuned neurons without orderly maps (Paik and Ringach 2011).

The critics tested the proposal against quantitative laws. Kaschube and colleagues analyzed orientation maps in the ferret, tree shrew, and galago, species separated since the basal radiation of placental mammals, and found a common layout, including a pinwheel density per hypercolumn area close to π; a symmetry-based class of self-organization models predicted the layout, but only if suppressive long-range interactions dominate development (Kaschube et al. 2010). Schottdorf and colleagues generalized the stochastic wiring model of Ringach and colleagues, derived closed-form predictions for perfectly hexagonal mosaics, added realistic disorder, and found that, independent of parameters, the predicted layouts deviated substantially from the experimentally observed laws; with the most realistic model of mosaic layout, the predicted maps resembled Gaussian random fields (Schottdorf et al. 2015). They concluded that the invariants of visual cortical layout are signatures of cortical optimization that generic random feedforward wiring cannot explain.

The retinal moiré theory predicted a qualitative feature, a hexagonal arrangement, that the data showed in some individuals. It did not reproduce the quantitative invariants, such as pinwheel density, that its rivals were built to explain. A relational mechanism that reproduces a qualitative pattern shared by many models has not yet been tested; the test is the quantitative prediction on which the models differ.

47.6 Lessons for Theory Building

Five lessons carry beyond grid cells and orientation maps.

Principle 47.4 (Relational mechanisms and their evidence)**.** Evidence for a relational neural mechanism is subject to five constraints.

1. Mathematical elegance and qualitative fit are not evidence for a mechanism over its rivals; many mechanisms produce hexagons.
2. The decisive tests are predictions on which mechanisms differ: membrane potential trajectories, dependence on a carrier rhythm, connectivity, population geometry, quantitative layout statistics.
3. A relational *variable* can be right while the proposed physical *reader and generator* of it is wrong; the grid phase survived the failure of the interference mechanism.
4. Comparisons across species are relational interventions of a sort: they change the substrate while the function stays fixed, as bats without continuous theta did for grid cells.
5. Each failed prediction teaches what a successful theory must also explain, such as phase precession, which the interference model still captures best.

The lessons apply with extra force to the question with which this book began. A moiré theory of consciousness would have to name the layers that interfere, the physical process that makes their relation explicit, and a prediction on which it differs from the theories of Chapter 44. The neural moiré theories of grids and orientation maps met that standard and were still found wanting in important respects. Chapters 60 and 62 state what a relational theory of consciousness would have to survive.

47.7 Audit

Canonical phenomenon

Grid cells in the medial entorhinal cortex fire on a triangular lattice of locations; orientation preference in primary visual cortex forms a quasi-periodic map of domains and pinwheels (Hafting et al. 2005; Paik and Ringach 2011).

Minimal model

Velocity-controlled oscillators whose beat phase is position (Proposition 47.2); moiré lattices from rotated or rescaled lattices (47.3) and (47.4); continuous attractor networks with toroidal state spaces.

What is relational here?

The phase difference between two oscillators (interference models), the relative orientation or scale of two lattices (moiré grids and retinal moiré), and the position of an activity bump on a torus (attractor models).

What physically reads it?

In interference models, a cell's firing threshold applied to the sum of oscillations; in attractor models, the recurrent network itself, whose connectivity constrains activity to the torus; in retinal moiré models, the statistical wiring from mosaics to cortex.

What existing field explains it?

The physics of moiré and beats (Chapters 2 and 3), the neurophysiology of theta and phase precession (O'Keefe and Recce 1993), attractor network theory (Fuhs and Touretzky 2006; Burak and Fiete 2009), and theories of cortical self-organization (Kaschube et al. 2010).

What remains unexplained?

How theta and phase precession fit into an attractor account of grids; what developmental mechanism produces the quantitative invariants of orientation maps.

A discriminating experiment

For grids, the tests of Table 47.1 have largely been run. For any proposed neural moiré mechanism, the discriminating experiment follows the same

template: identify the layers and the reader, derive a quantitative prediction that the rival mechanism does not make (a dependence of the pattern's scale on the mismatch between layers, for example), and manipulate the mismatch while measuring the scale.

47.8 Notes and References

Grid cells were discovered by Hafting et al. (2005), and their modular organization is shown by Stensola et al. (2012). Phase precession is reported by O'Keefe and Recce (1993); the interference model is proposed in O'Keefe and Burgess (2005) and developed in Burgess et al. (2007); moiré grids are proposed by Blair et al. (2007). Attractor models are Fuhs and Touretzky (2006) and Burak and Fiete (2009), and the models of the period are reviewed by Giocomo et al. (2011). The discriminating tests are Yartsev et al. (2011), Brandon et al. (2011), Koenig et al. (2011), Domnisoru et al. (2013), Schmidt-Hieber and Häusser (2013), Couey et al. (2013), and Gardner et al. (2022). For orientation maps, the mosaic proposals are Soodak (1987), Ringach (2007), and Paik and Ringach (2011); the universality result is Kaschube et al. (2010), which was followed by a published comment and response, and the quantitative test of random wiring is Schottdorf et al. (2015). The moiré formulas are derived in Chapter 2 and in Amidror (2009).

47.9 Exercises

Exercise 47.1. A place cell bursts at 9 Hz while theta runs at 8 Hz (illustrative values). Compute the phase advance per theta cycle and the number of theta cycles for a full 360-degree precession. If the rat crosses the place field in one second, how much of the observed precession range of O'Keefe and Recce (1993) does this account for?

Exercise 47.2. Simulate a random trajectory in a unit square and three velocity-controlled oscillators of Model 47.1 with directions 0, 60, and 120 degrees and gain β. Plot the locations where the product of the three interference signals exceeds a threshold and check the lattice spacing against Proposition 47.2. Add independent phase noise to each oscillator and measure how quickly the grid degrades; relate the result to the need for landmark resets.

Exercise 47.3. Using (47.3) and (47.4), compute the moiré spacing for two lattices of spacing 30 cm rotated by 3 degrees, and for aligned lattices of spacings 30 cm and 33 cm. What mismatch would produce a moiré grid of spacing 3 m, and how precisely would the mismatch have to be maintained?

Exercise 47.4. Explain why the joint activity of a module of grid cells should lie on a two-dimensional torus in a continuous attractor model, and why a

population of independent interference-based cells need not. What property of the data in Gardner et al. (2022) distinguishes the two predictions?

Exercise 47.5. Generate ON and OFF mosaics as noisy hexagonal lattices with slightly different spacings, compute their moiré pattern, and measure its period. Increase the positional disorder and report the disorder at which the moiré period can no longer be estimated. Relate your result to the conclusions of Schottdorf et al. (2015).

Exercise 47.6. Write down, in the form of Principle 47.4, what a moiré theory of consciousness would have to specify: the layers, the reader, the relational variable, and one quantitative prediction that differs from those of the theories in Chapter 44. Then design the test.

A functional perspective is a set of agent-centered variables that a system reads and uses. A phenomenal perspective is there being something it is like to be that system. Every major theory of consciousness names an additional property that is supposed to take us from the first to the second; none of them has yet been shown to do so. Relational neural measures correlate with the state of consciousness, and they have earned a place only where they outperform simpler measures. Part XI puts four architectures side by side and asks what changes at each step.

Part XI

The Comparative Ladder

What actually changes at each step?

48

The Radio

A radio receiver generates an oscillation of its own, mixes it with whatever arrives at the antenna, reads a difference frequency, and, in its modern forms, changes its own operating frequency when a channel is jammed. Written out in the vocabulary of this book, that is most of the reading loop, and most of what Part V found in the electric fish. The radio therefore makes the right first rung for a ladder that ends with human beings. It is an engineered reader whose every part is understood, and it forces a question that later rungs will keep asking in harder forms.

The thesis of this chapter is that an engineered system can read relations, can read the *sign* of a relation, and can act on what it reads so as to change the relation, all without anything that we are inclined to call meaning or experience. Whatever separates the electric fish from an adaptive radio, then, cannot lie in the mapping from mixtures to actions, because that mapping is shared. It must lie in how the reader came to exist, in what its readings are for, or in some internal property that the mapping does not fix. Part XI tracks those candidates rung by rung, and Table 48.1 starts the tally.

48.1 Carrier and Modulation

A transmitter emits a carrier, a sinusoid at an angular frequency ω_c, and writes a message onto it by modulation. In amplitude modulation the message $\mu(t)$ scales the carrier's envelope:

$$s(t) = A_c \left[1 + k\,\mu(t)\right] \cos \omega_c t, \qquad |k\,\mu(t)| < 1. \tag{48.1}$$

In frequency modulation the message shifts the instantaneous frequency, $s(t) = A_c \cos\!\big(\omega_c t + k \int^t \mu(t')\,\mathrm{d}t'\big)$. In both cases the message is not a property of the carrier taken alone. It is carried by a relation between the received waveform and a reference: the unmodulated envelope in the first case, the nominal carrier frequency in the second. A receiver extracts the message by making that relation explicit, which is the operation Chapter 4 called demodulation.

The carrier also defines the channel. Radio spectrum is shared by dividing it into frequency bands, and a receiver must select one band and suppress the

rest. Selection, demodulation, and suppression of interference are the three functions of every receiver, and each can be described as the reading of a relation.

48.2 Interference

Suppose a wanted carrier of amplitude A at ω_c and an unwanted carrier of amplitude A' at a nearby frequency ω' both reach the antenna. Writing $m = A'/A$ and $\Delta\omega = \omega' - \omega_c$, the received signal is

$$y(t) = A\big[\cos\omega_c t + m\cos(\omega' t + \psi_0)\big] = A\ \mathrm{Re}\Big\{\mathrm{e}^{\mathrm{i}\omega_c t}\big(1 + m\,\mathrm{e}^{\mathrm{i}(\Delta\omega\, t + \psi_0)}\big)\Big\}. \tag{48.2}$$

This is, symbol for symbol, the mixture that reaches a patch of the electric fish's skin when a neighbor discharges near its frequency, with m playing the role of the contamination ratio (Chapter 20). The envelope of y beats at $|\Delta\omega|$,

$$A_y(t) = A\sqrt{1 + m^2 + 2m\cos(\Delta\omega\, t + \psi_0)}, \tag{48.3}$$

and its phase, measured against the wanted carrier, wobbles at the same rate. An amplitude-modulation receiver that follows the envelope delivers the beat to its loudspeaker as a tone at $|\Delta\omega|/2\pi$, the whistle familiar to anyone who has tuned between two stations at night. Nothing in the receiver distinguishes that tone from a message: the interference has been read, faithfully, as if it were signal.

Interference is a relational phenomenon in exactly the sense of Chapter 1. Neither carrier is defective. The problem lies in their difference frequency, and it grows worse as that difference shrinks toward the bandwidth of the message, because the beat then falls inside the band the receiver must pass. The fish faces the same arithmetic: a neighbor whose frequency is close to its own imposes a slow beat on the carrier the fish uses to sense its surroundings (Chapter 20).

48.3 Demodulation and the Superheterodyne

The simplest demodulator for (48.1) is a rectifier followed by a low-pass filter. The rectifier is the nonlinearity; Chapter 4 showed how a nonlinear element puts energy at difference frequencies that a linear sum does not contain, and the low-pass filter keeps the slow terms and discards the rest. The envelope detector reads a relation, the envelope, that the Fourier spectrum of the input holds only implicitly.

Early receivers amplified and filtered at the frequency of the station itself, which made tuning awkward, since every stage had to be retuned together. Armstrong's superheterodyne receiver moved the problem to a fixed frequency

of the receiver's own choosing (E. Armstrong 1921). The incoming signal is multiplied by the output of a local oscillator at ω_{LO}:

$$\cos\omega_{\mathrm{RF}}t \cdot \cos\omega_{\mathrm{LO}}t = \tfrac{1}{2}\cos(\omega_{\mathrm{RF}} - \omega_{\mathrm{LO}})t + \tfrac{1}{2}\cos(\omega_{\mathrm{RF}} + \omega_{\mathrm{LO}})t. \tag{48.4}$$

A filter fixed at an intermediate frequency ω_{IF} keeps the difference term and discards the sum. To tune to a different station, the receiver changes only its local oscillator. All selectivity, gain, and demodulation then happen at ω_{IF}, a frequency defined by the relation between the world and the receiver's own oscillation (Figure 48.1).

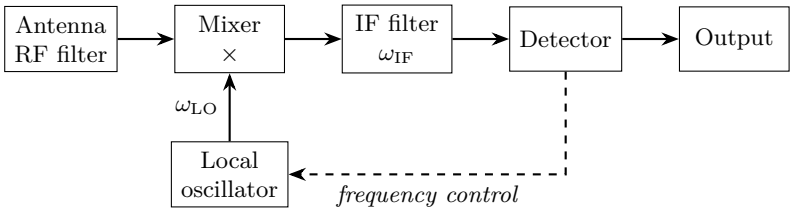


Figure 48.1: A superheterodyne receiver drawn as a reader. The mixer delivers the incoming signal together with the receiver's own oscillation; the fixed filter selects one difference frequency; the dashed path is automatic frequency control, which retunes the local oscillator to keep the difference constant.

Model 48.1 (The superheterodyne as a reading loop)**.** In the notation of Chapter 15, let $s_{\mathrm{A}} = \omega_{\mathrm{LO}}$ be the receiver's own frequency and s_{E} the set of carriers present at the antenna. The mixer implements $y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})$ by multiplication; the intermediate-frequency filter and detector implement $z = \mathsf{D}(y)$, which responds only to carriers whose difference from ω_{LO} lies in the passband; and automatic frequency control implements a policy $a = \pi(z)$ that adjusts ω_{LO} so as to hold the difference at ω_{IF}. The loop is closed: what the receiver reads determines the next value of the relation it reads.

The superheterodyne has a characteristic blind spot. Because cos is even, the mixer maps a carrier at $\omega_{\mathrm{LO}} + \omega_{\mathrm{IF}}$ and a carrier at $\omega_{\mathrm{LO}} - \omega_{\mathrm{IF}}$ onto the same intermediate frequency. The second is called the *image*, and a receiver with a single real mixer cannot tell a station from its image. Designers suppress the image with a filter placed before the mixer, or with a pair of mixers driven in quadrature (Horowitz and Hill 2015). The image problem is the sign problem of the electric fish in engineering dress: a single real channel registers the magnitude of a frequency difference and loses its sign.

48.4 Reading the Sign of a Frequency Offset

The quadrature solution is worth writing out, because it shows exactly what information the sign of a frequency difference requires. Multiply the input

$y(t) = \cos\big((\omega_{\mathrm{LO}} + \Delta\omega)t + \psi_0\big)$ by two copies of the local oscillator a quarter cycle apart and low-pass filter the products. Using (48.4) and its sine counterpart,

$$I(t) = \tfrac{1}{2}\cos(\Delta\omega\, t + \psi_0), \qquad Q(t) = \tfrac{1}{2}\sin(\Delta\omega\, t + \psi_0), \tag{48.5}$$

so the point $\big(I(t), Q(t)\big)$ circles the origin at rate $|\Delta\omega|$, counterclockwise when $\Delta\omega > 0$ and clockwise when $\Delta\omega < 0$. The sign of the offset is the sense of rotation. A two-phase lock-in amplifier computes the same pair of outputs against a reference supplied by the experimenter (Dicke 1946; Scofield 1994).

The following result states the general principle.

Theorem 48.2 (Sign of a detuning from sinusoidal channels)**.** *Let a reader have access to channels $z_k(t) = b_k + a_k\cos(\Delta\omega\, t + \psi_0 - \alpha_k)$, $k = 1, \dots, K$, with known $a_k \neq 0$, α_k, and b_k, and unknown $\Delta\omega$ and ψ_0.*

1. *If $K = 1$, the observations generated by $(\Delta\omega, \psi_0)$ coincide with those generated by $(-\Delta\omega,\ 2\alpha_1 - \psi_0)$ for all t. The sign of $\Delta\omega$ is not identifiable.*
2. *If $K = 2$ and $\alpha_2 - \alpha_1 \not\equiv 0 \pmod{\pi}$, then with $\tilde z_k = z_k - b_k$,*

$$\tilde z_1\, \dot{\tilde z}_2 - \dot{\tilde z}_1\, \tilde z_2 = a_1 a_2\, \Delta\omega\, \sin(\alpha_2 - \alpha_1) \tag{48.6}$$

at every instant, so $\operatorname{sgn}\Delta\omega = \operatorname{sgn}\big(\tilde z_1\dot{\tilde z}_2 - \dot{\tilde z}_1\tilde z_2\big) \cdot \operatorname{sgn}\big(a_1 a_2 \sin(\alpha_2 - \alpha_1)\big)$.

Proof. For $K = 1$: $\cos(-\Delta\omega\, t + 2\alpha_1 - \psi_0 - \alpha_1) = \cos\big(-(\Delta\omega\, t + \psi_0 - \alpha_1)\big) = \cos(\Delta\omega\, t + \psi_0 - \alpha_1)$, so the two parameter pairs produce identical channels. For $K = 2$, write $\psi = \Delta\omega\, t + \psi_0$. Then $\dot{\tilde z}_k = -a_k\Delta\omega\sin(\psi - \alpha_k)$, and $\tilde z_1\dot{\tilde z}_2 - \dot{\tilde z}_1\tilde z_2 = -a_1a_2\Delta\omega\big[\cos(\psi - \alpha_1)\sin(\psi - \alpha_2) - \sin(\psi - \alpha_1)\cos(\psi - \alpha_2)\big] = -a_1a_2\Delta\omega\sin(\alpha_1 - \alpha_2)$, which is (48.6). □

The left side of (48.6) is twice the rate at which the trajectory $(\tilde z_1, \tilde z_2)$ sweeps out signed area, a direct measure of its sense of rotation. For the quadrature pair (48.5), $\alpha_2 - \alpha_1 = \pi/2$, $a_1 = a_2 = \tfrac{1}{2}$, and the left side of (48.6) equals $\Delta\omega/4$. The theorem is stated for sinusoidal channels. A channel whose waveform is distorted by higher harmonics can carry some sign information in the asymmetry of its shape, and Exercise 48.3 explores the point.

The theorem places the radio and the fish side by side. Both read the sign of a frequency difference from two channels in approximate quadrature. The radio obtains its two channels by comparing the input with an internal reference, its own oscillator, split into two phases. The fish obtains its two channels, amplitude modulation and the phase difference between two patches of skin, from the mixture alone, with no reference to its own motor command (Heiligenberg 1991a; Kawasaki 1993); Chapter 21 derives its rule and shows that it is the same rotation criterion. The input-output function is shared. The route to it is different, and the difference is instructive: the fish has no oscillator it can consult, so it must find two unequal views of the same mixture on its own body.

48.5 Channel Selection, Feedback, and Adaptive Frequency Hopping

Receivers have long contained closed loops. Automatic gain control holds the output level steady as the input fades; automatic frequency control and the phase-locked loop hold the local oscillator in a fixed relation to the incoming carrier. Each loop reads a relation (a level against a set point, a frequency or phase against a reference) and adjusts the receiver's own state to hold the relation where it is wanted. These are the negative-feedback loops that Wiener (1948) took as the prototype of purposive behavior in machines and animals alike.

Adaptive frequency hopping brings the radio closest to the jamming avoidance response. Bluetooth devices hop rapidly among 79 radio channels of one megahertz each, in a band they share with wireless networks and other equipment (*Specification of the Bluetooth System, Version 1.2* 2003). Version 1.2 of the specification, adopted in 2003, introduced adaptive frequency hopping: devices assess the quality of each channel from their own reception, classify channels as good or bad, and hop only among the good ones (*Specification of the Bluetooth System, Version 1.2* 2003). A device that detects interference on a band therefore changes its own transmission frequencies to move away from it. Described at the level of input and output, that is what the fish does: it senses a neighbor near its frequency and moves its own frequency so as to increase the difference (Metzner 1993b).

Example 48.3 (Adaptive hopping as a reading loop). The environment s_{E} is the set of interferers occupying the band; the agent's state s_{A} is its current map of good and bad channels. Each transmission mixes the device's signal with whatever else occupies its channel, $y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})$. The receiver's error statistics make channel quality explicit, $z = \mathsf{D}(y)$. The classification rule is the policy, $a = \pi(z)$, and updating the hop set is the intervention, $\dot{s}_{\mathrm{A}} = \mathsf{F}(s_{\mathrm{A}}, a)$. The next mixture depends on the updated hop set, so the loop is closed. The device also keeps a running record of how each channel has performed, a memory of its own recent input.

The channel classification deserves a second look. The quantity being estimated is the reliability of the device's own readings on each channel. That is a reading about readings, a crude form of the self-monitoring that Chapter 39 studies under the name metacognition, and it is why the radio receives a *partial* in the corresponding row of the complete ladder (Chapter 52).

48.6 Cognitive Radio

Mitola and Maguire (1999) proposed the *cognitive radio*: a software-defined radio that models its own capabilities, its user's needs, and the state of the spectrum, and reconfigures itself accordingly. Haykin (2005) gave the idea an

engineering definition as a wireless system that is aware of its radio environment, learns from it, and adapts its operating parameters in real time, with the twin aims of reliable communication and efficient use of the spectrum. Haykin's cognitive cycle has three parts: analysis of the radio scene to find unused spectrum and estimate interference, estimation and prediction of channel state, and control of transmit power and spectrum access.

In the vocabulary of Part VII, a cognitive radio carries a detached model of its environment: it predicts channel states it has not yet sampled and acts on the predictions (Chapter 31). It does so to serve objectives set by its designers and users. The prediction is genuine and so is the adaptation, but the standard against which the radio's readings succeed or fail is external to it. A cognitive radio that misclassifies a channel has made an error relative to the purposes of the network operator. Nothing about the radio's own continued existence depends on it.

48.7 What the Radio Lacks

The previous sections establish that an engineered system can make relational variables explicit, read the sign of a relation, keep a memory of its recent input, estimate the reliability of its own readings, and act to change the relations it reads. The following principle draws the consequence.

Principle 48.4 (Input-output parity)**.** If two systems realize the same mapping from mixtures to actions at a given level of description, then any property possessed by one and not the other is not a property of that mapping.

The principle is close to a tautology, and its value lies in what it rules out. An adaptive radio and an electric fish both map "interferer near my frequency" to "move my frequency away." If the fish reads, means, or experiences something that the radio does not, the difference must lie elsewhere: in the internal organization that realizes the mapping, in the history that produced the reader, or in the relation between the readings and the continued existence of the system that makes them. A theory of consciousness that locates experience in sensorimotor function alone must either grant the adaptive radio whatever it grants the jamming avoidance response, or specify a further ingredient.

That settles nothing about the fish, but it narrows the search. The candidate differences are few, and each is taken up at a later rung.

1. *Who built the reader.* The radio's detectors, filters, and classifiers were designed and assembled by engineers. The fish grows its receptors, its decoding circuits, and its pacemaker, and maintains them. Pattee's semantic closure makes this difference central (Chapters 26 and 49).
2. *What the readings are for.* The radio's readings serve its users; its errors are errors relative to their purposes. The fish's readings serve a system whose persistence depends on them (Chapters 28 and 30).

3. *Internal organization.* Theories of consciousness disagree about which internal properties matter, whether global broadcast, integrated causal structure, higher-order representation, or something else, and several of them would assess a radio and a fish very differently even where their input-output functions coincide (Chapter 44).
4. *History.* The fish's reader is the product of selection; the radio's is the product of design.

Two answers that are commonly given should be set aside. The first is that we know how the radio works. We also know, to the level of identified nuclei and transmitter types, how the fish computes its jamming avoidance response (Heiligenberg 1991a; Metzner 1993b), and mechanistic knowledge cannot by itself decide the question for one system and not the other. The second is that the radio's behavior is simple. Simplicity invites attributions as easily as it defeats them: Braitenberg (1984) showed that vehicles with two sensors and two motors wired crosswise or directly behave in ways observers readily describe as fear, aggression, and love. Our readiness to attribute or withhold experience is data about us. It becomes evidence about the system only when it is backed by a theory that says which property matters and why.

The question with which Part XI begins can now be asked precisely. Why don't we ordinarily consider the radio conscious? The reasons we have concern closure, purpose, organization, and history; none of them concerns reading as such, and none has yet been shown to be the difference that makes a difference to experience.

48.8 The First Rung

Table 48.1 records what the radio has. Each later chapter of Part XI re-assesses these rows for its own rung and adds rows for the capabilities it introduces; the complete table is in Chapter 52.

Table 48.1: The comparative ladder, first rung. The radio has the reading loop. Entries: *yes*; *no*; *partial*; *derived* (only relative to a designer's or user's purposes); *contested* (the evidence is disputed); *?* (not established by the evidence reviewed in this book).

Capability	Radio
Introduced with the radio	
Separates a signal from interference	yes
Makes a relational variable explicit	yes
Reads the sign of a relation	yes
Acts to change a relation it reads	yes

The radio passes every row introduced so far. That is the point of starting here. Every capability in this first group is a capability of the reading loop as such, and the loop can be built from resistors, capacitors, crystals, and code.

48.9 Audit

Canonical phenomenon

A Bluetooth device detects interference on some of its channels and removes them from its hopping sequence, moving its own frequencies away from the interferer.

Minimal model

Two carriers beating at the antenna (48.2); the superheterodyne mixer (48.4); quadrature demodulation (48.5); the rotation criterion for the sign of a detuning (Theorem 48.2).

What is relational here?

The difference frequency between an incoming carrier and the receiver's own oscillator; the envelope of a mixture; the sign of a frequency offset, carried by the sense of rotation of a pair of quadrature channels; the quality of a channel relative to the device's own reception.

What physically reads it?

Mixers, filters, rectifiers, phase-locked loops, and the channel-classification logic, all designed and built by engineers. The quadrature mixer reads the sign of an offset against an internal reference.

What existing field explains it?

Communication engineering and electronics (E. Armstrong 1921; Horowitz and Hill 2015), signal processing, control theory and cybernetics (Wiener 1948), and the cognitive-radio literature (Mitola and Maguire 1999; Haykin 2005). Every mechanism in the chapter is explained by these fields.

What remains unexplained?

Nothing about how the radio reads. The residue is comparative: the radio shares the input-output function of the jamming avoidance response, so whatever the fish has that the radio lacks must be sought in closure, purpose, internal organization, or history.

A discriminating experiment

Build or simulate an adaptive-hopping radio and a model fish whose mappings from mixtures to frequency changes are matched, then ask each theory of consciousness in Chapter 44 to state, in advance, which of the two it credits and on what measurable property. A theory that cannot distinguish them, or that distinguishes them by a property it cannot measure, has been shown to rest on something other than sensorimotor function.

48.10 Notes and References

E. Armstrong (1921) describes the superheterodyne principle; Horowitz and Hill (2015) covers mixers, oscillators, phase-locked loops, and receiver design

at a working engineer's depth. The lock-in amplifier, which reads a signal's amplitude and phase against a supplied reference, descends from Dicke (1946); Scofield (1994) gives a clear frequency-domain account. Wiener (1948) is the founding text for treating feedback in machines and organisms with one vocabulary. Adaptive frequency hopping entered the Bluetooth standard with version 1.2 (*Specification of the Bluetooth System, Version 1.2* 2003). Mitola and Maguire (1999) introduced the cognitive radio and Haykin (2005) gave it its standard engineering formulation. Braitenberg (1984) remains the best short demonstration of how little machinery it takes to provoke mental attributions.

The quadrature construction of Theorem 48.2 is standard in communications; the observation that the jamming avoidance response solves the same sign problem without an internal reference is Heiligenberg's (Heiligenberg 1991a), and the independent solution in *Gymnarchus* is Kawasaki (1993).

48.11 Exercises

Exercise 48.1. A superheterodyne receiver has $f_{\mathrm{IF}} = 455\,\mathrm{kHz}$ and is tuned to a station at $1000\,\mathrm{kHz}$ with the local oscillator above the signal. Find f_{LO} and the image frequency. Explain, using Theorem 48.2(1), why a single real mixer cannot distinguish the station from its image, and how a second mixer driven a quarter cycle later resolves the ambiguity.

Exercise 48.2. Verify (48.5) and compute $I\dot{Q} - \dot{I}Q$. Add independent white noise to I and Q and simulate the estimate $\operatorname{sgn}\langle I\dot{Q} - \dot{I}Q\rangle$ over windows of increasing length. How does the error rate depend on $|\Delta\omega|$, the noise level, and the window length?

Exercise 48.3. Let a single channel carry $z(t) = \cos\psi + c\cos 2\psi$ with $\psi = \Delta\omega\, t + \psi_0$. Show that the observations for $(\Delta\omega, \psi_0)$ and $(-\Delta\omega, -\psi_0)$ coincide for every c. Now let $z(t) = \sin\psi + c\sin 2\psi$ and show that, for $c \neq 0$, no choice of ψ_0' makes the observations for $-\Delta\omega$ coincide with those for $\Delta\omega$. What does this imply about sign information in the waveform of a distorted modulation, and why does Theorem 48.2 restrict itself to sinusoidal channels?

Exercise 48.4. Write automatic frequency control as a reading loop $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$. Linearize it about lock and find the condition on the loop gain for the frequency error to decay. Compare with the local instability of the jamming avoidance response reported by Madhav et al. (2013): what feature of the fish's loop makes $\Delta f = 0$ repelling rather than attracting?

Exercise 48.5. Design a channel classifier for adaptive hopping that uses only the device's own packet-error counts. State its memory (how long a channel stays marked bad) and discuss the trade-off between reacting quickly to a new interferer and tolerating random errors. Which row of Table 48.1 does the memory correspond to in the next rung (Chapter 49)?

Exercise 48.6. State Principle 48.4 for a specific theory of consciousness from Chapter 44. Does the theory assign different verdicts to an adaptive radio and to the fish's jamming avoidance response? If so, identify the property it relies on and propose a measurement that could establish whether each system has it.

Exercise 48.7. Imagine a radio that detects its own component failures and replaces them from a store of spare parts it also manufactures. Which rows of the complete ladder (Chapter 52) would change? Which would not, and what does that suggest about the relation between self-repair and the capabilities introduced with the cell?

49

The Cell

A bacterium swimming in a sugar gradient reads a relation, acts on it, and changes the relation it reads, as the adaptive radio of Chapter 48 does. It adds something the radio lacks. The receptors, signaling proteins, and motors that make up its reader are built by the bacterium from its own genome and food, they wear out and are replaced by the same processes, and the whole apparatus exists because cells that read their surroundings this way left more descendants than cells that did not.

The thesis of this chapter is that the cell is the first rung on which a reading can be good or bad *for the system that performs it*. The reader is a product of the organization it serves, and its readings feed the processes that keep that organization, and the reader, in existence. With that closure a norm appears, and with the norm the possibility of error that is the system's own and not its designer's. Whether this deserves the name *meaning* depends on which theory of content one holds, and the chapter shows what each theory says about the same bacterium. It does not require consciousness, on any theory the book reviews.

49.1 Self-Production and the Boundary

A cell is a network of chemical processes that produces the components that realize it, including the membrane that separates it from its surroundings. F. Varela, H. Maturana, et al. (1974) called such an organization *autopoietic* and proposed criteria for recognizing it, and H. R. Maturana and F. J. Varela (1980) developed the idea into a theory of the living. Related formulations describe organisms as closed to efficient causation (Rosen 1991), or as systems whose constraints are mutually dependent, each maintained by the others (Montévil and Mossio 2015; Moreno and Mossio 2015). Chapter 27 compares these formulations. Here we need only the feature they share: the parts of a cell are produced and replaced by the activity of the cell.

The radio has no such feature. If a capacitor fails, the radio does not replace it; an engineer does, or the radio stops. The distinction does not concern complexity, since some engineered systems are far more complex than

a bacterium. It concerns where the processes that build and repair the parts are located. For the cell they are inside, part of the same organization they maintain.

The boundary matters for a reader in a specific way. A reader needs an inside and an outside: a surface across which the mixture $y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})$ is delivered, and a side on which the read variable $z = \mathsf{D}(y)$ is used. In a radio the boundary is a casing and a connector, fixed at manufacture. In a cell it is a membrane the cell continuously makes, and the receptors that sit in it are among the components the cell produces. The reader's location in the system is itself a product of the system.

49.2 Chemotaxis: Reading the Present Against the Past

Escherichia coli swims by rotating its flagella. Tracking individual cells in three dimensions, Berg and D. A. Brown (1972) showed that the motion consists of runs, nearly straight swims, interrupted by tumbles that reorient the cell at random, and that runs are extended when a cell happens to be moving up a gradient of attractant. The result is a biased random walk that climbs the gradient on average although no single run is aimed.

How does the cell know it is moving up the gradient? Macnab and Koshland (1972) built a "temporal gradient apparatus" that exposed bacteria to a sudden change from one uniform concentration of attractant to another, with no spatial gradient at all. A sudden decrease produced tumbling; a sudden increase suppressed it; and both responses gradually returned to the steady-state pattern. Their conclusion is the one this chapter needs: the apparent detection of a spatial gradient involves the detection of a temporal gradient experienced as a result of movement through space. The bacterium compares the concentration it senses now with the concentration it sensed a short while ago.

Physical limits make the temporal strategy the natural one. Berg and Purcell (1977) showed that counting noise limits how precisely a cell can measure a concentration: the least fractional error attainable in a time T is approximately $(TcaD)^{-1/2}$, where c is the concentration, a the cell's radius, and D the attractant's diffusion constant. A comparison across the length of a bacterium spans a distance of order a; a comparison over time spans the distance the cell swims in the comparison interval, which can be many body lengths. In a shallow gradient the first difference is smaller than the counting noise and the second need not be. Swimming converts a spatial difference too small to measure into a temporal difference that can be measured.

Model 49.1 (Minimal adaptive reader)**.** Let $c(t)$ be the attractant concentration at the cell and h an increasing function describing receptor occupancy.

Let the cell carry a slow internal variable $\mu(t)$, a stand-in for the methylation state of its receptors, and let the signal that controls tumbling be

$$u(t) = h(c(t)) - \mu(t), \qquad \tau\,\dot{\mu} = u(t). \tag{49.1}$$

Positive u suppresses tumbling and negative u promotes it. This is a toy model chosen for transparency; the biochemistry of the *E. coli* pathway is richer, and Barkai and Leibler (1997) give a quantitative model of it.

Two properties of Model 49.1 carry the chapter's argument.

Proposition 49.2 (Exact adaptation and derivative reading)**.** *In Model 49.1:*

1. *after a step change in c to a new constant value, $u(t)$ decays exponentially to exactly zero with time constant τ, for every function h and every step size;*
2. *if $c(t)$ varies slowly compared with τ, then $u(t) \approx \tau\, h'(c(t))\, \dot{c}(t)$.*

Proof. Differentiating $u = h(c) - \mu$ and using $\tau\dot{\mu} = u$ gives $\dot{u} = h'(c)\,\dot{c} - u/\tau$. For constant c after a step, $\dot{u} = -u/\tau$, so $u(t) = u(0^+)\,\mathrm{e}^{-t/\tau} \to 0$, whatever h is. For slowly varying c, set $\dot{u} \approx 0$ to obtain $u \approx \tau h'(c)\dot{c}$. □

The first property is exact adaptation: whatever the background level, the output returns to the same set point, so the cell responds to changes and not to levels. The variable μ integrates the output and subtracts it, which is integral feedback. In the real pathway the precision of adaptation is robust: Alon et al. (1999) varied the concentrations of the chemotaxis proteins and found that steady-state behavior and adaptation time changed while the precision of adaptation did not, as the model of Barkai and Leibler (1997) predicted. The second property says what the adapted cell reads. With $h = \ln$, the output approximates $\tau\, \mathrm{d}\ln c/\mathrm{d}t$, the relative rate of change of concentration, a relation between the present and the recent past.

The read variable also contains the cell's own motion. A cell swimming with velocity $\boldsymbol{v}$ through a static field $c(\boldsymbol{x})$ experiences

$$\dot{c} = \boldsymbol{v} \cdot \nabla c, \qquad \text{so} \qquad u \approx \tau\, h'(c)\; \boldsymbol{v} \cdot \nabla c. \tag{49.2}$$

The quantity that controls tumbling is the component of the gradient along the cell's own direction of travel. It is a variable of the form "world relative to my action," the minimal agent-centered measurement frame that Chapters 22 and 34 develop for the electric fish. The cell does not emit the energy it senses, so it falls short of active sensing in the full sense of Chapter 19, but its reading exists only because it moves; the ladder records this as *partial*.

Example 49.3 (The same loop, a different reader). In the notation of the reading loop, s_{E} is the attractant field and s_{A} the cell's position, heading, and methylation state. Swimming delivers c along the path, $y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})$; receptor occupancy and adaptation compute $z = \mathsf{D}(y) = u$; the motor switch

implements $a = \pi(z)$, run or tumble; and F moves the cell, changing the next concentration it will meet. Every element appeared in the adaptive radio. What differs is who made the elements and what they are for.

49.3 The Reader Is Built by the System It Serves

Howard Pattee spent a career asking how a physical system can contain a measuring device, a question Chapter 25 takes up at length. His answer for the cell is *semantic closure*: the cell constructs its measuring apparatus from genetic records, uses the results of measurement to control its dynamics, and thereby maintains the organization that constructs and interprets the records (Pattee 1969; Pattee 1995; Pattee 2001). The chemotactic cell instantiates the loop concretely. Genes specify the receptor and signaling proteins; the proteins read the temporal gradient; the reading controls motility; motility brings the cell to food; and food supplies the material and energy with which the cell builds the proteins again.

The contrast with the radio can now be stated without appeal to anything mysterious. In both systems a physical reader makes a relational variable explicit and a policy acts on it. In the radio, the processes that construct and maintain the reader lie outside the loop, in factories and engineers. In the cell, they lie inside it, and they depend on the reader's outputs. Chapter 26 examines whether this closure is the right criterion for measurement, and where Pattee's metaphor of the cell as an observer stops being useful.

49.4 Norms Without Minds

A cell that swims toward food persists, and a cell that swims away does not. Because the cell's continued existence depends on its readings, the readings can be assessed as better or worse *for the cell*. That is the minimal normativity the enactive tradition calls sense-making: an autonomous system, precarious and self-maintaining, regulates its interactions with its environment according to their consequences for its viability (Di Paolo 2005; Thompson 2007). The attractant is significant for the cell, and the significance is defined by the cell's organization, not by an observer. Jonas (1966) located the origin of this kind of concern in metabolism itself, which must keep exchanging matter with its surroundings to remain what it is.

Lyon's biogenic approach takes such regulation as the starting point for the study of cognition (Lyon 2006), and Levin has argued that the boundaries of the goals a system can pursue, the scale of its "self," are set by the physiological processes that bind its parts into one agent (Levin 2019). These are proposals about where to draw the lower boundary of cognition, and they differ. They agree on the point this rung needs: norms defined by a system's own organization appear with the cell, whatever one then decides to call them.

Normativity of this kind does not require consciousness. Nothing in exact adaptation, integral feedback, or the dependence of survival on motility requires that there be anything it is like to be the bacterium, and none of the theories reviewed in Chapter 44 claims that it does. Chapter 28 makes the same point from the side of the enactive literature.

49.5 Has Meaning Appeared?

Whether the chemotactic reading *means* anything depends on the theory of content one holds. Chapters 29 and 30 develop the theories; here we apply them to one case.

A teleosemantic theory assigns content by function, and function by selection history. The chemotaxis pathway exists because it steered ancestral cells toward nutrients, so its output concerns the direction of nutrients, and a reading that steers the cell toward a non-nutritive chemical that happens to bind the same receptor is a misreading (Millikan 1989; Neander 2017). An informational theory assigns content by what the reading carries information about under normal conditions (F. Dretske 1988), and it can make the same assignment if the normal conditions are specified. A deflationary view holds that the pathway is a mechanism whose outputs covary with concentration changes, and that "meaning" is our description of its role, adding nothing to the mechanism. An enactive view locates meaning in the system's own sense-making, and grants it to the cell on the grounds of the previous section.

The dispute shows up in a concrete case much discussed in the teleosemantics literature. Magnetotactic bacteria contain chains of magnetic particles that align them with the Earth's field; in the northern hemisphere the field points downward as well as north, so swimming along it carries the cell away from oxygen-rich surface water toward the sediment. A bar magnet held nearby can lead the cell the wrong way. Millikan (1989) argues that what the magnetic alignment represents is fixed by what the systems that use it need it to correspond to, the direction of oxygen-poor water, and that the bar magnet therefore produces a misrepresentation. On a purely informational reading, the alignment carries information about the magnetic field and nothing has gone wrong. Which description is right is the indeterminacy problem that Chapter 29 treats in general.

For the ladder we record what is not in dispute. The cell's readings can fail relative to its own organization in a way the radio's cannot, since the radio's failures are failures only relative to its users. Whether such failure deserves to be called misrepresentation is contested, and the ladder records the cell's entry as *contested* when the row appears with the fish (Chapter 50). Meaning in the minimal sense of information for this system has appeared, on every view except the deflationary one. Meaning in any richer sense, such as content the system could entertain or combine with other contents, has not.

49.6 The Second Rung

Table 49.1 adds the cell. The first four rows are re-assessed: the cell separates a chemical signal from background, makes a temporal relation explicit, reads its sign (up or down the gradient), and acts to change the relation it reads. The new rows record what the radio lacks.

Table 49.1: The comparative ladder through the cell. The rows introduced with the radio are re-assessed for the cell, and new rows are added. Entries: *yes*; *no*; *partial*; *derived* (only relative to a designer's or user's purposes); *contested* (the evidence is disputed); *?* (not established by the evidence reviewed in this book).

Capability	Radio	Cell
Introduced with the radio		
Separates a signal from interference	yes	yes
Makes a relational variable explicit	yes	yes
Reads the sign of a relation	yes	yes
Acts to change a relation it reads	yes	yes
Introduced with the cell		
Keeps a memory of its own recent input	yes	yes
Builds and repairs its own readers	no	yes
Its readings matter to its own persistence	no	yes
Shaped by natural selection	no	yes

The first new row is shared. An adaptive radio keeps a record of channel quality, and the cell keeps a record of recent concentration in its methylation state; both memories are what allow the present to be read against the past. The other three new rows separate the two systems, and none of them concerns reading as such. They concern where the reader comes from and what it is for. That is the pattern Part XI will keep finding: each rung's additions are real and measurable, and each leaves open whether it is the one that matters for experience.

49.7 Audit

Canonical phenomenon

Escherichia coli extends its runs when swimming up an attractant gradient, and responds to sudden uniform changes in concentration by suppressing or promoting tumbling before returning to its baseline.

Minimal model

The minimal adaptive reader (49.1): integral feedback yields exact adaptation and an output proportional to the rate of change of concentration (Proposition 49.2); swimming converts a spatial gradient into a temporal one (49.2).

What is relational here?

The relation between the present concentration and the recent past; through (49.2), the relation between the gradient and the cell's own direction of travel.

What physically reads it?

The receptor array and signaling pathway, whose methylation state stores the recent past. The reader is built and maintained by the cell it serves.

What existing field explains it?

Bacterial chemotaxis research (Berg and D. A. Brown 1972; Macnab and Koshland 1972; Barkai and Leibler 1997; Alon et al. 1999), the physics of chemoreception (Berg and Purcell 1977), the theory of autopoiesis and biological autonomy (F. Varela, H. Maturana, et al. 1974; H. R. Maturana and F. J. Varela 1980; Moreno and Mossio 2015), Pattee's semantic closure (Pattee 2001), enactive sense-making (Di Paolo 2005), and teleosemantics (Millikan 1989). Together they explain the mechanism and state the competing views of its significance.

What remains unexplained?

No mechanism is unexplained. What remains is conceptual: whether closure of the reader within the system it serves is what turns a reading into a meaning, or whether it is only a necessary condition for something that appears later. That question is not settled by any fact about *E. coli*.

A discriminating experiment

Compare a natural chemotactic strain with an engineered system that realizes the same reading loop but whose reader is supplied from outside (for instance, a cell-free vesicle carrying purified receptors and motors, or a synthetic cell whose sensing proteins are replenished by the experimenter). Every theory that grants meaning to the natural cell on the basis of closure must withhold it from the engineered system with the same input-output behavior. A theory that does not has located meaning in the loop, not in the closure, and should say so.

49.8 Notes and References

Berg and D. A. Brown (1972) introduced three-dimensional tracking of swimming bacteria and the run-and-tumble description; Macnab and Koshland (1972) established that bacteria detect temporal gradients experienced through their own movement; Berg and Purcell (1977) worked out the physical limits on concentration measurement by a small cell; Barkai and Leibler (1997) proposed that exact adaptation follows from the network's connectivity and needs no fine-tuning, and Alon et al. (1999) confirmed experimentally that the precision of adaptation is robust while other properties are not. J. Adler (1966) is the

classic demonstration of chemotaxis in bacteria. The integral-feedback reading of adaptation connects chemotaxis to the internal model principle of control theory (Francis and Wonham 1976).

On autopoiesis and autonomy, see F. Varela, H. Maturana, et al. (1974), H. R. Maturana and F. J. Varela (1980), and Moreno and Mossio (2015); on closure of constraints, Montévil and Mossio (2015); on the organism as closed to efficient causation, Rosen (1991). Pattee (2001) is the most compact statement of semantic closure; Pattee and Rączaszek-Leonardi (2012) collects the papers. On sense-making and the normativity of life, Di Paolo (2005), Thompson (2007), and Jonas (1966); on the biogenic approach, Lyon (2006); on scale-free cognition, Levin (2019). The theories of content applied in Section 49.5 are developed in Millikan (1989), F. Dretske (1988), and Neander (2017).

49.9 Exercises

Exercise 49.1. Solve (49.1) for a step $c(t) = c_0$ for $t < 0$ and $c(t) = c_1$ for $t \geq 0$, with $h = \ln$ and the system adapted before the step. Show that the peak response is $\ln(c_1/c_0)$, so the cell reads the fold change and not the absolute change. What does this imply about its response to the same absolute increase at low and high background concentrations?

Exercise 49.2. Replace $\tau\dot{\mu} = u$ in (49.1) by $\tau\dot{\mu} = u + \epsilon\,\mu$ with small $\epsilon \neq 0$. Show that adaptation is no longer exact and find the steady-state error after a step. Relate the result to the finding of Alon et al. (1999) that adaptation precision is robust while other properties vary.

Exercise 49.3. Using the counting-noise limit of Berg and Purcell (1977), compare the signal-to-noise ratio of a comparison across a cell of radius a with that of a comparison over the distance vT swum in time T, for a linear gradient. Express the ratio in terms of vT/a and state when temporal comparison wins.

Exercise 49.4. Simulate a run-and-tumble swimmer governed by (49.1) and (49.2) in a linear gradient. Verify that it climbs the gradient on average. Then give the simulated cell a sensor that measures c with no memory (set $\tau \to 0$) and show that it fails. Which relational variable did the memory make available?

Exercise 49.5. List every component of the reading loop in Example 49.3 and state, for each, which process builds and maintains it. Do the same for the adaptive-hopping radio of Chapter 48. Which rows of Table 49.1 follow from your two lists?

Exercise 49.6. For the magnetotactic bacterium of Section 49.5, write down the content of the magnetic alignment according to a teleosemantic theory and according to an informational theory. Describe an experiment on the bacterium whose outcome the two theories would describe differently, and say whether the outcome itself could discriminate between them.

50

The Electric Fish

Part V presented the mechanism: how a weakly electric fish generates the field it senses, reads the beat that a neighbor imposes on it, extracts the sign of a frequency difference, and changes its own frequency to change the relation. This chapter places that mechanism on the comparative ladder. The fish keeps everything the cell had, since it builds, repairs, and depends on its readers, and it adds a group of capabilities that neither the radio nor the bacterium has: it probes the world with its own output, it computes relations in a nervous system, it separates self-generated input from input caused by the world, and it can misread.

The thesis of this chapter is that the electric fish gives the most complete functional account of a reader that this book contains, and that it thereby makes one question unusually sharp. If perspective, in the minimal sense of variables indexed to the system's own action, is what active sensing provides, then the fish has a perspective. Whether there is anything it is like to be the fish is a separate question. The evidence on it is contested, and the fish's value to the argument is that the question can now be attached to an animal whose reading has been mapped from waveform to behavior.

50.1 What the Fish Adds

Each capability the fish adds was established in Part V; here we collect them.

1. *Active sensing. Eigenmannia* emits a continuous, quasi-sinusoidal electric organ discharge (EOD) at a few hundred hertz and senses the distortions of that field through electroreceptors in its skin (Heiligenberg and Bastian 1984; Heiligenberg 1991a). Objects near the fish change the field, and the fish locates them from the change (Lissmann and Machin 1958; Emde 1999; Caputi and Budelli 2006). The carrier it reads is its own (Chapter 19).
2. *Relational computation in a nervous system.* When a neighbor discharges near the fish's frequency, the fish must determine the sign of the frequency difference $\Delta f = f_e - f_s$, and it does so in identified circuits of the hindbrain, midbrain, and diencephalon (Chapters 20 and 21).

3. *Reafference.* The input the fish receives is a joint function of its own discharge and the world. Electric fishes show both of the solutions that Chapter 22 distinguishes: pulse-type mormyrids cancel predictable self-generated input with a corollary discharge and a plastic "negative image" (C. C. Bell 1981; C. Bell et al. 1997; Requarth and Sawtell 2011), while the jamming avoidance response of wave-type fishes needs no internal copy of the motor command at all (Heiligenberg 1991a; Kawasaki 1993).
4. *Error.* Combinations of amplitude and phase modulation that involve no neighbor at all can evoke frequency shifts with the tuning, sensitivity, and dynamics of the jamming avoidance response (Carlson and Kawasaki 2007). Chapter 23 examines whether this is misreading.
5. *World-directed action.* The jamming avoidance response changes the fish's own frequency so that the difference from its neighbor grows, which restores the carrier the fish needs for electrolocation.

The fish also shows how such a reader comes to exist. The jamming avoidance response is innate: a correct response emerges in young *Eigenmannia* at 1.2 to 1.6 cm body length, a developmental age of 24 to 32 days; fish raised in complete isolation develop it on the same schedule; and it emerges even under a stimulus regimen that rewards shifts in the wrong direction (Viete and Heiligenberg 1991). In a few fish an incorrect response, with shifts in the wrong direction, appeared before the correct one, which the authors attribute to initial errors in the wiring: a reader can be built wrong before it is built right. And the same computational rules evolved independently in the African *Gymnarchus niloticus*, implemented in different brain regions (Kawasaki 1993; Kawasaki 1996), a convergence that Chapter 24 uses to separate the algorithm from its implementation.

50.2 The Loop, Written Out

The reading loop $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$ of Chapter 15 can be written out for the fish element by element, with a source for each (Table 50.1). No other system in the book has every element identified at this level.

Two features of the table matter for the ladder.

First, the reading is distributed. Many patches of skin contribute, and the behavior emerges from a large population of imprecisely tuned neurons with no single critical bottleneck, which Heiligenberg called a neuronal democracy (Heiligenberg, Baker, et al. 1978; Heiligenberg and G. Rose 1985). The population achieves what no single neuron does: the fish detects timing differences between regions of its body at least as small as 400 ns (G. Rose and Heiligenberg 1985b; C. Carr et al. 1986).

Second, the sign of Δf is carried only by a relation between two relations. At a patch with contamination ratio m, the fish's own field and its neighbor's combine to give an amplitude modulation $A(t) \approx A_{\mathrm{s}}(1+m\cos\psi(t))$ and a phase

Table 50.1: The jamming avoidance response of *Eigenmannia* as a reading loop.

Element	Implementation	Sources
Mixing M	Superposition of the fish's own field and its neighbor's at each patch of skin, with a contamination ratio m that differs from patch to patch	Heiligenberg (1991a)
Transduction	P-units encode amplitude; T-units fire one phase-locked spike per EOD cycle and encode timing	Scheich et al. (1973)
Reading D	Separate pathways through the electrosensory lateral line lobe to the dorsal torus semicircularis; differential phase computed in lamina 6; deeper neurons combine amplitude and phase and are selective for the sign of Δf	Metzner and Heiligenberg (1991), G. Rose and Heiligenberg (1985a), C. Carr et al. (1986), Heiligenberg and G. Rose (1985), and G. Rose and Heiligenberg (1986b)
Policy π	Nucleus electrosensorius: stimulation of its "increase" and "decrease" subdivisions produces the slow rises and falls of the response	C. H. Keller et al. (1990)
Intervention F	Two motor pathways converge on the medullary pacemaker in a push-pull arrangement, one raising and one lowering the discharge frequency	Metzner (1993b)
Closure	The new f_s changes Δf, and with it the next mixture	Madhav et al. (2013)

modulation $\varphi(t) \approx m \sin \psi(t)$, where $\psi(t) = \Delta\omega\, t + \psi_0$ and the approximations hold for small m. To this order, neither modulation alone reveals the sign of $\Delta\omega$ (the amplitude never does; Chapter 21 shows that the phase carries a trace of it only at higher order). With two patches, the more contaminated one labeled 2, the differential phase $\delta = \varphi_2 - \varphi_1 \approx (m_2 - m_1) \sin \psi$ and the amplitude A_2 trace a closed loop, clockwise when $\Delta f > 0$ and counterclockwise when $\Delta f < 0$, with differential phase on the horizontal axis and amplitude on the vertical, and

$$\operatorname{sgn} \Delta f = -\operatorname{sgn}\langle \dot{A}_2\, \delta \rangle. \tag{50.1}$$

Chapter 21 derives the rule; it is the rotation criterion that Chapter 48 proved for quadrature channels, obtained here without an internal reference. When the amplitude at the more contaminated patch rises while that patch leads, the neighbor is lower and the fish raises its frequency; when the amplitude rises while that patch lags, the neighbor is higher and the fish lowers its frequency (Metzner 1993b; Heiligenberg and G. Rose 1986).

Example 50.1 (Reading the sign from two patches). Take illustrative contamination ratios $m_1 = 0.05$ and $m_2 = 0.2$ and a neighbor 4 Hz above the fish, so $\Delta\omega = 2\pi \times 4/\text{s}$. Then $A_2 \approx A_s(1 + 0.2\cos\psi)$ and $\delta \approx 0.15\sin\psi$, and averaging over a beat cycle gives $\langle \dot{A}_2\, \delta \rangle \approx -\frac{1}{2} A_s (0.2)(0.15)\Delta\omega < 0$. By (50.1), $\Delta f > 0$, and the fish should lower its frequency, which is what *Eigenmannia* does. Reversing the neighbor's offset reverses the sign of the average and of the response. Neither $A_2(t)$ alone nor $\delta(t)$ alone would have changed in any way the fish could detect: each is a sinusoid at the same rate in both cases.

The dynamics of the closed loop can be summarized by a model whose parameters are illustrative and whose qualitative shape is the empirical one.

Model 50.2 (Illustrative closed-loop jamming avoidance)**.** Let the fish's frequency move against the sign of the difference, with gain $k > 0$ and sensorimotor delay d:

$$\dot{f}_s(t) = -k\, g(\Delta f(t-d)), \qquad g(x) \approx x \text{ near } 0, \quad g(x) \to 0 \text{ for large } |x|. \tag{50.2}$$

For a neighbor at constant frequency, $\dot{\Delta f} = -\dot{f}_s$, so the difference grows away from zero: $\Delta f = 0$ repels. Madhav et al. (2013) tested the real loop by clamping the stimulus frequency to follow the fish's own and found exactly this character: near the equilibrium the behavior was fitted by a delayed first-order linear model, the equilibrium was unstable, and the global dynamics were nonlinear, including a "snap-through" bifurcation as a function of Δf.

50.3 Perspective in the Minimal Sense

The cell's reading contained its own motion (Chapter 49). The fish's reading contains its own output. Every variable the fish reads is defined relative to its discharge: the amplitude and phase modulations are modulations *of its own carrier*, and the differential phase is a comparison *between regions of its own body*. The measurement frame is centered on the fish by construction, before any model of a self is involved. This is the minimal agent-centered measurement frame of Chapter 22, a set of variables of the form "world relative to my action."

Definition 50.3 (Minimal functional perspective)**.** A system has a *minimal functional perspective* on a set of variables if each variable is defined relative to the system's own output or body, so that it changes when the system changes what it does, and if the system's behavior depends on those variables.

The definition is deliberately functional. It asks nothing about experience, and it is satisfied by any reader whose readings are indexed to its own action and used.

On this rung, then, perspective in a functional sense has appeared. The fish's relevant variables are indexed to the fish, and indexed physically, by

the geometry of its body and the timing of its own discharge. Nothing more is claimed. Chapters 34 and 37 show how much further a perspective can be articulated: allocentric maps, temporal centering, self-models, the logic of indexicals. And Chapter 43 states the gap that separates any functional perspective, however elaborate, from a phenomenal one.

The fish's solution also shows that a perspective need not include a copy of one's own command. A reader with an internal reference, like the quadrature receiver of Chapter 48, reads the world against a clock it carries. The wave-type fish reads the world against itself in a different way: its two views of the mixture come from two parts of its body, and the relation between them carries what it needs (Kawasaki 1993). The frame is the body.

50.4 Is There Anything It Is Like to Be the Fish?

Nagel (1974) chose the bat because echolocation is a sense unlike any of ours, and argued that the facts of what it is like to be a bat are not accessible from the outside. The electric fish sharpens the example. Electroreception has no human counterpart at all, and we understand the fish's reading of its electric world better than we understand almost any other sensory computation in any animal. If complete functional knowledge settled the phenomenal question, it would be settled here.

It is not settled, and the evidence divides. Fishes have nociceptors: Sneddon et al. (2003) identified receptors in the rainbow trout responsive to noxious stimuli and described behavioral changes after noxious stimulation. Braithwaite (2010) argues from such evidence that fish probably feel pain. Key (2016) argues that fish lack the neural architecture required for felt pain, so that nociception in fish proceeds without experience. Feinberg and Mallatt (2016) argue that sensory consciousness arose with the earliest vertebrates and is present in fishes. The New York Declaration on Animal Consciousness, signed in 2024 by scientists and philosophers working on the question, holds that the evidence indicates "at least a realistic possibility" of conscious experience in all vertebrates, fishes included (*The New York Declaration on Animal Consciousness* 2024). Birch et al. (2020) recommend replacing the single question with several, since consciousness may vary along distinct dimensions (perceptual richness, evaluative richness, integration at a time, integration across time, and self-consciousness) and a species may have a distinctive profile across them.

We take no side the evidence does not support. The book's contribution is to attach the question to the architecture. If there is something it is like to be *Eigenmannia*, which piece of Table 50.1 matters? The candidates can be listed. Active sensing supplies an agent-centered frame; Merker (2005) argues that consciousness presents a stable world from which the confounding effects of self-produced motion have been removed, and that it arose as a solution

to problems of decision making in mobile animals with centralized brains, a description the fish fits. The midbrain combines amplitude and phase from many body regions into one decision, an integration that theories emphasizing integrated causal structure would weigh. The loop is closed through the fish's own viability, which the enactive tradition would weigh. And the fish lacks, as far as the evidence reviewed here shows, the self-models and metacognitive capacities that higher-order theories require. If there is nothing it is like to be the fish, what is missing? The next rung, Chapter 51, lists the candidates, and Chapter 44 asks each theory which one it means.

Two limits hold whichever way the question is answered. The jamming avoidance response shows that a relation can be read by a physical system without an internal reference; it does not show that anything is experienced. And an adaptive radio realizes the same input-output function (*Specification of the Bluetooth System, Version 1.2* 2003; Haykin 2005), so whatever the fish has that the radio lacks cannot be the input-output function (Chapter 48).

50.5 The Third Rung

Table 50.2 adds the fish.

Table 50.2: The comparative ladder through the electric fish. Entries: *yes*; *no*; *partial*; *derived* (only relative to a designer's or user's purposes); *contested* (the evidence is disputed); *?* (not established by the evidence reviewed in this book).

Capability	Radio	Cell	Fish
Introduced with the radio			
Separates a signal from interference	yes	yes	yes
Makes a relational variable explicit	yes	yes	yes
Reads the sign of a relation	yes	yes	yes
Acts to change a relation it reads	yes	yes	yes
Introduced with the cell			
Keeps a memory of its own recent input	yes	yes	yes
Builds and repairs its own readers	no	yes	yes
Its readings matter to its own persistence	no	yes	yes
Shaped by natural selection	no	yes	yes
Introduced with the electric fish			
Probes the world with its own output	no	partial	yes
Separates self-caused from external input	no	no	yes
Computes relations in a nervous system	no	no	yes
Can misread relative to its own organization	derived	contested	yes

The entries for the new rows need comment. The radio receiver does not probe; a radar, which emits the pulses whose echoes it measures, would earn a *yes*, and that is a reminder that active sensing is easy to engineer. The cell earns a *partial* because its reading exists only through its own motion (Chapter 49). The fish's *yes* for separating self-caused from external input is carried by electric fishes as a group: mormyrids cancel reafference with a corollary discharge, and the wave-type jamming avoidance response achieves the separation without one. The fish's *yes* for misreading rests on phantom jamming, whose status as error Chapter 23 examines; the radio's errors are errors only relative to its users, and whether the cell's failures count as misrepresentation depends on the theory of content (Chapter 49).

Every one of the new rows describes something that has been measured. None of them, by itself, tells us whether there is a subject.

50.6 Audit

Canonical phenomenon

Eigenmannia, meeting a neighbor that discharges near its own frequency, raises its frequency if the neighbor is lower and lowers it if the neighbor is higher, using only the mixture of the two fields on its skin.

Minimal model

The two-patch model of amplitude and differential-phase modulation and the sign rule (50.1); the illustrative closed-loop model (50.2), whose unstable equilibrium matches the closed-loop experiments of Madhav et al. (2013).

What is relational here?

The frequency difference Δf; its sign, carried by the sense of rotation of amplitude against differential phase; and the measurement frame itself, which is defined relative to the fish's own discharge and body.

What physically reads it?

P- and T-unit afferents, the electrosensory lateral line lobe, sign-selective neurons of the torus semicircularis, and the nucleus electrosensorius, which drives two motor pathways to the pacemaker (Table 50.1). The fish grows and maintains all of them.

What existing field explains it?

Neuroethology of electric fishes (Heiligenberg 1991a; Metzner 1993b; Kawasaki 1993), the theory of reafference and corollary discharge (Crapse and Sommer 2008), control theory (Madhav et al. 2013), and the comparative literature on animal consciousness (Sneddon et al. 2003; Braithwaite 2010; Key 2016; Feinberg and Mallatt 2016; Birch et al. 2020; *The New York Declaration on Animal Consciousness* 2024). The mechanism is fully explained by the first three.

What remains unexplained?

Whether there is anything it is like to be the fish, and, if there is, which element of the loop matters. The functional account does not answer either question, and no theory reviewed in Chapter 44 derives an answer from the functional account alone.

A discriminating experiment

Ask each theory of consciousness to predict, in advance, how its preferred measure would differ between a fish performing the jamming avoidance response and the same fish under a manipulation that leaves the response intact while removing the property the theory names (for example, disrupting multisensory integration in the midbrain while sparing the electrosensory pathway, or clamping the loop open as in Madhav et al. (2013)). A theory whose prediction does not change under such a dissociation has not located experience in the property it names.

50.7 Notes and References

The facts in this chapter follow the sources collected for Part V: Heiligenberg (1991a) for the jamming avoidance response of *Eigenmannia* and its neural basis; Metzner (1993b) for the motor pathways; Kawasaki (1993) and Kawasaki (1996) for *Gymnarchus*; Carlson and Kawasaki (2007) for phantom jamming; Madhav et al. (2013) for the closed-loop dynamics; and Viete and Heiligenberg (1991) for the development of the response. On active electrolocation, see Lissmann and Machin (1958), Emde (1999), Caputi and Budelli (2006), and M. E. Nelson and MacIver (2006); on reafference in mormyrids, C. C. Bell (1981), C. Bell et al. (1997), and Requarth and Sawtell (2011).

Nagel (1974) is the classic argument that the character of an alien sense is not accessible from outside. The contested literature on fish experience includes Sneddon et al. (2003), Braithwaite (2010), Key (2016), Feinberg and Mallatt (2016), and the New York Declaration (*The New York Declaration on Animal Consciousness* 2024); Birch et al. (2020) propose the multidimensional framework, and Godfrey-Smith (2016) gives an accessible account of how the question looks from the evolution of other minds. Merker (2005) ties the origin of consciousness to the logistics of decision making in mobile animals whose own motion confounds their senses.

50.8 Exercises

Exercise 50.1. Using the small-m expressions for A_2 and δ, verify (50.1) and show that the trajectory (δ, A_2) circles clockwise when $\Delta f > 0$. Then show that the quantity $\delta\,\dot{A}_2 - \dot{\delta}\,(A_2 - A_{\mathrm{s}})$ is constant in time, and relate it to the rotation criterion of Chapter 48.

Exercise 50.2. Simulate (50.2) with $g(x) = x\,\mathrm{e}^{-x^2/(2\sigma^2)}$, a neighbor at constant frequency, and a small initial difference. Plot $\Delta f(t)$ for several delays d. Where does the difference settle, and why does it not grow without bound? Compare the qualitative behavior with the findings of Madhav et al. (2013).

Exercise 50.3. In an open-loop experiment the neighbor's frequency is fixed by the experimenter; in the closed-loop experiment of Madhav et al. (2013) it follows the fish's own frequency at a clamped difference. Explain why the closed-loop design can reveal an unstable equilibrium that an open-loop design cannot, and state which element of the reading loop the clamp replaces.

Exercise 50.4. Specify an adaptive radio whose mapping from interferer frequency to frequency change matches (50.2). List every row of Table 50.2 on which the radio and the fish differ, and for each, describe an experiment that could establish the difference without reference to behavior at the level of that mapping.

Exercise 50.5. Using the five dimensions proposed by Birch et al. (2020), write down what evidence would be needed to place *Eigenmannia* on each. For which dimensions does the neuroethology reviewed in Part V provide any relevant evidence at all?

Exercise 50.6. The chapter claims that the fish's measurement frame is agent-centered "by construction." Make the claim precise: define the frame as a set of variables and show that each is invariant under a common shift of the phase of both fields but changes when the fish changes its own discharge. Which variable in your set would an observer with an external clock read that the fish cannot?

51

The Mammalian Brain

A rat choosing between two arms of a maze pauses at the junction, looks one way and then the other, and while it does so, cells in its hippocampus fire in sequences that run ahead of it along the paths it has not yet taken. A rat that has just made an uncertain choice waits less long for a reward than one that has made an easy choice, and cells in its frontal cortex fire in proportion to how uncertain the choice was. A monkey with damage to its primary visual cortex can point to a light in its blind field and yet, given the option, classifies that light as "no light." Each of these animals reads its world through the kind of loop the electric fish showed in full. Each also reads something the fish, as far as the evidence reviewed in this book shows, does not: what is absent, how reliable its own readings are, and how its body stands.

The thesis of this chapter is that the capabilities the mammalian brain adds to the ladder are well documented, that each of them has been proposed by some theory as the ingredient that matters for consciousness, and that none of them is treated as necessary by every theory. Several have engineered analogs, and several are found in animals far from mammals. The question "which of these additions matters for phenomenality?" therefore has no answer independent of a theory of consciousness. It has a map, and this chapter draws it.

51.1 Integrating Several Senses in One Model

The posterior parietal cortex of monkeys represents space in several coordinate frames at once: locations defined relative to the eye, the head, the body, and the limb, combined from visual, auditory, somatosensory, and proprioceptive inputs and used to plan movements (Andersen, Snyder, et al. 1997). The representations are relational in the sense of Chapter 1: a location relative to the hand is a relation between two positions, and the transformations among frames are computed, not given. Chapter 34 develops the geometry.

When two senses report the same quantity, the brain must decide how much to trust each. Across many tasks, humans and other animals combine cues

in close to the way an ideal observer would, weighting each by its reliability (Knill and Pouget 2004). The ideal combination has a simple form.

Theorem 51.1 (Reliability-weighted combination)**.** *Let two cues report a quantity s as $\hat{s}_1 = s + \varepsilon_1$ and $\hat{s}_2 = s + \varepsilon_2$, with independent zero-mean errors of variance σ_1^2 and σ_2^2. Among unbiased linear combinations $\hat{s} = w\hat{s}_1 + (1-w)\hat{s}_2$, the variance is minimized by*

$$w^* = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \qquad \mathrm{Var}(\hat{s}) = \Big(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\Big)^{-1}. \tag{51.1}$$

Proof. $\mathrm{Var}(\hat{s}) = w^2\sigma_1^2 + (1-w)^2\sigma_2^2$. Setting the derivative $2w\sigma_1^2 - 2(1-w)\sigma_2^2$ to zero gives w^*, and substituting gives the stated variance. □

The weight in (51.1) depends only on the ratio of the two variances. A system that combines cues this way reads a relation between the reliabilities of its own readers and uses it to set how much each reader counts. That is a second-order relational variable, a relation among readings rather than among things in the world. It is also cheap to engineer: sensor-fusion filters do the same arithmetic. What is distinctive in the mammalian case is that the variances are estimated and updated by the same system, from its own experience, in dozens of modalities at once, and folded into a single model of where things are.

51.2 Hierarchical Models of the World

Predictive processing treats the cortex as a hierarchy of generative models, each level predicting the activity of the level below and passing forward only the error of its prediction (K. Friston 2010; Clark 2013). On this view the variables a mammal reads are mostly its own predictions, corrected by the senses where they fail. Chapter 33 presents the formal framework and its limits.

For the ladder, one consequence matters. A system with a generative model reads the world through relations between what it expected and what arrived. The fish reads relations between its own field and a neighbor's, both present in the mixture on its skin. The mammal adds a relation whose first term is internal and whose content is a guess about what should be there. The difference is a difference in what can be absent: a prediction can concern a state of the world that does not now obtain, and the prediction error can then drive learning or action.

51.3 Simulating What Is Absent

Tolman (1948) argued from the behavior of rats in mazes that they learn a map of the environment and not only a chain of responses. The hippocampal

place cells discovered later made the map a physiological object (O'Keefe and Nadel 1978). The map can be read when the animal is not at the place it represents.

Rats at a choice point sometimes pause and orient back and forth between the options, a behavior called vicarious trial and error because it resembles deliberation (Redish 2016). During such pauses, ensembles of hippocampal CA3 cells transiently encode paths ahead of the animal, sweeping first down one arm and then down the other (Johnson and Redish 2007). In a delay period spent running in a wheel, hippocampal cell assemblies fire in internally generated sequences that predict the animal's upcoming choice (Pastalkova et al. 2008). In each case the brain represents a trajectory the animal is not currently following.

Example 51.2 (A detached reading). In the notation of the reading loop, the fish's reader computes $z = \mathsf{D}(y)$ from a mixture y present on its skin. The rat at the junction computes, in addition, a sequence $\hat{R}(t + \Delta t)$ of predicted positions along a path it has not taken, from a stored map and without the corresponding input. The reading is detached from the present stimulus in the sense of Chapter 31: its content concerns what would be sensed if the animal acted otherwise.

Detached reading of this kind is the precondition for choosing among actions by their predicted consequences, which Chapter 32 treats as counterfactual simulation. Whether other animals share it to a degree that matters is an open empirical question, and the ladder records the fish's entry as *?* because the evidence reviewed in this book does not settle it. In humans the neural systems that support remembering the past overlap with those that support imagining the future (Schacter et al. 2007); Suddendorf and Corballis (2007) argue that the full capacity for such mental time travel is uniquely human, a claim that the rodent evidence has kept under debate.

51.4 Models of the Body

The mammalian brain represents its own body as well as the world. Proprioception, the sense of limb position, and interoception, the sense of the physiological condition of the body, feed dedicated pathways; Craig (2002) traced the interoceptive pathway in primates to a cortical representation in the insula. On the view that Seth (2013) develops, the brain's model of the body is a generative model like its model of the world, and emotions are among its predictions.

The body model is a model, and it can be wrong. In the rubber-hand illusion, a person who watches a rubber hand being stroked in synchrony with strokes on their own hidden hand comes to feel the touch in the rubber hand (Botvinick and J. Cohen 1998). The relation that drives the illusion is temporal coincidence between two sensory streams, seen touch and felt touch; when the

relation is present, the body model assigns ownership, and when it is broken by asynchrony, the illusion weakens. The model reads a relation and draws a conclusion about what belongs to the self. Chapters 36 and 41 take up what such models do and whether a model of the self is thereby a self.

51.5 Monitoring One's Own Readings

The adaptive radio of Chapter 48 estimates the quality of its channels, a crude reading about readings. Mammals do far more. Rats trained to categorize odor mixtures near a category boundary are less certain of their choices, and Kepecs et al. (2008) found neurons in the orbitofrontal cortex whose firing tracked a model of decision confidence; when rewards were delayed, rats waited longer after confident choices, so that the confidence estimate governed behavior. Monkeys given the option to decline a difficult perceptual decision in exchange for a smaller sure reward decline more often when the evidence is weak, and neurons in their lateral intraparietal area represent the degree of certainty (Kiani and Shadlen 2009). Comparative studies of uncertainty monitoring find related capacities across several species (Smith et al. 2003). These are the phenomena Chapter 39 calls metacognition.

A dissociation shows how far such monitoring can separate from first-order reading. Monkeys with a lesion of the primary visual cortex on one side can detect and localize stimuli in the affected half of the visual field when a stimulus appears on every trial. Tested in a task that also contained blank trials and a response for "no stimulus," they classified stimuli in the affected field as blanks, while classifying stimuli in the intact field correctly (Cowey and Stoerig 1995). Human patients with such lesions show the same pattern and report that they see nothing, which is why the phenomenon is called blindsight (Weiskrantz 1986). The first-order reading is intact; the second-order classification of that reading as "something is there" is not.

This is the closest the non-human evidence comes to a report about experience. It is not a verbal report, and it does not show that the monkey lacks visual experience in the blind field; it shows that the monkey's own classification of its input treats the blind-field input as it treats no input. The ladder records the mammal's entry for reporting on its own experience as *partial* on this basis, and Chapter 52 returns to what reports can and cannot show.

51.6 Which Additions Matter?

The mammal adds multimodal integration in one model, hierarchical prediction, counterfactual simulation, models of the body, and monitoring of its own readings. Theories of consciousness disagree about which, if any, of these carry the weight. Table 51.1 maps a representative set of positions onto the

capabilities, using each position's own statement of what matters; Chapter 44 examines the theories themselves.

Table 51.1: What several accounts take to matter for consciousness, and where that places the rungs of the ladder. Summaries follow the cited sources.

Account	What it points to
Global workspace (Dehaene and Changeux 2011; Mashour et al. 2020)	Global availability of information across specialized processors through a workspace of long-range cortical connections
Higher-order theories (Lau and D. Rosenthal 2011)	Representations of one's own first-order states; closest to the metacognitive capacities of Section 51.5
Integrated information (Tononi, Boly, et al. 2016)	The integrated causal structure of the physical substrate, which no functional capability on the ladder fixes by itself
Midbrain integration (Merker 2007)	An upper-brainstem system that integrates target selection, action selection, and motivation; the cortex elaborates contents
Egocentric simulation (Barron and Klein 2016)	An integrated, egocentric neural simulation of the mobile animal in space, found in the vertebrate midbrain and in analogous insect structures
Unlimited associative learning (Ginsburg and Jablonka 2019)	The capacity for open-ended associative learning as the marker of a transition to minimal consciousness
Neurobiological features (Feinberg and Mallatt 2016)	A set of features including hierarchical, multisensory, map-like representations, which the authors find in all vertebrates

Three patterns in the table matter for the ladder.

First, the accounts place the lower boundary of consciousness at different rungs. Accounts that locate consciousness in cortical workspaces or in higher-order representations make the mammalian additions central. Accounts that locate it in midbrain integration, in egocentric simulation, or in neurobiological features shared by all vertebrates place it at or below the fish. Barron and Klein (2016) argue that the capacity for subjective experience arose as a solution to problems of sensory reafference and navigation in mobile animals, which is the problem Chapters 22 and 50 found in the electric fish, and they extend it to insects. Merker (2007) argues from children born without most of the cerebral cortex that the upper brainstem can support primary consciousness, a conclusion that remains contested.

Second, several of the capabilities have been found far from mammals. Associative learning of the kind Ginsburg and Jablonka (2019) emphasize, egocentric integration of the kind Barron and Klein (2016) emphasize, and complex behavior in animals whose nervous systems evolved independently of

ours (Godfrey-Smith 2016) all argue against treating the mammalian brain as the only architecture worth asking about. Birch et al. (2020) propose that the question be split into dimensions, so that different animals can have different profiles rather than a place on a single scale.

Third, most of the capabilities have engineered analogs. Reliability-weighted cue combination, predictive models, planning by simulated rollouts, and confidence estimates are all standard in engineering. A theory that locates consciousness in a functional capability alone must say whether it grants consciousness to the engineered analog, as the principle of input-output parity in Chapter 48 required of the radio.

51.7 The Fourth Rung

Table 51.2 adds the mammalian brain.

Table 51.2: The comparative ladder through the mammalian brain. Entries: *yes*; *no*; *partial*; *derived* (only relative to a designer's or user's purposes); *contested* (the evidence is disputed); *?* (not established by the evidence reviewed in this book).

Capability	Radio	Cell	Fish	Mammal
Introduced with the radio				
Separates a signal from interference	yes	yes	yes	yes
Makes a relational variable explicit	yes	yes	yes	yes
Reads the sign of a relation	yes	yes	yes	yes
Acts to change a relation it reads	yes	yes	yes	yes
Introduced with the cell				
Keeps a memory of its own recent input	yes	yes	yes	yes
Builds and repairs its own readers	no	yes	yes	yes
Its readings matter to its own persistence	no	yes	yes	yes
Shaped by natural selection	no	yes	yes	yes
Introduced with the electric fish				
Probes the world with its own output	no	partial	yes	yes
Separates self-caused from external input	no	no	yes	yes
Computes relations in a nervous system	no	no	yes	yes
Can misread relative to its own organization	derived	contested	yes	yes
Introduced with the mammalian brain				
Integrates several senses in one model of the world	no	no	contested	yes
Simulates what is absent	no	no	?	yes
Models its own body and states	no	no	?	yes
Monitors confidence in its own readings	partial	no	?	yes

The fish's entries in the new rows record the limits of the evidence, not a verdict. Integration of several senses into one model is *contested* for the fish because the accounts in Table 51.1 that emphasize midbrain integration and vertebrate neurobiological features attribute it to all vertebrates, while the neuroethology of the jamming avoidance response reviewed in Part V concerns one sense. The other three entries are *?*: the evidence reviewed in this book neither shows nor excludes counterfactual simulation, a body model, or confidence monitoring in electric fish. The radio's *partial* for confidence monitoring is the channel-quality estimate of Chapter 48.

The mammalian rung is the first at which every row the book has introduced is *yes*. That makes it the natural reference class for theories of consciousness, most of which were built on mammalian and especially human data. It also makes it easy to mistake the mammal's full set of capabilities for the minimum that consciousness requires, which is a claim no theory has earned by the ladder alone.

51.8 Audit

Canonical phenomenon

A rat at a maze junction pauses and looks back and forth while hippocampal ensembles sweep ahead along the untaken paths; a rat waits longer for reward after a confident choice; a monkey with a visual cortex lesion localizes a light it classifies as "no light."

Minimal model

Reliability-weighted cue combination (Theorem 51.1); detached readings $\hat{R}(t + \Delta t)$ computed from a stored map (Example 51.2).

What is relational here?

Relations among the reliabilities of the system's own readers; relations between predicted and received input; relations between the animal's position and positions it is not occupying; temporal coincidence between seen and felt touch.

What physically reads it?

Posterior parietal and hippocampal circuits, orbitofrontal and parietal neurons that track confidence, and the pathways that carry interoceptive and proprioceptive signals, all grown and maintained by the animal.

What existing field explains it?

Systems and cognitive neuroscience (Andersen, Snyder, et al. 1997; Johnson and Redish 2007; Kepecs et al. 2008; Kiani and Shadlen 2009), Bayesian models of perception (Knill and Pouget 2004), predictive processing (K. Friston 2010; Clark 2013), the neuropsychology of blindsight (Weiskrantz 1986; Cowey and Stoerig 1995), and the comparative literature on animal

consciousness (Merker 2007; Barron and Klein 2016; Ginsburg and Jablonka 2019; Birch et al. 2020). They explain every capability in the chapter.

What remains unexplained?

Which of the capabilities, if any, is required for there to be something it is like to be the animal. The accounts in Table 51.1 give incompatible answers, and the capabilities themselves do not decide among them.

A discriminating experiment

Choose a pair of accounts that place the boundary at different rungs, for instance a higher-order account and a midbrain-integration account. Find a manipulation in a mammal that removes the capability one account requires while sparing the other (for instance, disrupting confidence monitoring while sparing egocentric integration), and have each account predict, in advance, the effect on a report-independent marker of conscious state. Opposite predictions make the experiment discriminating.

51.9 Notes and References

On multimodal spatial representation in the parietal cortex, see Andersen, Snyder, et al. (1997); on ideal-observer cue combination, Knill and Pouget (2004). Predictive processing is introduced by K. Friston (2010) and Clark (2013). The cognitive map runs from Tolman (1948) to O'Keefe and Nadel (1978); vicarious trial and error and its hippocampal correlates are reviewed by Redish (2016), with the primary findings in Johnson and Redish (2007) and Pastalkova et al. (2008). On prospection and mental time travel, see Schacter et al. (2007) and Suddendorf and Corballis (2007). On interoception, Craig (2002) and Seth (2013); on the rubber-hand illusion, Botvinick and J. Cohen (1998). On metacognition in animals, Kepecs et al. (2008), Kiani and Shadlen (2009), and Smith et al. (2003); on blindsight, Weiskrantz (1986) and Cowey and Stoerig (1995).

The accounts in Table 51.1 are developed in Dehaene and Changeux (2011), Mashour et al. (2020), Lau and D. Rosenthal (2011), Tononi, Boly, et al. (2016), Merker (2007), Barron and Klein (2016), Ginsburg and Jablonka (2019), and Feinberg and Mallatt (2016); Birch et al. (2020) propose the multidimensional framework, and Godfrey-Smith (2016) discusses minds that evolved independently of ours.

51.10 Exercises

Exercise 51.1. Extend Theorem 51.1 to n independent cues. Then suppose the system does not know the variances and must estimate them from its own errors over time. Write an update rule for the weights, and identify which part of your rule reads a relation among the system's own readers.

Exercise 51.2. Specify an engineered system that satisfies every row of Table 51.2 marked *yes* for the mammal except those in the group introduced with the cell. Which rows could you not satisfy without making the system build and repair its own parts? What does your answer suggest about how the rows depend on one another?

Exercise 51.3. In the task of Cowey and Stoerig (1995), the monkey localizes blind-field stimuli when every trial contains a stimulus, and classifies them as blanks when blank trials are included. Model the monkey as a system with a first-order reader and a second-order classifier, and state the minimal assumptions under which your model reproduces both findings. Does your model require any assumption about experience?

Exercise 51.4. Design an analysis that would distinguish hippocampal sequences that represent alternatives being evaluated from sequences that merely replay recent trajectories. Specify the behavioral variable each hypothesis predicts the sequences should correlate with, and the null model you would use.

Exercise 51.5. For each account in Table 51.1, state which rungs of the ladder it grants consciousness, citing the property it relies on. Where the account's stated property is not measured by any row of the ladder, propose a row that would measure it.

Exercise 51.6. The rubber-hand illusion depends on temporal coincidence between seen and felt touch. Model ownership as an inference that weights this coincidence against a prior that one's hand is where it feels to be. Predict how the strength of the illusion should change as the asynchrony grows, and state what result would count against your model.

52

Humans

Everything this book knows about consciousness it learned, in the end, from what people said. Reports of seeing and not seeing, of pain and its absence, of confidence and doubt, are the primary data of consciousness science; the neural measures of Part X were validated against them, and the animal evidence of the previous rungs was interpreted by analogy with them. The capacities that make such reports possible are the ones humans add to the ladder: language, explicit self-reference, a narrative self, recursive metacognition, cultural symbol systems, and reports about experience itself.

The thesis of this chapter is that these additions transform what can be known about consciousness, and that on most theories they reveal consciousness more than they constitute it. They also make it easy to overestimate what reports show. People report fluently on the causes of their own behavior and are often wrong; people with brain damage respond accurately to what they say they cannot see; a patient who cannot move or speak can show, in a scanner, brain activity indistinguishable from that of healthy volunteers following instructions. The last rung of the ladder therefore asks the ladder's hardest question in its clearest form: which capacities reveal consciousness, and which, if any, constitute it?

52.1 Language and Shared Intentionality

Hauser et al. (2002) distinguish the faculty of language in a broad sense, which includes sensory-motor and conceptual systems that other animals share, from the faculty in a narrow sense, and they hypothesize that the narrow faculty consists only of the computational mechanism of recursion, the one component of language they propose is uniquely human. The hypothesis remains debated, but the distinction it draws is useful here. Much of what language requires is found on lower rungs; what may be new is a way of combining finitely many elements into unboundedly many structured expressions.

Tomasello (2014) locates the distinctively human form of thinking elsewhere, in shared intentionality: the capacity to form joint goals and joint attention with others and, later in development, to participate in the collective conventions

and norms of a culture. On this view language is one product of a more basic capacity to take part in shared projects, and it brings with it a public medium in which one person's readings of the world can be made available to another.

For the ladder, language matters in three ways. It gives the system a medium for representing relations of unbounded complexity, including relations among its own states. It makes those representations available to others, so that a reading can be checked, corrected, and accumulated across people and generations. And it provides the channel through which experience is reported, which is the capacity on which the scientific study of consciousness depends.

52.2 "I": Explicit Self-Reference

The fish's measurement frame is centered on the fish by the physics of its sense (Chapter 50); the mammal's body model is centered on the mammal by construction (Chapter 51). Humans can also refer to the center explicitly. Perry (1979) made the point with a story about following a trail of sugar around a supermarket in search of the shopper with the torn bag, until he realized that he was that shopper. Nothing he believed about the world changed when he realized it except a belief whose content he could express only with the word "I," and that belief changed what he did. Chapter 37 develops the logic of such self-locating beliefs.

The fish's variables are indexical without the fish representing that they are. A human can represent the index itself, reason about it, and discover that a description of someone in the world picks out oneself. This is a real addition to the ladder: a system that can represent "the one who is reading" as one of the things it reads about. It is also exactly the kind of recursion that Chapter 41 warns is cheap, since a program can refer to itself. The addition is well defined; its relevance to experience is what theories dispute.

52.3 The Narrative Self

Gallagher (2000) distinguishes a *minimal self*, a sense of oneself as the owner and agent of one's experience and action, present from moment to moment without temporal extension, from a *narrative self*, which involves personal identity and continuity across time and is sustained by memory and by the stories people tell about themselves. The minimal self has candidate precursors on earlier rungs: the fish's agent-centered frame, the mammal's body model and sense of agency. The narrative self is, as far as the evidence shows, human.

Several authors argue that the narrative self is a construction and not a thing. Dennett (1991) describes the self as a center of narrative gravity, an abstraction posited to organize the stories a person tells, useful in the way a center of gravity is useful and equally lacking in independent existence. Metzinger (2003) argues that no such things as selves exist in the world, and

that what exists are self-models whose contents the system experiences as a self because it cannot recognize them as models. Hofstadter (2007) describes the "I" as a self-reinforcing pattern of symbolic self-reference. These are views about what the narrative self is. Whether a narrative self is needed for experience is a different question, and most accounts answer it negatively: people with severe amnesia, who have lost much of the material of their narratives, are not thereby thought to have lost consciousness.

52.4 Recursive Metacognition

Flavell (1979) introduced *metacognition* as knowledge and monitoring of one's own cognitive processes. The mammal of Chapter 51 monitors the confidence of its decisions and acts on the estimate. Humans do so as well, and can also state the estimate, reason about why it might be wrong, and monitor the monitoring: "I was confident, but I am usually overconfident about this kind of judgment." Fleming et al. (2012) review the computational and neural basis of such judgments and the evidence that metacognitive accuracy, how well confidence tracks correctness, varies across people and tasks and can be dissociated from first-order performance.

Definition 52.1 (Orders of reading)**.** A *first-order* reading has a variable of the world as its content. A reading is of order $n+1$ if its content is a property (such as the reliability, the source, or the occurrence) of a reading of order n.

The adaptive radio reached order two in a narrow form, estimating the quality of its channels (Chapter 48). The mammal reaches order two in many domains. Humans reach higher orders and can express them, which is why the higher-order theories of Chapter 40 can be tested on humans at all. Whether reaching some order is what makes a state conscious is the claim of higher-order theories (Lau and D. Rosenthal 2011); whether it is only what makes a conscious state reportable is the objection to them.

52.5 Reports and Their Limits

Reports are indispensable and unreliable. Nisbett and Wilson (1977) reviewed evidence that people often lack introspective access to the processes behind their judgments and choices: subjects can be unaware of a stimulus that influenced a response, or unaware that it did, and when asked to explain their behavior they offer explanations that match those that outside observers would guess from common-sense theories of behavior. The reports are sincere, fluent, and frequently wrong about causes. Schwitzgebel (2008) extends the doubt from causes to contents, arguing that people are unreliable even in describing their current conscious experience, for example the clarity of peripheral vision or the character of their emotions.

Dennett (2003) proposes a method for using reports without trusting them: heterophenomenology treats a subject's reports as data about what the subject believes and says about their experience, to be explained, without assuming in advance that the beliefs are true or false. The method has been criticized for building in a verdict against phenomenal consciousness, and defended as neutral; either way it separates two things the naive use of reports runs together, the report and what it is about.

Two dissociations show how far report and experience can come apart, in opposite directions.

Example 52.2 (Response without reported experience). Patients with damage to the primary visual cortex report that they see nothing in part of the visual field, yet when required to guess they can detect, localize, and discriminate stimuli there well above chance (Weiskrantz 1986). The first-order reading is present and used; the report denies experience. Monkeys with the corresponding lesion show a parallel pattern in a nonverbal task (Cowey and Stoerig 1995, see also Chapter 51). Whether blindsight involves no experience, or degraded experience that the patient does not classify as seeing, remains debated, and the debate turns on how much weight the report should carry.

Example 52.3 (Experience without behavioral report). A patient who met the clinical criteria for the vegetative state was asked, in a scanner, to imagine playing tennis and to imagine moving through the rooms of her home. Her patterns of brain activity were indistinguishable from those of healthy volunteers given the same instructions (Owen et al. 2006). She made no behavioral report, yet her brain followed spoken instructions over extended periods. The minimally conscious state had already been defined as a condition of limited but reproducible evidence of awareness (Giacino et al. 2002); covert command-following of this kind extends the evidence beyond what behavior shows.

Together the examples show that report is neither necessary for the best evidence of consciousness nor sufficient for accurate knowledge of it. The capacity to report is a reader of the system's own states, and like every reader in this book it makes some variables explicit and can misread.

52.6 Reveal or Constitute?

Each human addition can be asked the same question: does it reveal consciousness, as a window reveals a room, or does it help constitute consciousness, as the walls constitute the room? The major positions answer differently.

Block (1995) distinguishes access consciousness, information available for report, reasoning, and the control of action, from phenomenal consciousness, what it is like to be in a state, and argues that the two can come apart. On this view reports and metacognition belong to access and reveal phenomenal states that exist independently of them. Higher-order theories hold that a state is

conscious when it is the object of a suitable higher-order representation (Lau and D. Rosenthal 2011), so recursive metacognition is close to constitutive, although the higher-order state need not be linguistic. Chalmers (1996b) argues that no functional capacity, human or otherwise, entails phenomenal consciousness, so the human additions reveal at most, and the question of what constitutes experience lies beyond function. Illusionism holds that what needs explaining is our strong disposition to judge that we have phenomenal properties (Frankish 2016); on this view the human capacity for introspective report is the very thing a theory of consciousness must explain, and it opens no window onto properties beyond itself.

The ladder cannot choose among these positions. It can show what each commits to. If report constitutes consciousness, the mammal and everything below it are unconscious, a conclusion almost no one accepts. If report only reveals it, then consciousness may extend down the ladder as far as the property that constitutes it, and that property has not been named by the report. If illusionism is right, the human rung is where the explanandum lives, and the lower rungs pose a different question: which of them has the kind of self-model that generates such judgments.

52.7 The Complete Ladder

Table 52.1 completes the comparison.

Four lessons follow from reading the table as a whole.

First, the reading loop is the ladder's floor. Every rung, engineered or living, has the four capabilities of the first group. Reading relations, including their sign, and acting to change them is common and cheap, and no theory of consciousness treats it as sufficient.

Second, each group of capabilities can occur without the groups introduced after it, while the later groups presuppose some of the earlier ones. The radio has the reading loop without building its own readers, and the cell builds its own readers without a nervous system; but no system in the table monitors the reliability of its readings without first having readings. The ladder is cumulative in the sense of Part XI's design: each rung adds capacities that presuppose earlier ones.

Third, the entries that are not *yes* or *no* mark where knowledge stops. The *contested* and *?* entries in the fish's column are where comparative neuroscience has most to do, and they are where the accounts of Chapter 51 disagree about the lower boundary of consciousness.

Fourth, and most important, no row of the table is labeled "experience." Every row is a capability that can be defined and measured. Each theory of consciousness points to one or more of them, or to a property of internal organization that no row captures, and says that experience goes with it. The

Table 52.1: The complete comparative ladder. Entries: *yes*; *no*; *partial*; *derived* (only relative to a designer's or user's purposes); *contested* (the evidence is disputed); *?* (not established by the evidence reviewed in this book).

Capability	Radio	Cell	Fish	Mammal	Human
Introduced with the radio					
Separates a signal from interference	yes	yes	yes	yes	yes
Makes a relational variable explicit	yes	yes	yes	yes	yes
Reads the sign of a relation	yes	yes	yes	yes	yes
Acts to change a relation it reads	yes	yes	yes	yes	yes
Introduced with the cell					
Keeps a memory of its own recent input	yes	yes	yes	yes	yes
Builds and repairs its own readers	no	yes	yes	yes	yes
Its readings matter to its own persistence	no	yes	yes	yes	yes
Shaped by natural selection	no	yes	yes	yes	yes
Introduced with the electric fish					
Probes the world with its own output	no	partial	yes	yes	yes
Separates self-caused from external input	no	no	yes	yes	yes
Computes relations in a nervous system	no	no	yes	yes	yes
Can misread relative to its own organization	derived	contested	yes	yes	yes
Introduced with the mammalian brain					
Integrates several senses in one model of the world	no	no	contested	yes	yes
Simulates what is absent	no	no	?	yes	yes
Models its own body and states	no	no	?	yes	yes
Monitors confidence in its own readings	partial	no	?	yes	yes
Introduced with humans					
Refers to itself in language	no	no	no	no	yes
Tells a narrative about itself	no	no	no	no	yes
Reports on its own experience	no	no	no	partial	yes

ladder does not show where experience begins. It shows which property each theory must point to, and what evidence would test the choice.

52.8 Audit

Canonical phenomenon

People report on their experience in language; the reports are sometimes wrong about causes, and in blindsight and covert awareness report and experience appear to come apart in opposite directions.

Minimal model

Orders of reading (Definition 52.1): a report is a reading of order two or higher whose content is a property of first-order readings, subject like any reading to error.

What is relational here?

Relations between the system and its own states: the reliability of a reading, the source of an action, the continuity of a life, and, in explicit self-reference, the identity between "the one described" and "the one describing."

What physically reads it?

Language, metacognitive monitoring, and memory-based self-narration, grown by each person and shaped by a culture that supplies the words.

What existing field explains it?

Linguistics and comparative cognition (Hauser et al. 2002; Tomasello 2014), philosophy of the self (Gallagher 2000; Dennett 1991; Metzinger 2003), research on metacognition (Flavell 1979; Fleming et al. 2012), the psychology of introspection (Nisbett and Wilson 1977; Schwitzgebel 2008), neuropsychology (Weiskrantz 1986; Owen et al. 2006), and philosophy of consciousness (Block 1995; Chalmers 1996b; Frankish 2016). They explain the capacities and state the dispute.

What remains unexplained?

Which capacities constitute consciousness and which reveal it. The positions of Section 52.6 give incompatible answers, and no fact about the human additions decides among them.

A discriminating experiment

Use a paradigm in which first-order performance is matched across two conditions while metacognitive access differs (as in blindsight, or in masking designs that equate accuracy while varying confidence), and measure a report-independent marker of conscious state. Higher-order accounts predict that the marker follows metacognitive access; accounts on which report only reveals predict that it follows first-order processing. Preregistered predictions from both make the outcome informative either way.

52.9 Notes and References

On the faculty of language and the recursion hypothesis, see Hauser et al. (2002); on shared intentionality, Tomasello (2014). The essential indexical is from Perry (1979). On the minimal and narrative self, see Gallagher (2000); on the self as a construction, Dennett (1991), Metzinger (2003), and Hofstadter (2007). Metacognition was introduced by Flavell (1979); Fleming et al. (2012) review its computational and neural basis. The classic case against the accuracy of reports about the causes of behavior is Nisbett and Wilson

(1977); Schwitzgebel (2008) argues that reports about current experience are unreliable too; Dennett (2003) defends heterophenomenology and Dennett (1991) applies it. On blindsight, Weiskrantz (1986) and Cowey and Stoerig (1995); on covert awareness and disorders of consciousness, Owen et al. (2006) and Giacino et al. (2002). The distinction between access and phenomenal consciousness is from Block (1995); the argument that no functional capacity entails phenomenal consciousness is developed in Chalmers (1996b); illusionism is stated by Frankish (2016); higher-order theories are reviewed by Lau and D. Rosenthal (2011).

52.10 Exercises

Exercise 52.1. Classify each of the following by its order in the sense of Definition 52.1: a thermostat's reading of temperature; the adaptive radio's estimate of channel quality; a rat's confidence estimate in Kepecs et al. (2008); a person's report "I was confident, but I am usually overconfident about this." At which order does each system in Table 52.1 stop, as far as the evidence shows?

Exercise 52.2. Design an experiment in the style of Nisbett and Wilson (1977) in which a stimulus influences a choice and participants are asked why they chose. Specify how you would compare the participants' explanations with those of observers who did not take part, and what pattern of results would show that the reports were based on common-sense theories rather than access to the process.

Exercise 52.3. Model blindsight as a system with an intact first-order reader and a second-order classifier whose input from the first-order reader is attenuated. Show how the model can produce above-chance forced-choice performance together with a report of "nothing seen." Then state what additional finding would be needed to decide whether the patient has degraded experience or none.

Exercise 52.4. The covert-awareness paradigm of Owen et al. (2006) infers consciousness from task-appropriate brain activity in response to instructions. List the assumptions this inference requires. Which of them would a proponent of each position in Section 52.6 accept, and which would they reject?

Exercise 52.5. For each position in Section 52.6, mark on Table 52.1 the rows it treats as constitutive of consciousness, the rows it treats as merely revealing it, and the rows it treats as irrelevant. Which position places the boundary of consciousness lowest on the ladder, and on what row does it rely?

Exercise 52.6. Propose one additional row for Table 52.1 that is not a capability of reading, remembering, modeling, or reporting, but a property of internal organization that some theory treats as decisive. Define it so that it can be measured, and fill in the five entries with the evidence you would need for each.

Radio, cell, fish, mammal, human: each step adds capacities that can be named and measured. Self-maintenance appears with the cell, agent-centered measurement with active sensing, counterfactual simulation and metacognition with the mammalian brain, report and narrative with language. The ladder does not show where experience begins. It shows which property each theory would have to point to, and that is the more useful result. Part XII turns the book's method on the book's own name.

Part XII

Is There a Metaxic Principle?

Does anything remain after subtraction?

53

Candidate Unifications

Fifty-two chapters have produced a list. A relation can be a variable of a configuration; a physical operation can make it explicit; the operation can be implemented in a causal structure that supports the right counterfactuals; an animal can produce the carrier it reads; a loop can close through action; a system can build and maintain its own reader; the variables it reads can be centered on itself; and a reader can take other readers as its input. Seven ingredients recur throughout: relational variables, endogenous readout, counterfactual sensitivity, active sensing, closure, agent-centered variables, and recursion. This chapter asks whether any combination of them is a new theoretical object, or whether the combinations are lists with names.

The answer is the second, and the chapter shows why case by case. Every pair of ingredients we examine already has an owner, often a theorem decades old. The best candidate the book has produced, a reader of reafferent relations that works without an internal reference and is built by the system it regulates, is a precise conjunction with instances in more than one substrate, and the one generalization it supports belongs to signal theory. A conjunction defines a class, and a class is useful for comparison. A new theoretical object needs a law of its own.

53.1 Seven Ingredients

Table 53.1 collects the ingredients with the chapters that develop them and the fields that already study them. The working definitions are those of the chapters cited; nothing is redefined here.

A candidate unification combines some of these and claims that the combination does more than its parts.

Definition 53.1 (Candidate unification)**.** A *candidate unification* is a proposed theoretical object defined as a conjunction of two or more ingredients, together with the claim that the conjunction supports inferences that no ingredient supports alone.

Four tests decide whether the claim holds.

Ingredient	Working definition (where developed)	Field that studies it
Relational variable	A function of a configuration and of no single component (Chapter 1)	Physics, dynamical systems
Endogenous readout	A physical operation inside the system that makes a relational variable dynamically explicit (Chapters 15 and 16)	Measurement, signal processing
Counterfactual sensitivity	The readout's causal organization mirrors the state transitions it is said to implement (Chapter 17)	Philosophy of computation
Active sensing	The system generates part of the signal it senses (Chapters 19 and 22)	Sensorimotor neuroscience, neuroethology
Closure	The reader is built and maintained by the system whose behavior it regulates (Chapters 26 and 27)	Theoretical biology
Agent-centered variables	Read variables whose origin is the system itself (Chapters 34 and 37)	Cognitive science, philosophy of language
Recursion	Readers whose inputs include other readers (Chapters 39 and 41)	Metacognition research, philosophy of mind

Table 53.1: The seven ingredients that recur through the book.

Principle 53.2 (Tests for a genuine unification)**.** A candidate unification is genuine to the extent that:

1. it is defined precisely enough that membership can be decided for a given system;
2. it has instances in at least two substrates that share no mechanism;
3. it supports a counterfactual generalization about its instances that is false of systems lacking one of its ingredients;
4. that generalization cannot be derived from an existing theory that owns one of the ingredients.

The first three tests are easy to pass with careful definitions and a good comparative biology. The fourth is where candidates fail, and it is the book's method in compressed form: subtract the established explanation, then look at what remains.

53.2 Unifications That Already Exist

Several combinations of the ingredients have been unified before, some by theorem and some by framework. Any claim that Metaxics unifies them has to say what it adds to these.

53.2.1 The good regulator

Conant and Ashby considered a regulator that acts on a system in order to keep outcomes within a desired set, measured optimality by the entropy of the outcomes, and proved that a regulator that is both optimal and as simple as possible must be a model of the system it regulates, in the precise sense that its behavior is a mapping of the system's states (Conant and Ross Ashby 1970). The title of the paper states the conclusion: every good regulator of a system must be a model of that system. The theorem joins two of our ingredients, endogenous readout and closure through regulation, and derives a representational conclusion from a control premise. Its assumptions are narrow (optimality defined by outcome entropy, a deterministic regulator, maximal simplicity), and the theorem supports much less than its title once they are dropped. Within its assumptions it is exact, and it was published more than half a century before this book.

53.2.2 The internal model principle

Francis and Wonham proved a sharper result for linear control: a feedback regulator that tracks references and rejects disturbances robustly, in the sense that small changes in the plant do not destroy regulation, must contain in its feedback path a model of the dynamics that generate those exogenous signals (Francis and Wonham 1976).

Example 53.3 (Integral action and oscillators). A controller that must hold a regulated variable at a set point despite a constant disturbance of unknown size needs integral action: an integrator is a model of the dynamics $\dot{d} = 0$ that generate constant signals. A controller that must reject a sinusoidal disturbance of known frequency ω_0 and unknown amplitude and phase needs an undamped oscillator at ω_0 in its loop, for the same reason. The controller reproduces the generator of what it must cancel.

The principle sorts two electric fish that Part V kept apart. Pulse-type mormyrids cancel the predictable sensory consequences of their own discharges with a corollary discharge whose effect in the electrosensory lobe is shaped by experience into a negative image of the expected input (C. C. Bell 1981; C. Bell et al. 1997; Requarth and Sawtell 2011). That negative image is an internal model in Francis and Wonham's sense: a learned replica of the dynamics generating the signal to be cancelled. The wave-type jamming avoidance response contains no such model. *Gymnarchus* has no corollary discharge reaching its sensory areas, and *Eigenmannia* computes the response from the sensory mixture alone (Kawasaki 1994; Heiligenberg 1991a). The difference follows from the task. The mormyrid must remove its own signal to hear the world; the jamming avoidance response uses the relation between its own signal and the world's, and removing either term would destroy the variable it needs (Chapter 22).

53.2.3 Semantic closure, active inference, and sensorimotor contingencies

Three frameworks cover larger combinations. Pattee's semantic closure joins endogenous readout, counterfactual sensitivity, and closure: an organization that constructs its own measuring devices and uses their results to maintain itself (Pattee 1995; Pattee 2001). Active inference joins active sensing, closure, agent-centered variables, and world models under the claim that perception, action, and self-maintenance all minimize a single quantity, variational free energy (K. Friston 2010); its proponents present it as a normative framework rather than as one empirical model (Parr et al. 2022). Sensorimotor contingency theory joins active sensing with agent-centered variables and adds a claim about experience: that seeing is the exercise of mastery over the lawful ways sensation changes with movement (O'Regan and Noë 2001).

Existing unification	**Ingredients joined**	**Form**
Good regulator (Conant and Ross Ashby 1970)	Endogenous readout, closure	Theorem
Internal model principle (Francis and Wonham 1976)	Endogenous readout, closure, prediction of exogenous signals	Theorem
Semantic closure (Pattee 2001)	Endogenous readout, counterfactual sensitivity, closure	Framework
Active inference (Parr et al. 2022)	Active sensing, closure, agent-centered variables, world models	Framework
Sensorimotor contingencies (O'Regan and Noë 2001)	Active sensing, agent-centered variables, a claim about experience	Theory
Self-model theory (Metzinger 2003)	Recursion, agent-centered variables, a claim about experience	Theory

Table 53.2: Existing unifications of the book's ingredients.

Table 53.2 shows the coverage. Active inference alone contains four of the seven ingredients. Whatever Metaxics claims to unify, it must say what it adds to a framework that already claims to unify most of the list.

53.3 Ladders That Have Already Been Drawn

Part XI compared radio, cell, fish, mammal, and human, asking what changes at each step. The form of that comparison is borrowed, and the borrowing should be explicit.

Braitenberg's *Vehicles* builds a sequence of imaginary machines, each a few sensors wired to a few motors, and shows that crossed or uncrossed, excitatory or inhibitory connections produce behavior that an observer is tempted to describe as fear, aggression, or love (Braitenberg 1984). His lesson, which he called the law of uphill analysis and downhill invention, is that building a mechanism that produces a behavior is much easier than inferring the mechanism from the behavior. Dennett's Tower of Generate-and-Test sorts organisms by how they generate and test behavioral options: Darwinian creatures by natural selection among fixed designs, Skinnerian creatures by reinforcement of what worked, Popperian creatures by testing options against an internal model before acting, and Gregorian creatures by using culturally inherited thinking tools (Dennett 1995; Dennett 1996). Ginsburg and Jablonka propose a ladder of learning capacities and place a transition to minimal consciousness at the emergence of unlimited associative learning (Ginsburg and Jablonka 2019).

Ladder	**What varies along it**	**Proposed threshold**
Braitenberg's vehicles	Wiring between sensors and motors	None; the point is the observer's temptation
Dennett's tower	How behavioral options are generated and tested	None for consciousness
Ginsburg and Jablonka	Learning capacity	Unlimited associative learning
This book (Part XI)	What relation is read, by what reader, in which loop	None; each theory is asked to name one

Table 53.3: Comparative ladders. The axis of this book's ladder is new; the form is not.

Table 53.3 places the book's ladder among them. Its axis differs from the others: it varies the relation read and the loop that reads it, where Braitenberg varies wiring, Dennett the generation of options, and Ginsburg and Jablonka learning. The difference is useful, because it lets the ladder ask each theory of consciousness to point at a rung. A new axis on an old instrument is a contribution to method; a ladder does not become a law by acquiring an axis.

53.4 Testing a Conjunction

The strongest candidate the book has produced combines three ingredients that Part V found together in one animal.

Definition 53.4 (Reference-free reafferent reader)**.** A reading loop $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$ is a *reference-free reafferent reader* if:

1. the read variable $z = \mathsf{D}(y)$ depends on a relation between a signal the system produces and a signal from outside it;

2. D receives no copy of the motor command that produces the system's own signal;
3. D, π, and F are built and maintained by the system itself.

The definition passes the first test: for a given system, each clause can be checked. Table 53.4 checks five systems.

System	Relation read (clause 1)	Motor copy to the reader? (clause 2)	Self-built? (clause 3)
Eigenmannia jamming avoidance	Sign of Δf between own and neighbor's discharge	No (Heiligenberg 1991a)	Yes
Gymnarchus jamming avoidance	Same, by the same rules	No; no corollary discharge to sensory areas (Kawasaki 1993; Kawasaki 1994)	Yes
Horseshoe bat Doppler-shift compensation	Echo frequency relative to the emitted call (Schnitzler 1968; Schnitzler and Denzinger 2011)	Vocal-motor signals reach neurons implicated in compensation (Metzner 1993a)	Yes
Mormyrid electrosensory lobe	Own discharge's sensory consequences	Yes; corollary discharge (C. C. Bell 1981)	Yes
Adaptive frequency-hopping radio	Interference on the channels it uses (*Specification of the Bluetooth System, Version 1.2* 2003)	Yes; it holds its own hop sequence	No

Table 53.4: Instances and non-instances of the reference-free reafferent reader.

The horseshoe bat is the instructive neighbor. It emits calls with a long constant-frequency component, and in flight the echo returns Doppler-shifted upward; the bat lowers the frequency of its subsequent calls so that echoes return near a reference frequency, where its cochlea and auditory pathway devote an expanded, sharply tuned representation, the auditory fovea (Schnitzler 1968; Schnitzler and Denzinger 2011). Like the fish, it reads a relation between its own signal and a returned one and changes its own signal to control that relation. Unlike the fish, it does not read the relation reference-free: neurons in the paralemniscal tegmentum respond differently to the bat's own vocalizations than to acoustically identical playback, which indicates that vocal-motor signals reach neurons implicated in compensation (Metzner 1993a). The bat also controls the relation in the opposite direction: it holds the echo frequency near a set point, where the fish drives $|\Delta f|$ away from zero.

The second test is passed by the two electric fish, whose electrosensory systems evolved independently and implement the same computational rules with different brain structures (Kawasaki 1996; Kawasaki and Y. Guo 1996). The third and fourth tests need a generalization. The natural candidate concerns what any member of the class must compute.

Proposition 53.5 (Envelopes do not carry the sign)**.** *Let a receptor site receive* $y(t) = A_{\mathrm{s}}\,[\cos \omega_{\mathrm{s}} t + m \cos(\omega_{\mathrm{e}} t + \psi_0)]$ *with* $0 < m < 1$*. The envelope* $A(t) = A_{\mathrm{s}} \sqrt{1 + m^2 + 2m \cos(\Delta\omega\, t + \psi_0)}$ *is unchanged under* $(\Delta\omega, \psi_0) \mapsto (-\Delta\omega, -\psi_0)$*. Any reader whose input is the envelope at one or more sites therefore cannot recover* $\operatorname{sgn} \Delta f$*.*

Proof. The envelope depends on $\Delta\omega$ and ψ_0 only through $\cos(\Delta\omega\, t + \psi_0)$, and cosine is even. □

The sign is recoverable once the reader has a second observable in quadrature with the envelope. With an internal reference for the phase of its own carrier, the phase modulation at one site suffices. Without one, the differential phase between two sites with different contamination ratios suffices, and the joint behavior of amplitude and differential phase then carries the sign at every instant (Chapter 21). This gives the candidate generalization:

Conjecture 53.6 (A law of reference-free readers)**.** *Every reference-free reafferent reader of the sign of a frequency difference combines at least two observables in quadrature, taken at receptor sites with different mixing ratios.*

The conjecture passes the third test: it is false of readers with an internal reference, which can recover the sign from a single site. It fails the fourth. A real signal has a spectrum symmetric in frequency, so the sign of a frequency offset can be identified only from a pair of observables that together form a complex (analytic) representation (Oppenheim and Schafer 2010); this is the reason communication receivers that must tell a positive frequency offset from a negative one carry in-phase and quadrature channels. Appendix B derives the symmetry. The fish's solution is a biological realization of a requirement that signal theory states for any substrate, and the conjecture adds a biological instance to a known theorem.

The verdict on the strongest candidate is therefore mixed in a precise way. The class it defines is real, since it separates the fish from the bat, the mormyrid, and the radio along a line that matters for how the reading is done. Its law belongs to signal processing.

53.5 Recursion and the Self

The remaining ingredient, recursion, has been unified with the self by theories that Part IX examined: higher-order theories, self-model theory, and attention

schema theory (Chapters 40 to 42). Hofstadter's strange loop makes self-reference the organizing principle of the self (Hofstadter 2007), and Metzinger's self-model theory makes the transparency of a system's model of itself the source of the experience of being a self (Metzinger 2003). Recursion is cheap in engineering: a state estimator that also estimates the covariance of its own errors reads one of its own readings, and nobody takes that to be a self. A unification that adds recursion to the other six ingredients inherits the burden that Part IX placed on these theories, which is to say what property beyond recursion does the work.

53.6 The Verdict

Table 53.5 applies the four tests of Principle 53.2 to the candidates this book has met.

Candidate (ingredients)	**Owner**	**Tests met**	**Verdict**
Physical readout (relational variable, readout)	Signal processing, measurement	1, 2, 3	Owned
Implementation (readout, counterfactual sensitivity)	Philosophy of computation	1, 2, 3	Owned
Reafference (active sensing, readout)	Neuroethology	1, 2, 3	Owned
Semantic closure (readout, closure)	Theoretical biology	1, 2	Owned
Reference-free reafferent reader (Definition 53.4)	Signal theory owns its law	1, 2, 3	Useful class, borrowed law
Active inference (four ingredients)	Computational neuroscience	1, 2	Owned; generality disputed
Self-reading (recursion, agent-centered variables)	Philosophy of mind, metacognition research	1, 2	Owned

Table 53.5: Candidate unifications against the four tests.

No candidate passes all four tests. The pattern is consistent: each candidate is either a combination already made by someone else or a class whose governing generalization is derivable from a theory that owns one of its parts. This is the result the book's method predicts, and it is a useful result, because it states where a claim of novelty would have to be made. It would have to be a generalization about closed reading loops that is false of open ones and cannot be derived from control theory, or an experimental operation whose

outcomes no owner predicts. Chapter 56 states those requirements as criteria, and Chapter 58 proposes the operation most likely to meet one of them.

53.7 Audit

Canonical phenomenon

Two electric fish that evolved their electric sense independently read the sign of a frequency difference between their own discharge and a neighbor's, without a copy of their own motor command; a horseshoe bat reads the relation between its call and its echo with one.

Minimal model

Candidate unifications as conjunctions of ingredients (Definition 53.1), four tests (Principle 53.2), the reference-free reafferent reader (Definition 53.4), and the sign-blindness of envelopes (Proposition 53.5).

What is relational here?

The frequency difference Δf between a self-produced signal and an external one, and its sign; in the bat, the difference between call and echo frequency.

What physically reads it?

In the fish, midbrain (*Eigenmannia*) or hindbrain (*Gymnarchus*) circuits that combine amplitude and differential phase; in the bat, audio-vocal circuits that also receive vocal-motor signals.

What existing field explains it?

Control theory (the good regulator and the internal model principle), signal processing (quadrature representation of frequency offsets), neuroethology (jamming avoidance, Doppler-shift compensation, corollary discharge), theoretical biology (semantic closure), computational neuroscience (active inference), and comparative ladders from Braitenberg, Dennett, and Ginsburg and Jablonka.

What remains unexplained?

The class of reference-free reafferent readers separates systems that other classifications lump together. No law specific to the class has been found; the one it supports is a theorem of signal theory.

A discriminating experiment

Search for a counterexample to Conjecture 53.6: an animal that resolves the sign of a frequency offset from a single receptor site without an internal reference. Finding one would falsify the conjecture as stated and would mean that the animal exploits structure the two-sinusoid model omits, such as the waveform asymmetry of the phase modulation at large contamination (Chapter 21). The discriminating question would then be whether signal theory, applied to the richer signal, predicts the animal's performance.

53.8 Notes and References

The good regulator theorem is due to Conant and Ross Ashby (1970); the internal model principle to Francis and Wonham (1976). Both belong to the cybernetic and control-theoretic traditions discussed in Chapter 54, and both are cited far more often than their assumptions are stated. The mormyrid negative image is described by C. C. Bell (1981) and reviewed by C. Bell et al. (1997) and Requarth and Sawtell (2011); its status as an internal model is argued in Chapter 22. Semantic closure is developed in Pattee (1995) and Pattee (2001); active inference in K. Friston (2010) and, as a textbook, Parr et al. (2022); sensorimotor contingency theory in O'Regan and Noë (2001).

The comparative ladders are Braitenberg (1984), Dennett (1995), Dennett (1996), and Ginsburg and Jablonka (2019). Doppler-shift compensation in horseshoe bats was described by Schnitzler (1968); Schnitzler and Denzinger (2011) review its relation to the auditory fovea; Metzner (1993a) recorded the audio-vocal neurons that distinguish self-produced from played-back calls. The convergence of the jamming avoidance response in *Gymnarchus* and *Eigenmannia* is established by Kawasaki (1993) and reviewed by Kawasaki (1996). The spectral symmetry of real signals is standard; Oppenheim and Schafer (2010) is a reference text.

53.9 Exercises

Exercise 53.1. A plant obeys $\dot{x} = -ax + u + d(t)$ with $a > 0$ and disturbance $d(t) = D\sin\omega_0 t$ of unknown amplitude D. Show that proportional-integral control $u = -k_p x - k_i \int x\,\mathrm{d}t$ leaves a steady-state oscillation in x. Add to the controller an undamped oscillator driven by x and show, by computing the closed-loop transfer function from d to x, that the steady-state error at ω_0 vanishes. Relate the result to the internal model principle.

Exercise 53.2. For a thermostat, *Escherichia coli* chemotaxis (Macnab and Koshland 1972), horseshoe bat Doppler-shift compensation, the *Eigenmannia* jamming avoidance response, and a phone's adaptive frequency-hopping radio, mark which of the seven ingredients of Table 53.1 each system has. Which pairs of systems share the most ingredients, and do they share mechanisms?

Exercise 53.3. For $y(t) = \cos\omega_\mathrm{s} t + m\cos(\omega_\mathrm{e} t + \psi_0)$ with small m, show that the phase modulation relative to the fish's own carrier is $\varphi(t) \approx m\sin(\Delta\omega\, t + \psi_0)$ and that the instantaneous-frequency deviation $\dot{\varphi}$ is approximately $m\Delta\omega\cos(\Delta\omega\, t + \psi_0)$. Compute the time average of $\dot{\varphi}\,(A - A_\mathrm{s})$ and show that its sign is $\operatorname{sgn}\Delta\omega$. What kind of internal reference does a reader need in order to measure $\dot{\varphi}$ at a single site, and how does this route compare with the differential-phase route of Chapter 21?

Exercise 53.4. Place *Eigenmannia*, a horseshoe bat, and a human on Dennett's tower and on Ginsburg and Jablonka's learning ladder. Identify one place

where the two ladders disagree about which of two organisms is "higher," and state what each ladder is measuring at that point.

Exercise 53.5. Design a Braitenberg vehicle with two receptors on opposite sides of its body, a self-generated oscillating field, and a motor that sets its own oscillation frequency, such that the vehicle moves its frequency away from that of a nearby vehicle. Specify the receptor signals and the wiring. Which clause of Definition 53.4 does your vehicle fail, and could it be made to pass?

Exercise 53.6. Write the jamming avoidance response as gradient descent of the fish's own frequency on a cost function of Δf. State the cost function, derive the rule for the sign of $\dot{f}_{\mathrm{s}}$, and discuss whether this rewriting is a prediction of active inference or a redescription of the behavior. What observation would distinguish the two?

54

Failed Grand Unifications

Metaxics is not the first attempt to find one science of structure beneath many substrates. General system theory, catastrophe theory, the edge of chaos, cybernetics, and several unifications built on the word "information" each promised to connect physics, biology, and mind through a shared form, and each attracted able people. Most did not become the science they promised. Their failures were specific, and each failure yields a test that a new unification can apply to itself before others apply it.

The thesis of this chapter is that grand unifications have failed in six identifiable ways. General system theory catalogued analogies without theorems of its own. Catastrophe theory had a genuine theorem and applied it where its control parameters could not be measured. The edge of chaos rested on a suggestive result that did not survive replication. Cybernetics produced theorems and engineering, and its umbrella dissolved while those survived. Information metaphysics equivocates among distinct concepts that share a word. Theories of consciousness have often fixed a measure before validating it. The last section turns the six tests on this book.

54.1 General System Theory

Ludwig von Bertalanffy argued that the same general principles recur across fields that study organized wholes, and that a general system theory could state those principles once (Bertalanffy 1968). His evidence was a set of *isomorphies*: the same law of exponential growth or decay in fields as different as population biology and radioactivity, the same logistic law of saturating growth in several fields, and the behavior of open systems, which exchange matter with their surroundings and can reach the same final state from different initial conditions, a property he called equifinality.

The isomorphies are real. They are also what one expects from a small stock of simple equations. A quantity whose rate of change is proportional to itself grows or decays exponentially whatever the quantity is; add a saturating term and the logistic curve follows. The recurrence is a fact about which differential equations are simplest, and mathematics explains it without a

science of systems. What general system theory did not produce was a theorem whose hypotheses mention systems as such and whose conclusion could be false of a system that satisfies them. Without one, it remained a program and a vocabulary.

Principle 54.1 (First test: a theorem of one's own)**.** A unification must state at least one result whose hypotheses are expressed in its own terms and whose conclusion does not follow from the mathematics its instances already share.

Chapter 53 applied this test to the book's best candidate and found that its law belongs to signal theory.

54.2 Catastrophe Theory

Catastrophe theory began with a real theorem. René Thom showed that for families of gradient systems depending on at most four control parameters, the generic singularities of the potential are equivalent, after smooth changes of coordinates, to one of seven elementary forms (Thom 1975). The simplest with interesting behavior is the cusp.

Model 54.2 (The cusp)**.** A state x moves down the gradient of the potential

$$V(x; a, b) = \tfrac{1}{4}x^4 + \tfrac{1}{2}ax^2 + bx, \tag{54.1}$$

where a and b are control parameters. Equilibria satisfy $x^3 + ax + b = 0$. For $4a^3 + 27b^2 < 0$ there are three equilibria, two stable and one unstable; outside that region there is one. The boundary $4a^3 + 27b^2 = 0$ is the bifurcation set.

The cusp explains, in one picture, three qualitative features: bistability (two stable states for the same control values), hysteresis (the state at given control values depends on the path taken to them), and sudden jumps (as b is swept at fixed $a < 0$, the occupied equilibrium disappears at the fold and the state jumps to the other branch). In bifurcation theory the cusp survives as a normal form whose parameters are physical parameters of a model, and there it is a working tool (Kuznetsov 2004).

E. C. Zeeman applied the cusp far beyond physics: to animal aggression, with fear and rage as the two control parameters of a dog's behavior, to stock-market crashes, to prison disturbances, and to other social and biological phenomena (Zeeman 1977). The applications drew a detailed critique from Sussmann and Zahler (1978), who argued that the models assumed potentials and smoothness that nothing in the phenomena justified, that their control parameters were not measured, and that qualitative agreement with bistability and jumps was cheap because many mechanisms produce those features. The dispute was reported in *Science* under the headline that the emperor had no clothes (Kolata 1977).

The lesson concerns the step from theorem to application. Thom's classification is exact, and its hypotheses are about potentials and control parameters. An application that cannot identify and measure the control parameters has not satisfied the hypotheses; it has matched the theorem's qualitative consequences to a phenomenon that shares them. The proposal that consciousness is a moiré effect has exactly this structure. Moiré patterns are real and fully explained (Chapter 2); experience is unified, slow relative to its substrate, and settled by competition; and the resemblance licenses nothing until the gratings, the mismatch, and the reading operation are identified in a brain and measured.

Principle 54.3 (Second test: measurable control parameters)**.** An application of a unifying mathematical form must identify, and be able to measure, the variables that play the roles of the form's parameters. Qualitative agreement with the form's consequences is not evidence that the form applies.

54.3 The Edge of Chaos

Christopher Langton studied cellular automata through a single parameter λ, the fraction of entries in a rule table that map to a state other than a designated quiescent one. As λ increases from zero, typical rules move from frozen or periodic behavior toward chaotic behavior, and Langton argued that the rules capable of supporting complex computation cluster near the transition, at the "edge of chaos" (Langton 1990). Norman Packard used a genetic algorithm to evolve cellular automata for a density-classification task and reported that the evolved rules moved toward the critical values of λ (Packard 1988). Together these results suggested a general principle: systems that compute, or that evolution shapes to compute, sit near a phase transition between order and chaos.

Mitchell et al. (1993) repeated the evolutionary experiment and could not reproduce the result. The rules their genetic algorithm evolved for the same task clustered close to $\lambda = 1/2$, the value that the symmetry of the density task under exchange of 0s and 1s favors, and not near the critical regions that had been reported. They concluded that the original interpretation was not correct: for this task, the experiment gave no support to the claim that evolution drives computation toward the edge of chaos. Exercise 54.3 shows one reason: rules with the same λ can behave in entirely different ways.

The idea did not disappear. Its descendant in neuroscience, the hypothesis that cortex operates near a critical point, commits to measurable signatures such as scale-free distributions of neuronal avalanches (Beggs and Plenz 2003; Chialvo 2010), and it makes testable claims about unconsciousness as a departure from critical dynamics (Tagliazucchi et al. 2016). It remains debated. The difference between the descendant and the original slogan is the difference the chapter is about: a measured signature, stated in advance, that can come out the wrong way.

Principle 54.4 (Third test: replication and re-analysis)**.** A result that motivates a unification must survive independent replication and re-analysis with the null models appropriate to it before it carries weight outside the model class in which it was found.

54.4 Cybernetics: A Partial Success

Norbert Wiener proposed a single science of control and communication in animals and machines, organized around feedback (Wiener 1948). W. Ross Ashby gave it precise results. The law of requisite variety states that a regulator can reduce the variety of outcomes only by matching it with variety of its own: only variety can destroy variety (Ashby 1956). In one information-theoretic form, under Ashby's assumptions about how disturbance D, regulator response R, and outcome E are related,

$$H(E) \geq H(D) - H(R), \tag{54.2}$$

so a regulator with too few states cannot hold outcomes steady against a rich enough disturbance, however it is wired. The good regulator theorem followed (Conant and Ross Ashby 1970) (Chapter 53).

Cybernetics is the partial success among the unifications in this chapter, and its fate is instructive. The parts that are still taught became theorems or engineering: feedback control, information theory, bounds such as (54.2). The ambition to be one science of control and communication across animal and machine did not survive as a discipline; its results were absorbed by control theory, information theory, computer science, and neuroscience, each of which kept the part it could use.

Principle 54.5 (Fourth test: survival as theorems)**.** A unification should expect to survive, if at all, as the theorems and operations it contributes to existing fields. Its name is the part most likely to be lost.

The test is a forecast for this book. If Metaxics contributes anything durable, the likeliest form is an experimental operation or a result absorbed into neuroethology, the philosophy of computation, or consciousness science, and not a discipline under its own name.

54.5 Information Metaphysics

The word *information* carries at least four distinct concepts, and unifications built on it tend to slide among them.

1. *Statistical information*, Shannon's measure, a property of ensembles of messages and of channels. Shannon bracketed meaning at the outset: "Frequently the messages have meaning; that is they refer to or are

correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem" (Shannon 1948, p. 379).

2. *Physical information*, the claim that stored and processed information is always carried by physical degrees of freedom, so that manipulating it has physical costs (Landauer 1991).
3. *Semantic information*, the content of a signal: what it is about. Dretske built an account of content on statistical information and needed further conditions to get meaning from correlation (F. I. Dretske 1981) (Chapter 29).
4. *Intrinsic information*, a quantity defined over a system's own causal structure and, in integrated information theory, identified with consciousness (Tononi 2004; Albantakis et al. 2023). Chalmers proposed, as explicit speculation, that information has a physical and a phenomenal aspect (Chalmers 1996b, ch. 8).

Each concept is precise in its home. Integrated information theory, for example, defines its quantity carefully and states its identity claim openly, which makes it a theory to be tested (Chapter 44), whatever one thinks of the claim. The failure mode arises when an argument moves between senses without marking the move: it borrows rigor from the first sense, where theorems exist, and significance from the fourth, where the phenomenon of interest lives. A sentence such as "the brain integrates information, and consciousness is integrated information" can be read four ways, and it is true under different readings in each half.

This book is exposed to the same slide. Chapter 28 speaks of information *for* a system; Chapter 16 of variables available to a reader; Chapter 44 of integrated information. Each use has to say which sense it means.

Principle 54.6 (Fifth test: one sense per use)**.** Every use of "information" (and of its relatives: reading, representation, meaning) must specify which concept it names, and an argument that changes senses must say where.

54.6 Premature Theories of Consciousness

Theories of consciousness face a special difficulty: the phenomenon is accessible to measurement only through report and behavior, so any proposed neural signature can be validated only against other signatures. The difficulty invites a pattern in which a measure is proposed, found to correlate with the state of consciousness in some conditions, and treated as confirmation of the theory that motivated it.

Crick and Koch proposed in 1990 that synchronized oscillations in the gamma range might bind the features of a visual object into a single conscious percept (Crick and Koch 1990). They framed it as a step toward a theory, and it served for years as a target for experiments. Many theories have

since accumulated evidence, each largely on its own terms and rarely tested directly against the others (Seth and Bayne 2022; Cogitate Consortium et al. 2025). The first large preregistered adversarial collaboration, which pitted integrated information theory against global neuronal workspace theory with predictions agreed by their proponents in advance, recorded 256 participants with functional magnetic resonance imaging, magnetoencephalography, and intracranial electroencephalography (Melloni et al. 2023; Cogitate Consortium et al. 2025). Its results aligned with some predictions of each theory and challenged key tenets of both: the absence of sustained synchronization within posterior cortex counted against integrated information theory, and the general absence of ignition at stimulus offset, with limited prefrontal representation of some conscious content, counted against the workspace theory (Cogitate Consortium et al. 2025).

The lesson is procedural. Before the adversarial test, each theory had evidence; after it, each had a documented failure of a stated prediction. A theory of consciousness becomes a scientific theory at the point where its proponents commit, in advance, to outcomes that would count against it.

Principle 54.7 (Sixth test: preregistered adversarial prediction)**.** A unification that makes claims about consciousness must state, before the data are collected, which outcomes would count against it, and must be tested against a rival that predicts differently.

Chapter 61 designs such tests for the relational hypotheses of this book.

54.7 What the Failures Teach

Table 54.1 collects the six tests and records where Metaxics stands on each.

The table is sobering and useful. The book passes the tests it can pass by construction (measured parameters for its model system, a vocabulary rule), and it has not yet faced the tests that decide whether a unification is a science (a theorem of its own, a replicated result, a preregistered prediction that could fail). The failures of the past do not show that the attempt is hopeless. They show where the attempt will be judged.

54.8 Audit

Canonical phenomenon

Four historical programs (general system theory, catastrophe theory, the edge of chaos, cybernetics) and two recurrent patterns (information metaphysics, premature theories of consciousness), each proposing one account of structure across substrates.

Unification	What was genuine	What failed	Metaxics on the same test
General system theory	Cross-field analogies; open systems	No theorem of its own	Fails so far; its best law is owned (Chapter 53)
Catastrophe theory	Thom's classification	Applications without measured control parameters	Passes for the fish (Δf is set and measured); untested for consciousness
Edge of chaos	A suggestive result in one model class	Did not survive replication	No result yet to replicate
Cybernetics	Feedback, requisite variety, the good regulator	The umbrella dissolved	Likely fate; plan for it
Information metaphysics	Four precise concepts	Equivocation among them	Rule adopted (Principle 54.6)
Consciousness theories	Proposals that directed experiments	Measures fixed before validation	Tests designed, not run (Chapter 61)

Table 54.1: Six ways grand unifications have failed, and the corresponding test for this book.

Minimal model

The cusp catastrophe (54.1); Langton's λ; Ashby's requisite-variety bound (54.2).

What is relational here?

In each program, a relation among variables claimed to recur across substrates: a shared differential equation, a shared singularity, a shared position near a phase transition, a shared feedback structure.

What physically reads it?

Mostly the theorist: in the failed programs, the recurrence was detected by an outside observer comparing descriptions. Cybernetics is the exception, because feedback is a relation that the regulated system itself uses.

What existing field explains it?

History and philosophy of science: the critiques of catastrophe theory (Sussmann and Zahler 1978; Kolata 1977), the re-analysis of the edge of chaos (Mitchell et al. 1993), and the methodology of research programmes discussed in Chapter 56.

What remains unexplained?

None in the history. For the book, the residue is the set of tests it has not

yet faced.

A discriminating experiment

Apply Principle 54.3 to the claim that consciousness is a relational pattern among neural processes: name the relata, the mismatch, and the reading operation, and state a measurement that could show them absent in a conscious subject. If no such measurement can be specified, the claim is a Zeeman-style fit.

54.9 Notes and References

General system theory is set out in Bertalanffy (1968). Catastrophe theory is due to Thom (1975); its applications to biology and the social sciences are collected in Zeeman (1977); the critique is Sussmann and Zahler (1978), and the controversy is reported in Kolata (1977). The cusp as a normal form in modern bifurcation theory is treated in Kuznetsov (2004). The edge-of-chaos hypothesis is Langton (1990) and Packard (1988); the re-analysis is Mitchell et al. (1993); Mitchell (2009) gives a broader history of complexity science. Criticality in neural systems is introduced by Beggs and Plenz (2003) and reviewed by Chialvo (2010).

The founding texts of cybernetics are Wiener (1948) and Ashby (1956); the good regulator theorem is Conant and Ross Ashby (1970). On the several senses of information: Shannon (1948) for the statistical theory, Landauer (1991) for its physical embodiment, F. I. Dretske (1981) for semantic content, and Tononi (2004) and Albantakis et al. (2023) for integrated information; Chalmers (1996b, ch. 8) presents the double-aspect speculation. On theories of consciousness and their testing, see Crick and Koch (1990), Seth and Bayne (2022), the protocol of Melloni et al. (2023), and the results of Cogitate Consortium et al. (2025).

54.10 Exercises

Exercise 54.1. For the cusp (54.1), show that the equilibrium equation $x^3 + ax + b = 0$ has three real roots exactly when $4a^3 + 27b^2 < 0$. Fix $a = -3$ and compute the values of b at which the state jumps as b is swept upward and then downward. Sketch the hysteresis loop.

Exercise 54.2. Zeeman's model of canine aggression uses fear and rage as control parameters. Propose operational measures for both, describe an experiment that would test the model's prediction of hysteresis, and state an outcome that would falsify the model. Then do the same for the claim that consciousness is a moiré effect.

Exercise 54.3. For elementary cellular automata (two states, three-cell neighborhoods, quiescent state 0), λ is the fraction of the eight neighborhoods mapped to 1. Compute λ for rules 30, 90, 110, and 184 from their binary

expansions. Simulate each from a random initial condition and describe its behavior. What does the comparison say about λ as a predictor of behavior?

Exercise 54.4. A disturbance takes eight equally likely values and a regulator can choose among two responses. Use (54.2) to bound the entropy of outcomes from below. Construct an outcome table that attains the bound, and one that does worse.

Exercise 54.5. Label each occurrence of "information" in the following passage with one of the four senses of Section 54.5, and identify the step where the argument changes sense: "Neurons transmit information at a rate measurable in bits per second. The brain integrates this information into a unified model of the world. Because experience is unified, experience is integrated information."

Exercise 54.6. Write a preregistration for a test of the hypothesis that conscious states correlate with the degree of partial phase alignment among several cortical rhythms, beyond what the power of each rhythm predicts. State the measure, the comparison conditions, the null model, and the outcomes that would count against the hypothesis.

55

What Existing Fields Already Explain

Almost every step of this book belongs to a field that was there first. Moiré fringes belong to wave optics, locking to nonlinear dynamics, relative phase to coordination dynamics, implementation to the philosophy of computation, active sensing to neuroethology, closure to theoretical biology, representation to the philosophy of mind, and the hard problem to consciousness science and philosophy. This chapter says so precisely, in one table, and argues that saying so is the book's best defense against a specific failure, which we call disciplinary amnesia: the process by which a synthesis rediscovers a field's results under a new name and loses the precision the field had earned.

The thesis is that the progression of this book is owned step by step, and that what no field owns lies at the joints between steps. The joints have boundary owners too. The one thing that remains unowned is the comparative use of a single question, who or what reads the relation, at every joint, with one model system followed throughout. That is a method, and a method is a legitimate contribution; it is not yet a discipline.

55.1 Disciplinary Amnesia

Fields forget. New specialties split from old ones along some distinction, and the distinctions they split on tend to recur inside each branch, so that positions abandoned in one generation return in the next under different names (Abbott 2001). Syntheses that cross several fields are especially exposed, because their authors know some of the fields they cross less well than the fields' practitioners do.

Definition 55.1 (Disciplinary amnesia)**.** *Disciplinary amnesia* is the presentation of a result, concept, or method that an existing field already possesses as though it were new, under a new name and usually with less precision than the owning field achieved.

Amnesia has three costs. It loses precision, because the renamed concept is usually stated more loosely than the original. It duplicates effort, because the original field's refinements, counterexamples, and measurement techniques have to be rediscovered or are never recovered. It produces false novelty, which misleads readers about where the frontier is.

Example 55.2 (Renaming reafference). Parts V and VIII describe the variables an actively sensing animal reads as *world relative to its own action*, and Chapter 22 calls the resulting structure an agent-centered measurement frame. The phrase is a gloss. The concept it names is the reafference principle of von Holst and Mittelstaedt, which distinguishes sensory input caused by one's own action from input caused by the world (Holst and Mittelstaedt 1950), together with the active-sensing literature's account of animals that emit the energy they sense (M. E. Nelson and MacIver 2006). Used as a gloss, the phrase is harmless. Used as a new concept, it would sever the reader from more than seventy years of experiments on corollary discharge (Crapse and Sommer 2008).

The remedy is a map that sets each step of the book beside the field that owns it.

55.2 The Map

Table 55.1 is the map. Each row names a step of the book, the field that owns it, a canonical result or tool from that field, and what this book adds. The last column is short on purpose.

Table 55.1: Each step of the book beside the field that already explains it.

Step (chapter)	Owner	Canonical result or tool	Added here
Relational observables (Chapter 1)	Mechanics, symmetry, network science	Relative coordinates and reduction; higher-order interactions (Landau and Lifshitz 1976; Battiston et al. 2020)	Vocabulary
Moiré (Chapter 2)	Wave optics, moiré metrology	Moiré period and magnification; moiré measurement of strain (Amidror 2009; Post et al. 1994)	Nothing
Beats and new frequencies (Chapters 3 and 4)	Signal processing, nonlinear systems, auditory physiology	Envelope identity; square-law and heterodyne detection; Volterra series; combination tones (Oppenheim and Schafer 2010; Schetzen 1980; Goldstein 1967)	The reader question in physical terms

Step (chapter)	Owner	Canonical result or tool	Added here
Balance, resonance, commensurability (Chapters 5 to 7)	Dimensional analysis, mechanics, nonlinear dynamics	Buckingham's Π theorem; resonance curves; circle map and Arnold tongues (Buckingham 1914; Landau and Lifshitz 1976; Arnold 1983)	A three-way distinction
Magic angles (Chapter 8)	Condensed-matter physics	Flat bands in twisted bilayers (Bistritzer and MacDonald 2011; Cao, Fatemi, Fang, et al. 2018)	A worked lesson in relational explanation
Synchronization (Chapter 9)	Nonlinear dynamics	Kuramoto model; phase reduction (Kuramoto 1984; Pikovsky et al. 2001)	Nothing
Relative phase (Chapter 10)	Coordination dynamics	Haken–Kelso–Bunz model and its confirmed predictions (Haken et al. 1985; J. A. S. Kelso 1995)	Nothing
Metastability, chimeras (Chapters 11 and 12)	Nonlinear dynamics, computational neuroscience	Transient coordination; coexistence of coherence and incoherence (Tognoli and J. A. S. Kelso 2014; Abrams and Strogatz 2004)	Nothing
Order parameters (Chapter 13)	Synergetics	Slaving principle and adiabatic elimination (Haken 1983b)	Nothing
Causal emergence (Chapter 14)	Philosophy of science, information theory	Effective information and its critics (E. P. Hoel et al. 2013; Dewhurst and Philosophy Documentation Center 2021)	Nothing
Readers and reader-relative variables (Chapters 15 and 16)	Measurement, signal processing, information theory	Detectors, channels, and bottlenecks (Cover and Thomas 2006; Tishby et al. 2000)	The organizing question
Implementation and polycomputation (Chapters 17 and 18)	Philosophy of computation, unconventional computing	Counterfactual and mechanistic accounts; polycomputing (Chalmers 1996a; Piccinini 2015; Bongard and Levin 2023)	Emphasis on readers inside the system
Active sensing and reafference (Part V)	Neuroethology, sensorimotor neuroscience	Reafference principle; corollary discharge; the jamming avoidance response from receptor to pacemaker (Holst and Mittelstaedt 1950; Crapse and Sommer 2008; Heiligenberg 1991a)	The fish as a thread through the whole book

Step (chapter)	Owner	Canonical result or tool	Added here
Closure and meaning (Part VI)	Theoretical biology, philosophy of representation	Semantic closure; autopoiesis; teleosemantics (Pattee 2001; H. R. Maturana and F. J. Varela 1980; Millikan 1984)	Nothing
World models (Part VII)	Computational neuroscience, philosophy of mind	Internal models; predictive processing (Wolpert et al. 1995; K. Friston 2010; Clark 2013)	Nothing
Perspective (Part VIII)	Cognitive science, philosophy of language	Reference frames; the essential indexical; centered worlds (Andersen, Snyder, et al. 1997; Perry 1979; Lewis 1979)	Active sensing as a route to agent-centered variables, partly owned by O'Regan and Noë (2001)
Self-reading (Part IX)	Metacognition research, philosophy of mind	Confidence and monitoring; higher-order and self-model theories (Fleming et al. 2012; D. M. Rosenthal 2005; Metzinger 2003)	Nothing
Phenomenality (Part X)	Consciousness science and philosophy	The hard problem; theories and their adversarial tests (Chalmers 1995; Block 1995; Cogitate Consortium et al. 2025)	A sharper statement of where the gap sits

Read down the last column and the pattern is plain. Nine of the eighteen rows add nothing. The rest add a vocabulary, a question, a distinction, an emphasis, a thread, or a sharper statement of a problem; none adds a result the owning field lacks.

55.3 Physics and Dynamics

The first ten rows are owned outright. Moiré fringes are a solved problem in optics, with a literature of measurement techniques built on them (Amidror 2009; Post et al. 1994). Beats and difference frequencies are textbook signal processing, and the fact that a linear sum contains no component at the difference frequency until something nonlinear receives it is standard content in any account of detection (Oppenheim and Schafer 2010; Schetzen 1980). Resonance, commensurability, and locking belong to mechanics and nonlinear dynamics, which separated them long before this book insisted on the separation (Landau and Lifshitz 1976; Arnold 1983; Pikovsky et al. 2001). The collective

variables of Part III (order parameters, relative phase, metastable coordination) belong to synergetics and coordination dynamics, whose founding models made predictions that experiments confirmed (Haken 1983b; Haken et al. 1985; J. A. S. Kelso 1995).

What these fields do not ask is which subsystem responds to the relation. They have no need to. In each case the relation acts through dynamics: the beam that couples Huygens's clocks transmits a force that depends on their relative motion, and the order parameter enslaves the modes that make it up. Asking who reads such a relation is asking a question the physics does not pose, and the physics is complete without it. That is the first place where the reader question enters: at the boundary of a field that is finished on its own terms.

55.4 Computation, Sensing, and Control

The reader question is posed, in different words, by three fields that own the middle of the book.

The philosophy of computation asks when a physical system implements a computation, and its best answers already require what Part IV required of a reader: a causal organization that supports the right counterfactuals (Chalmers 1996a), or a mechanism whose components have the function of processing the relevant vehicles (Piccinini 2015). The polycomputing literature asks which of the many computations describable in a substrate are exploited by the substrate itself (Bongard and Levin 2023). Neuroethology asked, and answered for one animal, which circuits extract a relation from a sensory mixture: the jamming avoidance response was mapped from electroreceptors to the medullary pacemaker in the 1980s and early 1990s (Heiligenberg 1991a; Metzner 1993b). Control theory treats closed loops in which what an agent senses depends on what it does, and has been applied to the jamming avoidance response itself (Cowan et al. 2014; Madhav et al. 2013).

The contribution of the book in this region is organizational. It asks the question of implementation, of polycomputation, of neuroethology, and of control in one vocabulary and about one animal. Each field would recognize its own answers in the corresponding chapters.

55.5 Biology, Representation, and Perspective

Theoretical biology owns closure: Pattee's account of organizations that build and use their own measuring devices (Pattee 2001), the autopoietic account of self-producing organization (H. R. Maturana and F. J. Varela 1980), and later accounts of biological autonomy. The philosophy of representation owns the step from correlation to content and the possibility of error, from informational semantics to teleosemantics (F. I. Dretske 1981; Millikan 1984; Neander 2017).

Computational neuroscience owns internal models and prediction (Wolpert et al. 1995; K. Friston 2010). Cognitive science owns egocentric and allocentric reference frames (Andersen, Snyder, et al. 1997; Burgess 2006), and the philosophy of language owns the logic of *here*, *now*, and *I* (Perry 1979; Lewis 1979).

One row in this region adds more than an emphasis. Part VIII argued that active sensing constructs agent-centered variables before any self-model exists, because a variable of the form "world relative to my own emission" has the agent as a term by construction. The claim is partly owned: sensorimotor contingency theory makes the lawful relation between movement and sensation central to perception (O'Regan and Noë 2001), and the reafference literature supplies the mechanism (Holst and Mittelstaedt 1950; Crapse and Sommer 2008). What is less clearly owned is the connection this book draws between such variables and the formal peculiarities of indexicals. That connection is a joint between rows, and it is the kind of place where the book's residue lives.

55.6 Mind and Experience

The last two rows are owned as problems rather than as solutions. Metacognition research measures confidence and monitoring in humans and animals (Fleming et al. 2012); higher-order and self-model theories claim that such self-reading is what makes states conscious (D. M. Rosenthal 2005; Metzinger 2003). Consciousness science and philosophy own the distinction between access and phenomenal consciousness and the hard problem (Block 1995; Chalmers 1995), and they now own a method for testing theories against each other (Cogitate Consortium et al. 2025). The book adds a sharper statement of where the gap sits: at the arrow from functional perspective to phenomenal perspective, after every earlier arrow has been given a physical account (Chapters 43 and 63). Sharpening a problem is a contribution; it does not solve the problem, and it does not found a field.

55.7 Using the Map

The map is a checking tool: before a claim is called new, locate its row, read the owner's canonical result, and see what is left. The procedure is easiest to see on the claim that started this book.

Example 55.3 (Difference as well as integration). The first form of the book's question proposed that experience might depend on differentiation among many partially independent processes as well as on their integration: that the conscious state could be the pattern of their discrepancies, the way a moiré is the pattern of two gratings' mismatch. Three rows of the map bear on it. The phenomenality row points to neural complexity, a measure designed by Tononi, Sporns, and Edelman to be high exactly when a system combines functional

segregation with integration (Tononi, Sporns, and Edelman 1994), and to the dynamic-core proposal that conscious states are those in which a large set of neural processes is at once highly integrated and highly differentiated (Tononi 1998). The metastability row points to coordination in which regions tend toward and away from phase alignment without locking, which is integration without permanent synchronization (Tognoli and J. A. S. Kelso 2014). The synchronization row supplies the order parameters with which both kinds of claim are measured. After subtraction, the intuition is owned twice over, by a measure and by a dynamical regime. What remains is an empirical question with a definite form: does a relational measure of mismatch among rhythms predict the state of consciousness better than complexity, metastability, or the power of the rhythms themselves (Chapters 46 and 60)?

The example shows the typical outcome. The map rarely kills an idea outright; it converts a claim of novelty into a comparison with named rivals, which is the form in which an idea can be tested.

55.8 The Joints

If the rows are owned, the unowned material must lie between them. Three joints recur.

1. *Readout and implementation.* Which variables of a physical mixture are dynamically explicit for a given subsystem? Measurement theory treats detectors; the philosophy of computation treats implementation; neither treats the question of which relational variables a subsystem can use as its central problem (Chapter 16).
2. *Active sensing and indexicality.* When does a variable read through one's own emission become a variable about oneself? Neuroethology and the philosophy of language each hold one side.
3. *Closure and reading.* When a system builds its own reader, what does that add to the reading? Pattee owns this joint more fully than any other author (Pattee 1995; Pattee 2001).

Each joint has a boundary owner, usually an author who worked across two rows. What no one owns is the use of a single question at every joint, with one model system followed through all of them.

Principle 55.4 (Rule of the map)**.** A claim that crosses rows of Table 55.1 must cite the owner of each row it crosses and must preserve each row's precision: a quantity defined in one row may be used in another only after its definition has been restated there.

The rule would have caught the most common failure of syntheses like this one, which is to define a quantity carefully in one field and then use it loosely in another. The phrase "relational information," for example, can name a

mutual information between two signals in one row and an organism's use of a relation in another; the rule requires saying which (Chapter 54).

55.9 Audit

Canonical phenomenon
The eighteen steps of this book set beside the fields that explain them.

Minimal model
The map of Table 55.1, the definition of disciplinary amnesia (Definition 55.1), and the rule of the map (Principle 55.4).

What is relational here?
At every row, a variable of a configuration: a moiré period, a relative phase, a frequency difference, a relation between action and sensation, a relation between a system and its own states.

What physically reads it?
Varies by row, from none (the physics of moiré and beats) through engineered and neural readers (detectors, the fish's midbrain) to self-reading systems. The map's point is that each kind of reader already has a field that studies it.

What existing field explains it?
All of the fields in the second column of the table.

What remains unexplained?
The joints between rows, and the comparative use of one question at all of them with one model system. The first is shared with boundary owners; the second is a method.

A discriminating experiment
Delete the book's own vocabulary and rewrite each chapter's central claim in the vocabulary of its owning field. If a claim loses content in the translation, it marks a candidate contribution; if nothing is lost anywhere, the book's contribution is organizational.

55.10 Notes and References

On how disciplines split, forget, and rediscover, see Abbott (2001). The owners in Table 55.1 are cited in the rows themselves; the chapters named in the first column give fuller references. For the reafference principle, Holst and Mittelstaedt (1950) is the founding paper and Crapse and Sommer (2008) the comparative review; for active sensing, M. E. Nelson and MacIver (2006). The strongest boundary owners are Pattee (2001) for closure and reading,

O'Regan and Noë (2001) for action and perception, and Chalmers (1996a) for implementation.

55.11 Exercises

Exercise 55.1. Add a row to Table 55.1 for a relational phenomenon from a field you know well. Name the owner, cite a canonical result, and state honestly what the framework of this book would add.

Exercise 55.2. For three rows of the table, identify a result or method in the owning field that this book does not use, and describe how it would sharpen the corresponding chapter.

Exercise 55.3. Rewrite the central claim of Chapter 22 without the phrase "agent-centered measurement frame," using only the vocabulary of von Holst and Mittelstaedt and of the active-sensing literature. Does anything get lost? If so, say exactly what.

Exercise 55.4. Choose one of the three joints of Section 55.8. State a question at that joint that neither adjacent field has answered, describe an experiment or derivation that would answer it, and name the boundary owner whose work comes closest.

Exercise 55.5. Find two uses of the word "information" in different chapters of this book that belong to different rows of the table. For each, restate the definition as the rule of the map requires, and decide whether the two uses could be confused.

Exercise 55.6. Choose a synthesis you have read that claims a new field or a new unification. Apply Definition 55.1 to its three central concepts, and for each identify either its owning field or evidence that it has none.

56

What Would Metaxics Have to Add?

A field earns an independent existence by supplying something the existing fields cannot. The preceding chapters showed that the book's candidate unifications are owned (Chapter 53), that past unifications failed in instructive ways (Chapter 54), and that nearly every step of the progression belongs to a field that was there first (Chapter 55). This chapter states what would change that verdict.

The thesis is conditional. Metaxics would deserve to be called a science if it supplied at least one of five things: a new formalism, a new invariance principle that transfers across substrates, a new experimental operation, a new quantitative prediction, or a new empirical regularity. For each we state the best candidate the book has produced, as a conjecture or hypothesis with its failure condition. None of the five is yet satisfied, and the honest conclusion is that Metaxics is at present a synthetic perspective: a way of organizing existing sciences around one question, with a model system, a method, and an agenda.

56.1 What Makes a Field

Philosophy of science offers several criteria for when a body of work is a science, and they emphasize different things. Kuhn located a mature science in a shared paradigm: exemplary solutions that define which puzzles are worth solving and what counts as solving them (Kuhn 1962). Popper required that a theory forbid something, so that some possible observation would refute it (Popper 1959). Lakatos judged research programmes by their trajectory: a programme is progressive when it predicts novel facts that are then corroborated, and degenerating when it only accommodates facts discovered by others (Lakatos 1970). Fodor argued that a special science earns its autonomy when its kind terms enter laws that cannot be stated in the vocabulary of a lower-level science, which multiple realization makes possible: the same higher-level kind

can be realized by physically different mechanisms (Fodor 1974). Disciplines are also social objects, with departments, journals, and training (Abbott 2001); that sense of "field" is not the one at issue here.

These criteria agree on one point that matters for this book. A synthesis that reorganizes existing results, however illuminating, does not become a science by doing so. It becomes one when it adds something that its sources could not have produced.

Principle 56.1 (Five ways to add)**.** A research perspective becomes a candidate science when it supplies at least one of the following.

A. A *new formalism*: an object or variable that existing frameworks cannot express without loss.
B. A *new invariance principle*: a theorem whose hypotheses are stated in the perspective's own terms and which transfers across substrates.
C. A *new experimental operation*: a way of manipulating a system that existing fields do not use and whose outcomes they do not predict.
D. A *new quantitative prediction*: a measurable quantity for which the perspective's predicted value differs from every rival's, $P_{\text{Metaxics}} \neq P_{\text{existing}}$.
E. A *new empirical regularity*: a pattern found repeatedly across otherwise unrelated systems and not already expected by an existing theory.

The criteria are ordered by how hard they are to fake. A formalism can be notation in disguise; a theorem can be a known theorem in new words; an operation can be an old technique renamed. A prediction that differs from every rival's and survives a preregistered test cannot be faked, and neither can a regularity that no existing theory expected.

56.2 A New Formalism

The book's formal object is the reading loop $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$: a mixture $y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})$ of agent and environment, a reader $z = \mathsf{D}(y)$, a policy $a = \pi(z)$, and an intervention $\dot{s}_{\mathrm{A}} = \mathsf{F}(s_{\mathrm{A}}, a)$ that changes the next mixture (Chapter 15).

Control theory writes the same loop in different letters. A plant with state x (containing both s_{A} and s_{E}), a measurement equation $y = h(x)$ in which the measurement may depend on the controller's own state, a controller, and an actuator: nothing in $\mathcal{L}$ lacks a counterpart. Biologists already use this vocabulary for animal behavior, including the jamming avoidance response (Cowan et al. 2014; Madhav et al. 2013), and the internal model principle shows that control theory has results about what such loops must contain (Chapter 53).

Conjecture 56.2 (A loop invariant)**.** *There is a quantity defined on the pair* (M, D) *that measures how much of the read variable* z *depends on the agent's*

own contribution to the mixture, and this quantity predicts features of closed-loop behavior (stability, the direction in which the agent drives the relation) that standard observability and information measures do not.

Failure condition. The conjecture fails if every candidate quantity reduces to an existing one: a sensitivity of the measurement map to the agent's own state, an observability measure, or a mutual information between the agent's actions and its inputs. We expect it to fail. Until a quantity is written down and shown to predict something the existing measures miss, the reading loop is a notation, a useful one for this book and not a new formalism.

56.3 A Transferable Theorem

The best candidate theorem came from the fish. A reader that has no internal reference for its own carrier must determine the sign of a frequency difference from the sensory mixture alone, and the mixture's envelope does not carry that sign.

Proposition 56.3 (Sign-blindness of envelopes)**.** *For the mixture*

$$y(t) = A_{\mathrm{s}} \left[\cos \omega_{\mathrm{s}} t + m \cos(\omega_{\mathrm{e}} t + \psi_0)\right],$$

the envelope at any site depends on $(\Delta\omega, \psi_0)$ *only through* $\cos(\Delta\omega\, t + \psi_0)$*, and is therefore identical for* $(\Delta\omega, \psi_0)$ *and* $(-\Delta\omega, -\psi_0)$*.*

The proof is one line (Chapter 53), and the fish's solution, combining amplitude with differential phase across sites of different contamination, is derived in Chapter 21. The result transfers across substrates: it holds for any receiver of a sum of two sinusoids, whether fish, bat, radio, or cochlea. It passes the transfer test. It fails the ownership test, because it is a special case of the symmetry of real spectra, which signal theory states for every substrate at once and which is the reason engineered receivers use in-phase and quadrature channels to distinguish positive from negative frequency offsets (Oppenheim and Schafer 2010).

A second candidate fares no better. A memoryless polynomial reader of degree n receiving two tones produces only combination frequencies $|p\,\omega_1 \pm q\,\omega_2|$ with $p + q \le n$, and generically all of them, so different readers expose different relational variables (Chapter 4). This is the reader dependence the book was built on, and it is intermodulation theory, a chapter of nonlinear systems analysis (Schetzen 1980). No theorem the book has proved is its own.

56.4 A New Experimental Operation

The strongest candidate is an operation rather than a result. Part XIII develops three manipulations: changing a relation while holding its relata fixed

(Chapter 57), changing a reader while holding its input fixed (Chapter 58), and opening or closing the loop through which a system alters the relation it reads (Chapter 59). Combined, they give a factorial design.

Hypothesis 56.4 (Reader relativity in closed loops)**.** In a closed reading loop, changing the reader D while the mixture M, policy π, and intervention F are held fixed changes both the effective variable z and the fixed points of the loop; changing the relation while holding D fixed changes z along the reader's tuning. The two manipulations dissociate.

Failure condition. The hypothesis fails for a given system if no admissible change of reader alters z for fixed input, or if every effect attributed to the reader is explained by a change in the mixture or in the policy.

Each manipulation exists already. Stimulus control, lesions, pharmacology, and adaptation are standard in neurophysiology, and the jamming avoidance response has been studied with all of them, including a frequency clamp that holds Δf fixed while the fish tries to change it (Metzner 1993b; Madhav et al. 2013). What may be new is the systematic crossing of relation, reader, and loop in one design, applied across substrates with a common readout. That is a protocol, and a protocol can be a real contribution; it becomes a new operation only if it finds effects that the owning fields did not predict. This is the criterion on which the book has the best chance.

56.5 A New Quantitative Prediction

A prediction counts only if it differs from every rival's. For consciousness, the rivals make specific claims: ignition and global broadcasting for the workspace theory, integrated causal structure for integrated information theory, and simple spectral features for the indices used in anesthesia monitoring (Chapters 44 and 46).

Hypothesis 56.5 (Relational measures beyond their relata)**.** There is a relational measure of cortical dynamics (for example, the phase relation between rhythms, or the width of their entrainment tongues) whose value changes between conscious and unconscious states when the spectra and marginal statistics of each constituent rhythm are held fixed, by surrogate construction or by matched stimulation.

Failure condition. The hypothesis fails if every relational effect disappears once the constituent measures are matched. There is a precedent in clinical monitoring: in 39 patients during induction of anesthesia, the bispectral parameter used in the bispectral index was no better than an analogous power-spectrum parameter at separating the awake from the anesthetized state, and an amplitude-independent bicoherence measure did not change at all (Miller et al. 2004). It also fails as a *Metaxics* prediction if a rival theory predicts

the same effect, since a prediction shared with a rival does not distinguish the perspective.

The hypothesis is stated without a number because none can be stated honestly before pilot data exist. A quantitative prediction in the sense of criterion D would specify the measure, the conditions, the matched null, and the expected effect size in advance (Chapters 60 and 61). The book has not yet made one.

56.6 A New Empirical Regularity

The comparative chapters suggest a regularity. Actively sensing animals that evolved independently control a self-generated carrier so as to keep a relational variable in the range where their reader performs well.

Example 56.6 (Two carrier controllers). The electric fish drives the difference between its frequency and its neighbor's away from zero, where the beat would jam its electrolocation, and two lineages that evolved electrosensation independently do so with the same computational rules (Kawasaki 1993; Kawasaki 1996). Horseshoe bats lower the frequency of their calls during flight so that Doppler-shifted echoes return near a reference frequency, where the cochlea and auditory pathway devote an expanded, sharply tuned representation, the auditory fovea (Schnitzler 1968; Schnitzler and Denzinger 2011). In both, the animal changes its own emission to set a relation between emission and return, and in both the target of the control is a region where the animal's reader works well.

Hypothesis 56.7 (Carrier control for reader performance)**.** In actively sensing animals, the self-generated carrier is controlled so as to hold the relation between carrier and return in the region of best performance of the animal's reader.

Failure condition. The hypothesis fails if an actively sensing animal controls its carrier in a way that moves the relation out of its reader's best region, or if the regularity is already the content of an existing theory.

The second failure condition is met. Active-sensing research already holds that animals shape the energy they emit, and their movements, to suit their sensors (M. E. Nelson and MacIver 2006), and the auditory fovea literature explains Doppler-shift compensation in exactly these terms (Schnitzler and Denzinger 2011). The strong version of the regularity also fails on the details. The fish reads its relation without a copy of its motor command, while the bat's compensation circuitry receives vocal-motor signals (Metzner 1993a); the fish pushes the relation away from a value, while the bat holds it at one. Two instances with different mechanisms and opposite objectives are a comparison, and the comparison is useful, but they do not establish a regularity that an existing theory failed to expect.

Fodor's criterion suggests where a stronger regularity would have to come from. The convergence of the jamming avoidance response in two lineages is evidence that its computational description picks out a real kind, one realized by different neural mechanisms. A science of relational readers would need many such kinds, entering laws that hold across their realizations. One kind with two realizations is where such a science would start, if it starts at all.

56.7 What a Synthetic Perspective Is For

A perspective that satisfies none of the five criteria can still do useful work, and it is worth saying what, so that the verdict is not mistaken for a dismissal.

It can choose model systems. The electric fish was chosen because it exhibits every step of the reading loop in a form that has been measured, and following one animal through fourteen parts exposes connections that a survey organized by discipline would scatter across unrelated literatures.

It can sharpen questions. The question of consciousness, restated at the end of a progression in which every earlier arrow has a physical account, is narrower and easier to test than the question the book began with (Chapter 63).

It can design experiments. The crossed designs of Part XIII come from asking one question of every step, and they are the most likely source of a result that would satisfy a criterion.

It can prevent amnesia, by keeping the map of Chapter 55 in view.

What a synthetic perspective should not do is acquire the social apparatus of a discipline (a journal, a society, a curriculum under its own name) before it has results that need them. The social form of a field follows its intellectual achievements in the healthy case (Abbott 2001), and the order matters: an institution built around a name tends to defend the name.

56.8 The Verdict

Table 56.1 summarizes.

No criterion is satisfied. Metaxics is, at present, a synthetic perspective. In Lakatos's terms it is a programme with a hard core (the question of how a relation becomes a variable for a system that participates in it) and a positive heuristic (subtract the established explanation, name the residue, design the test), and it has not yet predicted a novel fact. That is a respectable position for an introductory book, and the only honest one available.

Three developments would change the verdict. A quantity defined on reading loops that predicts closed-loop behavior better than existing measures would satisfy criterion A. A crossed relation-by-reader experiment that finds an interaction no owning field predicted would satisfy criterion C. A preregistered relational measure that separates conscious from unconscious states after its

Criterion	Best candidate	Failure condition	Status
A. Formalism	A loop invariant on (M, D) (Conjecture 56.2)	Reduces to observability or information measures	Unwritten; expected to fail
B. Theorem	Sign-blindness of envelopes (Proposition 56.3)	Derivable from spectral symmetry	Fails ownership
C. Operation	Crossing relation, reader, and loop (Hypothesis 56.4)	No effects beyond those the owning fields predict	Best candidate; untested as a protocol
D. Prediction	Relational measures beyond their relata (Hypothesis 56.5)	Effects vanish with matched constituents, or rivals predict them	Untested; no number yet
E. Regularity	Carrier control for reader performance (Hypothesis 56.7)	Already expected by active-sensing theory	Fails ownership

Table 56.1: The five criteria and the book's best candidate for each.

constituents are matched, and that no rival theory predicted, would satisfy criterion D. Part XIII designs toward all three.

56.9 Audit

Canonical phenomenon

A research perspective asking whether it has earned the status of a science.

Minimal model

Five criteria (Principle 56.1); a conjecture, a proposition, and three hypotheses, each with a failure condition.

What is relational here?

In each candidate, a relation between an agent's own signal or state and an external one, read by a physical system inside the loop.

What physically reads it?

The fish's midbrain for the theorem and the regularity; experimental designs for the operation; cortical dynamics, still unidentified, for the prediction.

What existing field explains it?

Philosophy of science on paradigms, falsifiability, research programmes, and special sciences (Kuhn 1962; Popper 1959; Lakatos 1970; Fodor 1974); control theory for the formalism; signal theory for the theorem; neurophysiological method for the operation; consciousness science for the prediction; active-sensing research for the regularity.

What remains unexplained?

The crossed relation-by-reader operation, and a relational prediction about consciousness that survives matched controls. Both are untested.

A discriminating experiment

Run the crossed design of Hypothesis 56.4 in the jamming avoidance response: clamp Δf at several values while altering the reader by adaptation or pharmacology, and ask whether the reader manipulation changes the fixed points of the closed loop in a way that the mixture and the policy alone do not predict.

56.10 Notes and References

The criteria for a science draw on Kuhn (1962), Popper (1959), and Lakatos (1970); the argument for the autonomy of special sciences is Fodor (1974); the social account of disciplines is Abbott (2001). Control-theoretic treatments of animal behavior, including the jamming avoidance response, are reviewed by Cowan et al. (2014) and applied by Madhav et al. (2013). Spectral symmetry is standard (Oppenheim and Schafer 2010); intermodulation belongs to nonlinear systems theory (Schetzen 1980). The bispectral index result is Miller et al. (2004). Doppler-shift compensation is described by Schnitzler (1968) and reviewed with the auditory fovea by Schnitzler and Denzinger (2011); the audio-vocal neurons are reported by Metzner (1993a); active sensing is reviewed by M. E. Nelson and MacIver (2006).

56.11 Exercises

Exercise 56.1. Write the hard core, protective belt, and positive heuristic of Metaxics in Lakatos's sense. Identify one auxiliary hypothesis that a defender could modify to protect the core from a failed prediction, and say whether modifying it would be progressive or degenerating.

Exercise 56.2. Translate the reading loop $\mathcal{L}$ into the state-space form of control theory, with a plant, a measurement equation, a controller, and an actuator. Identify each of M, D, π, and F in the translation, and state whether anything is lost.

Exercise 56.3. For the reader $\mathsf{D}(x) = x^3$ receiving $x(t) = \cos \omega_1 t + \cos \omega_2 t$, list every frequency present in the output and its amplitude. Verify that each has the form $|p\,\omega_1 \pm q\,\omega_2|$ with $p + q \le 3$.

Exercise 56.4. Turn Hypothesis 56.5 into a preregistration: choose a relational measure, two conditions that differ in the state of consciousness, a surrogate method that holds each rhythm's spectrum fixed, and a smallest effect size of interest. State the outcome that would falsify the hypothesis.

Exercise 56.5. Find a third actively sensing animal (an echolocating dolphin, a whisking rodent, or another of your choice) and test Hypothesis 56.7 against it. Does the animal control its carrier or its movement to keep a relational variable in its reader's best range? Cite the evidence.

Exercise 56.6. Fodor's argument requires that a higher-level kind enter laws. State one law in which the computational kind "reads the sign of a frequency difference from amplitude and differential phase" appears, and say whether the law holds across *Eigenmannia* and *Gymnarchus*.

General systems theory, catastrophe theory, and the edge of chaos each promised a science of structure across substrates, and each overreached in a specific, instructive way. Mapped against existing fields, almost every step of this book is already owned by someone. A new discipline would need a new formal object, a transferable theorem, a new experimental operation, a distinctive prediction, or a new regularity. Part XIII asks what experiments could supply one, and what results would show that none exists.

Part XIII

Research Programme

How could the relational claims be tested, and how could they fail?

57

Relational Intervention

The fourth criterion of Chapter 1 asked for intervention sensitivity: to show that a relation, and not some property of its relata, does causal work, one must change the relation while holding the relata as fixed as physically possible. Most of the book's relational claims have not been tested this way. This chapter describes how they can be.

The thesis is that a relational intervention is an intervention that changes a relation while preserving a stated list of intrinsic properties of each relatum, and that the list is what makes the intervention interpretable. Relations are fixed by their relata, so every intervention on a relation changes some relatum in some respect; the logical problem this raises is solved by saying which respects are held fixed. Physics and neuroethology supply clean cases: the twist angle between two crystal layers, the electric fish's mimic stimuli, and a clamp that holds the fish's frequency difference fixed while the fish tries to change it. Brain stimulation supplies instructive, less clean ones. And a relational intervention, however clean, shows that a relation matters without showing that anything reads it.

57.1 The Logic of Intervening on a Relation

On the interventionist account of causation, X is a cause of Y when some possible intervention on X would change Y, where an intervention sets X from outside its usual causes, breaks those causes' influence on it, and affects Y only through X (Woodward 2003). Pearl's notation writes such an intervention as $\mathrm{do}(X = x)$ and gives rules for computing its consequences from a causal model (Pearl 2009).

For a relation $r = \mathcal{R}(x_1, x_2)$, the account meets a difficulty. Because r is a function of the relata, no intervention can change r without changing at least one of x_1 and x_2. Every intervention on a relation is therefore also an intervention on something else, a situation the philosophical literature calls fat-handedness: the manipulation cannot be confined to the variable of interest (Baumgartner and Gebharter 2016). If a phase difference is changed by shifting

one oscillator, the shifted oscillator has changed, and one might attribute any effect to it.

The difficulty dissolves once the state of each relatum is divided into properties that the manipulation preserves and properties it changes. Chapter 1 showed that under a symmetry acting on all components at once, only relations are frame-independent. A symmetry acting on one component alone does something complementary: it leaves that component's intrinsic properties unchanged and alters its relations to the others.

Definition 57.1 (Relation-selective intervention)**.** Let x_1 and x_2 be relata, $r = \mathcal{R}(x_1, x_2)$ a relation, and ι_1, ι_2 stated sets of intrinsic statistics of each relatum (for signals: frequency content, amplitude distribution, waveform). An intervention is *relation-selective* with respect to (ι_1, ι_2) if it changes the distribution of r while leaving every statistic in ι_1 and ι_2 unchanged.

The canonical example is a time shift.

Proposition 57.2 (Time shifts are relation-selective)**.** *Let $x_1(t)$ and $x_2(t)$ be jointly stationary signals with cross-spectrum $S_{12}(\omega)$. The intervention $x_2(t) \mapsto x_2(t - \tau)$ leaves the power spectrum and the marginal distribution of each signal unchanged, leaves the coherence $|S_{12}|^2/(S_{11}S_{22})$ unchanged, and multiplies $S_{12}(\omega)$ by $\mathrm{e}^{\mathrm{i}\omega\tau}$, shifting the cross-spectral phase at frequency ω by $\omega\tau$.*

Proof. A delay multiplies the Fourier transform of x_2 by $\mathrm{e}^{-\mathrm{i}\omega\tau}$, a factor of unit modulus. The power spectrum S_{22} depends on $|\hat{x}_2|^2$ and is unchanged; stationarity makes the marginal distribution independent of time, so it is unchanged too. The cross-spectrum is formed from $\hat{x}_1\hat{x}_2^*$ and acquires the factor $\mathrm{e}^{\mathrm{i}\omega\tau}$; its modulus, and therefore the coherence, is unchanged. □

The proposition is the template for everything in this chapter: find a transformation of one relatum under which its intrinsic statistics are invariant, apply it, and the relation changes while nothing on the list ι does. Surrogate-data methods run the same logic in the opposite direction. Randomizing the Fourier phases of each signal independently destroys the relations between signals while preserving each signal's spectrum, and so provides the null model against which a relational statistic must be judged (Theiler et al. 1992) (Appendix J).

Principle 57.3 (State the invariants)**.** A relational intervention must list the intrinsic statistics it preserves and the relations it changes, including relations to components that were not measured. A claim that an effect is due to a relation is exactly as strong as that list.

The clause about unmeasured components matters. Shifting x_2 in time changes its relation to x_1, and also its relation to every other signal $x_3, x_4, \ldots$

in the system. If one of those carries the effect, the intervention was relation-selective with respect to the wrong relation.

57.2 A Catalogue

Table 57.1 lists the relations that experiments have manipulated selectively, with what each manipulation holds fixed.

Relation	**Manipulation**	**Held fixed**	**Example**
Phase	Stimulate two sites in phase or in anti-phase; time a stimulus to the phase of an ongoing rhythm	Frequency and intensity at each site	Dual-site tACS (Polanía et al. 2012; Helfrich et al. 2014); closed-loop auditory stimulation (Ngo et al. 2013)
Frequency difference	Set the frequency of one oscillator	The other oscillator	Mimic stimuli in electric fish (Heiligenberg 1991a)
Clamped relation	Feedback holds the relation at a set value while the system acts	The system's own dynamics	Frequency clamp of the jamming avoidance response (Madhav et al. 2013)
Delay	Insert a delay between an action and its sensory consequence	The action and the content of the feedback	Delayed auditory feedback of speech (B. S. Lee 1950)
Orientation	Rotate one layer relative to another	Each layer's lattice	In-situ rotation of a heterostructure (Ribeiro-Palau et al. 2018)
Coupling	Add an artificial conductance between cells	Each cell's intrinsic currents	Dynamic clamp (Sharp et al. 1993; Netoff et al. 2005)
Topology	Set which oscillators are coupled, and how strongly within and between groups	Each oscillator	Mechanical and chemical oscillator networks (Martens et al. 2013; Tinsley et al. 2012)
Sensorimotor contingency	Change how sensation depends on movement	The stimulus itself	Closed-loop refuge tracking in electric fish (Biswas et al. 2018)

Table 57.1: Relations that have been manipulated selectively, and what each manipulation holds fixed.

Three rows show the logic at its cleanest.

Orientation. The electronic properties of stacked two-dimensional crystals depend strongly on the angle between their lattices (Chapter 8), and until recently each angle required a separately fabricated device, so that angle was confounded with every other difference between devices. Ribeiro-Palau and colleagues shaped the top boron nitride layer of a graphene heterostructure so that an atomic force microscope tip could push it, varied the twist angle in situ in steps as small as 0.2°, and followed the device's properties through the rotation (Ribeiro-Palau et al. 2018). Each layer's lattice is unchanged; the relation between them is the only variable.

Coupling. The dynamic clamp uses a computer to inject simulated conductances into real neurons in real time; its inventors used it to build reciprocal inhibitory synapses between two stomatogastric neurons that are not naturally coupled, forming a new network at will (Sharp et al. 1993). The same method couples biological neurons to one another or to model neurons, so that the coupling, and therefore the phase relation it produces, can be set while the cells' intrinsic properties are left alone (Netoff et al. 2005).

Phase. In sleeping humans, sounds timed to arrive in phase with the up states of the slow oscillation enhanced the oscillation, the spindle activity coupled to it, and the consolidation of declarative memory, while the same sounds delivered out of phase were ineffective (Ngo et al. 2013). Stimulus content, intensity, and number were the same in both conditions. The relation between the stimulus and the ongoing rhythm made the difference.

57.3 The Fish as the Model Case

The electric fish has been studied with relational interventions for as long as its jamming avoidance response has been known. In the classic preparation the experimenter supplies both terms of the mixture: a sinusoidal mimic of the fish's own discharge, which Heiligenberg's papers call S1, and a mimic of a neighbor's, S2 (Heiligenberg 1991a). Each mimic is a sinusoid of stated frequency and amplitude, and those intrinsic properties are set independently. The relations are then under the experimenter's control: the frequency difference Δf, and, through the geometry of the electrodes, the contamination ratio m at each region of the fish's skin and hence the pattern of amplitude and differential-phase modulation that the fish receives (Chapter 21).

Because the mimic of the fish's own discharge is driven by the experimenter rather than by the fish's pacemaker, this preparation also opens the loop. The fish's responses then show what the sensory mixture alone supports. Experiments of this kind established that the jamming avoidance response needs no internal reference to the pacemaker command: the relation carried by the mixture suffices (Heiligenberg 1991a), and the same holds for the independently evolved response of *Gymnarchus* (Kawasaki 1993).

The loop can also be closed on the experimenter's terms.

Definition 57.4 (Relation clamp)**.** A *relation clamp* is a feedback arrangement that measures a system's contribution to a relation and adjusts the other term so that the relation stays at a value set by the experimenter, whatever the system does.

Madhav and colleagues built a relation clamp for the jamming avoidance response (Figure 57.1): the frequency of the jamming stimulus was tied to the fish's own measured frequency, so that Δf was held at a chosen value while the fish tried to change it (Madhav et al. 2013). The response, which is unstable on its own because the fish drives $|\Delta f|$ away from zero, was stabilized. Near equilibrium a delayed first-order linear model fit the behavior; clamping Δf at a range of values revealed the global nonlinear dynamics, and a model with competing sensory and motor components predicted a "snap-through" bifurcation as a function of Δf, which the experiments then confirmed (Madhav et al. 2013).

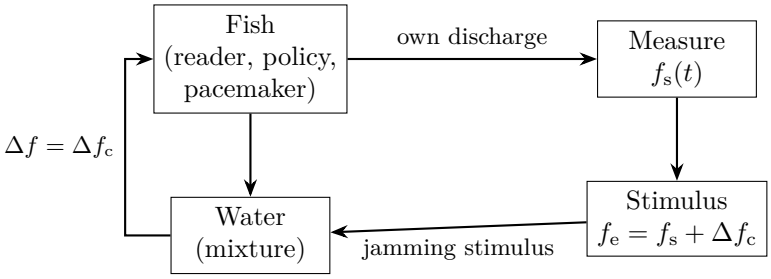


Figure 57.1: A relation clamp. The stimulus generator follows the fish's measured frequency so that the relation Δf stays at the experimenter's value Δf_c however the fish changes its discharge, after Madhav et al. (2013).

The clamp converts the relation from a dependent variable into an independent one. Without it, Δf is an outcome of the fish's own behavior and cannot be varied freely; with it, Δf is set, and the fish's response to each value can be measured. The same move underlies the most successful interventions in cellular physiology, where a quantity that a system normally controls is held fixed by feedback so that the system's response to it can be studied.

Example 57.5 (Perturbing a locked relation)*.* Two oscillators with detuning $\Delta\omega$ and coupling K obey, in the phase approximation, $\dot{\phi} = \Delta\omega - K\sin\phi$ for their phase difference ϕ. For $|\Delta\omega| < K$ the phase difference locks at $\phi^* = \arcsin(\Delta\omega/K)$. A brief kick that displaces ϕ without changing either oscillator's frequency is a relation-selective intervention, and the return is exponential with rate $K\cos\phi^* = \sqrt{K^2 - \Delta\omega^2}$. The rate falls to zero as $|\Delta\omega|$ approaches K: recovery from a relational perturbation slows as the locked state approaches its loss. In human bimanual coordination, perturbing the relative phase of the two hands and measuring recovery is exactly this intervention, and the predicted slowing near the transition was one of the confirmed predictions of coordination dynamics (Schöner and J. A. S. Kelso 1988) (Chapter 10).

57.4 Relational Interventions in Brains

Rhythmic stimulation of the brain sets a relation between an intrinsic rhythm and an external drive. Bursts of transcranial magnetic stimulation at the frequency of a parietal alpha rhythm entrained it, with effects that depended on the phase of the ongoing rhythm at stimulation onset (Thut et al. 2011); rhythmic visual flicker entrained alpha more strongly at higher intensity and at frequencies closer to each person's own alpha frequency, the pattern that the Arnold tongue of Chapter 7 predicts (Notbohm et al. 2016). Large-scale simulations of spiking networks predict the same structure for alternating-current stimulation, with an Arnold tongue centered on the network's resonance frequency (Ali et al. 2013), and further modeling shows that periodic drive shapes intrinsic oscillations through resonance, entrainment, and nonlinear acceleration (Herrmann et al. 2016).

These experiments manipulate the relation between drive and rhythm, but they also change the rhythm's relata: the drive adds input. Dual-site stimulation is closer to relation-selective. Theta-band alternating current delivered simultaneously to left prefrontal and parietal cortex, in phase or in anti-phase, improved memory-matching reaction times relative to placebo in the first condition and impaired performance in the second (Polanía et al. 2012). Gamma-band stimulation of the two hemispheres in phase enhanced interhemispheric synchronization and in anti-phase impaired it, with correlated changes in the perception of an ambiguous motion display (Helfrich et al. 2014). In both, each site received the same frequency and intensity of stimulation in both conditions; only the relative phase differed.

The list of invariants for such experiments must include the routes by which stimulation acts. Experiments in rats and humans showed that reported effects of transcranial alternating current on the motor system can be produced by stimulation of peripheral nerves in the skin, whose rhythmic activity then entrains cortex, rather than by current reaching the cortex directly (Asamoah et al. 2019). If the scalp is a relatum, a phase manipulation between two scalp sites changes relations among peripheral inputs as well as among cortical regions, and Principle 57.3 requires the design to separate them.

57.5 What a Relational Intervention Shows

A clean relational intervention supports a hypothesis of a definite form.

Hypothesis 57.6 (Relational causal relevance)**.** For a system with relata x_1, x_2, relation r, and outcome Y: under relation-selective interventions with respect to stated invariants (ι_1, ι_2), the distribution of Y depends on the value to which r is set.

Confirming the hypothesis shows that the relation is causally relevant to Y, relative to the invariants. It does not show that anything in the system reads

the relation. The beam that couples Huygens's clocks responds to their relative motion without being a reader in any sense the book has developed (Chapter 1); the twist angle of a heterostructure changes its electronic properties through band structure, not through a reader. A relation can matter through dynamics alone.

Showing that a relation is read requires two further manipulations. The first holds the relation fixed and changes the system's reader, which is the subject of Chapter 58. The second opens and closes the loop through which the system changes the relation it reads, which is the subject of Chapter 59. The fish is the one system in this book in which all three kinds of manipulation have been made: mimic stimuli set the relation, stimulation of the nucleus electrosensorius and lesions of the prepacemaker nuclei alter how the read relation is used (C. H. Keller et al. 1990; Metzner 1993b), and the relation clamp controls the loop.

57.6 Audit

Canonical phenomenon

A relation changed while its relata are held fixed: a twist angle varied in a single device, a phase relation between stimulation sites or between a sound and a sleep rhythm, and the frequency difference between an electric fish and a mimic neighbor, including when it is clamped.

Minimal model

Relation-selective interventions (Definition 57.1); the time-shift proposition (Proposition 57.2); the relation clamp (Definition 57.4); the locked phase equation of Example 57.5.

What is relational here?

Phase, frequency difference, delay, orientation, coupling, topology, and the dependence of sensation on movement.

What physically reads it?

In the fish, the midbrain circuits that compute the sign of Δf; in the brain-stimulation experiments, unidentified. In the physical examples, none: the relation acts through dynamics.

What existing field explains it?

Interventionist theories of causation (Woodward 2003; Pearl 2009); the fat-handedness debate (Baumgartner and Gebharter 2016); surrogate-data methods (Theiler et al. 1992); neuroethology of the jamming avoidance response (Heiligenberg 1991a; Madhav et al. 2013); the physiology of entrainment and brain stimulation.

What remains unexplained?

None in the method. The residue is empirical: most relational claims in

this book, including every claim about consciousness, have not been tested with relation-selective interventions.

A discriminating experiment

For any relational measure proposed as a correlate of consciousness, apply a relation-selective intervention that preserves each constituent's spectrum (dual-site stimulation with matched intensity, varying only relative phase) and record whether the measure and the report change together. If report changes only when constituent statistics change, the relational measure is not the causally relevant variable.

57.7 Notes and References

The interventionist account of causation is developed by Woodward (2003); Pearl (2009) gives the formal calculus; Baumgartner and Gebharter (2016) discuss fat-handed interventions. The logic of surrogate data is due to Theiler et al. (1992). In-situ control of the twist angle is reported by Ribeiro-Palau et al. (2018). The dynamic clamp is introduced by Sharp et al. (1993) and used to build hybrid networks by Netoff et al. (2005); oscillator networks with controlled coupling are described by Martens et al. (2013) and Tinsley et al. (2012). Phase-specific stimulation is reported by Ngo et al. (2013), Polanía et al. (2012), and Helfrich et al. (2014); entrainment by rhythmic stimulation by Thut et al. (2011) and Notbohm et al. (2016); its modeling by Ali et al. (2013) and Herrmann et al. (2016); the peripheral confound by Asamoah et al. (2019). Mimic stimuli in the jamming avoidance response are described throughout Heiligenberg (1991a); the relation clamp is Madhav et al. (2013); closed-loop control of active sensing movements is Biswas et al. (2018). Delayed auditory feedback was reported by B. S. Lee (1950).

57.8 Exercises

Exercise 57.1. Prove Proposition 57.2 with explicit Fourier conventions. Then show that a time shift is not relation-selective with respect to the invariant "the cross-correlation between x_2 and a third signal x_3," and describe a design in which that failure would matter.

Exercise 57.2. For $\dot{\phi} = \Delta\omega - K \sin \phi$, derive the locking condition, the locked phase ϕ^*, and the relaxation rate after a small kick. Plot the relaxation time against $\Delta\omega/K$ and identify the signature that an experimenter perturbing the relation would see as the system approaches unlocking.

Exercise 57.3. Design a relation-selective intervention on two cortical rhythms recorded with EEG. List the intrinsic statistics of each rhythm your intervention preserves, the relations it changes (including relations to unmeasured regions

and to peripheral inputs), and the control conditions that would let you attribute an effect to the targeted relation.

Exercise 57.4. In the mimic preparation for the jamming avoidance response, list the intrinsic properties of S1 and S2 that the experimenter sets. Which relations are set directly, and which only through electrode geometry? Using the two-patch model of Chapter 21, show how changing the geometry changes the contamination ratios and therefore the amplitude and differential-phase modulations the fish receives.

Exercise 57.5. Model the fish's frequency as $\dot{f}_\mathrm{s} = -k\tanh(\Delta f/w)$ with $\Delta f = f_\mathrm{e} - f_\mathrm{s}$ and $k, w > 0$. Show that without a clamp, $\Delta f = 0$ is an unstable equilibrium of the relation. Now impose the clamp $f_\mathrm{e}(t) = f_\mathrm{s}(t) + \Delta f_\mathrm{c}$ and show that the fish's rate of frequency change becomes a function of Δf_c alone, measurable as an open-loop quantity.

Exercise 57.6. The closed-loop sleep study timed identical sounds either in phase or out of phase with the slow oscillation. State the invariants of this design, identify one relation other than stimulus phase that also differs between the conditions, and propose a control condition that would rule it out.

58

Reader Intervention

The Preface stated the book's central claim in two factors: which relational variables become dynamically explicit depends jointly on the physical mixture and on the transduction performed by the receiving system. Chapter 57 showed how to vary the first factor, by changing a relation while holding its relata fixed. This chapter varies the second. A reader intervention changes the reader while the input is held fixed, and asks whether the same physical mixture now yields a different effective variable.

The thesis is that reader intervention is the direct test of reader dependence, and arguably the most distinctively metaxic operation the book proposes. It comes in four forms: replacing the reader, lesioning or stimulating it, modulating it through adaptation, and rewriting it through learning. In a closed loop it can change which relation a system maintains, as well as what it reports. The fish shows how hard the operation is to perform cleanly: most classical manipulations of its jamming avoidance circuit act on the policy and the motor pathway, downstream of the reader, and a manipulation of the geometry of the mixture can mimic a change of reader so exactly that only a recording from the reader tells the two apart.

58.1 The Operation

In the notation of the reading loop, a reader is the physical operation $z = \mathsf{D}(y)$ that makes a variable of the mixture $y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})$ dynamically explicit (Chapter 15).

Definition 58.1 (Reader intervention)**.** A *reader intervention* replaces D by D' while the distribution of the input y is held fixed and, in a closed loop, while M, π, and F are unchanged. Its effect is the change in the effective variable, $\mathsf{D}'(y) - \mathsf{D}(y)$, and in whatever depends on it.

Together with the relational intervention of Chapter 57, the definition gives a double dissociation: vary the input at a fixed reader, and vary the reader at a fixed input. If the effective variable changes in the second case,

reader dependence is demonstrated for that system; if it never changes for any admissible change of reader, the variable belongs to the input and the reader is transparent to it.

The first practical difficulty is to know which operator a manipulation has changed.

Principle 58.2 (Locate the intervention)**.** Every manipulation of a reading loop must be assigned to the operator it changes: the mixture M, the reader D, the policy π, or the intervention F. The boundary between M and D lies where the variable of interest stops being implicit in a physical mixture and becomes explicit in the state of a component whose activity covaries with it; the boundary between D and π lies where that explicit variable is converted into a command.

The boundaries are modeling choices, but they are not arbitrary: each is fixed by where a named variable first appears. Table 58.1 assigns manipulations of the electric fish and of other systems to operators.

Operator	Manipulation	Source
M (mixture)	Raise water conductivity, degrading electrosensory information; change electrode geometry and so the contamination at each patch of skin	(Stamper, Roth, et al. 2012; Heiligenberg 1991a)
D (reader)	Change a sensory neuron's adaptation state; let a cerebellum-like circuit learn a negative image of predictable input	(Brenner et al. 2000; C. C. Bell 1981; Enikolopov et al. 2018)
π (policy)	Stimulate the "increase" or "decrease" subdivision of the nucleus electrosensorius, writing the read variable directly; lesion one of the two prepacemaker pathways	(C. H. Keller et al. 1990; Metzner 1993b)
F (intervention)	Alter pacemaker modulation pharmacologically; replace the fish's own discharge with an experimenter-driven mimic	(Kawasaki and Heiligenberg 1990; Heiligenberg 1991a)

Table 58.1: Manipulations assigned to the operators of the reading loop.

58.2 Replacement: Swapping the Reader

The purest reader intervention is a replacement, and engineering supplies one that needs nothing beyond signal theory.

Example 58.3 (One signal, two readers). A single-sideband transmission of one tone is the signal $y(t) = \cos\big((\omega_c + \omega_m)t\big)$, a carrier frequency ω_c offset by a modulating frequency ω_m with the carrier itself suppressed. An envelope

detector reads the magnitude of the signal's analytic representation, which here is constant: its output contains no trace of ω_m. A product detector multiplies y by a locally generated $\cos \omega_c t$ and low-pass filters the result:

$$y(t) \cos \omega_c t = \tfrac{1}{2} \cos \omega_m t + \tfrac{1}{2} \cos\left((2\omega_c + \omega_m)t\right) \xrightarrow{\text{low-pass}} \tfrac{1}{2} \cos \omega_m t.$$

The input is identical. One reader extracts the modulating tone; the other extracts nothing.

The modulating tone is a relation between the received frequency and the reference carrier, and the product detector can read it because it supplies the reference. The example is the engineering counterpart of Chapter 4, where a linear filter finds no difference frequency in a sum of two tones and a square-law detector does, and of Chapter 21, where the fish, lacking any internal reference, reads the sign of a frequency difference by comparing regions of its own body. Each is a replacement of D with y fixed, and in each the effective variable changes.

58.3 Lesion and Stimulation in the Fish

The jamming avoidance circuit has been manipulated at several points (Chapter 21). Electrical stimulation of the "increase" and "decrease" subdivisions of the nucleus electrosensorius produces the slow rises and falls of discharge frequency characteristic of the response, without any sensory stimulus (C. H. Keller et al. 1990). Selective lesions and pharmacology of the two prepacemaker nuclei showed that frequency rises and falls are controlled by separate motor pathways in a push-pull arrangement, with no single population of final recognition units (Metzner 1993b).

Assigned by Principle 58.2, these act at the reader's output and downstream of it. Stimulating the nucleus electrosensorius writes the output of the sign computation directly and bypasses the computation that produces it; lesioning a prepacemaker pathway removes one direction of response further downstream. They show how the read variable is used; they do not show what happens when the reader itself is changed.

A clean reader intervention in the fish would alter the combination of amplitude and differential phase in the midbrain while leaving the diencephalic and prepacemaker pathways intact. The comparative evidence suggests where such a manipulation would bite. In *Sternopygus*, the one known gymnotiform genus without a jamming avoidance response, recordings of electrolocation circuitry suggested a substrate from which the response could have evolved (G. Rose, C. Keller, et al. 1987), and *Gymnarchus* and *Eigenmannia* implement the same computational rules in different brain regions (Kawasaki 1996). On this evidence the reader is the part of the loop whose computation is conserved across lineages while its anatomy is not, which makes it the part an experiment would most need to isolate.

58.4 Modulation: Adaptation Changes the Reader

Sensory neurons change their input-output relation with the statistics of their recent input. In the fly's visual motion system, when the dynamic range of the input changes, the neuron's input-output relation rescales to match it, and the rescaling is set so as to maximize the information the response carries about the input (Brenner et al. 2000). Adaptation of this kind is widespread and takes many forms (Wark et al. 2007).

Adaptation is a reader intervention that nature performs continually. Present the same test stimulus after adaptation to a narrow range of inputs and after adaptation to a wide range, and the output differs: the same y yields a different z because D has changed. Two consequences follow for experimental design. Adaptation is a tool: history-dependent changes of the reader can be induced deliberately and then probed with fixed inputs. Adaptation is also a confound: an experiment that claims to hold the input fixed while comparing conditions must hold the reader's adaptation state fixed too, or it has changed D without saying so.

58.5 Learning: Rewriting the Reader

The most informative reader interventions in electric fish come from the pulse-type mormyrids, whose electrosensory lobe receives a corollary discharge of each electric organ command. Bell showed that when a sensory stimulus is paired with the command, the corollary discharge response in the lobe changes into a negative image of the paired sensory response, so that the predictable consequence of the fish's own discharge is subtracted (C. C. Bell 1981). The subtraction is learned by synaptic plasticity in a cerebellum-like circuit (C. Bell et al. 1997; Requarth and Sawtell 2011).

Seen through the reading loop, the learned negative image is a rewriting of D. The input to the lobe contains the fish's reafferent signal and any external signal; before learning, the lobe's output carries both; after learning, it carries the residual, which is the external part. The effective variable has changed from "all input" to "input not predicted from my own command" while the input itself is unchanged. The change has behavioral consequences. Negative images improve both neural coding and behavioral detection of prey-like stimuli, and manipulating the synaptic plasticity that builds them produces specific changes in circuit output that disrupt coding and detection (Enikolopov et al. 2018). That is a reader intervention with a measured behavioral effect, and it is the clearest example in the literature of the operation this chapter describes.

The contrast with the jamming avoidance response is instructive. The mormyrid reader learns to cancel the relation between its own signal and its input, because it needs the world without itself. The wave-type jamming avoidance reader needs exactly that relation and has no corollary discharge to

cancel it with (Kawasaki 1994) (Chapter 22). Rewriting one reader removes the self from the variable; the other reader keeps the self as one of the variable's terms.

58.6 Reader Intervention in a Closed Loop

In an open loop, a reader intervention changes what the system reports. In a closed loop, it can change what the system does to the relation, and so which relation it maintains. A minimal model of the jamming avoidance response makes the point.

Model 58.4 (A reader in the loop)**.** The fish's frequency changes according to $\dot{f}_\mathrm{s} = -k\,\rho(\Delta f)$ with $k > 0$, where ρ is the reader's output, an odd function of $\Delta f = f_\mathrm{e} - f_\mathrm{s}$ with $\rho'(0) > 0$ (for example $\rho(\Delta f) = \tanh(\Delta f / w)$). The neighbor's frequency f_e is constant, so $\dot{\Delta f} = -\dot{f}_\mathrm{s} = k\,\rho(\Delta f)$.

Proposition 58.5 (The reader's sign sets the maintained relation)**.** *In Model 58.4, $\Delta f = 0$ is an unstable equilibrium and $|\Delta f|$ grows from any small nonzero value. If the reader is replaced by $-\rho$, with M, π, and F unchanged, $\Delta f = 0$ becomes asymptotically stable and the fish converges on its neighbor's frequency.*

Proof. Linearizing at $\Delta f = 0$ gives $\dot{\Delta f} \approx k\rho'(0)\,\Delta f$, which grows because $k\rho'(0) > 0$. Replacing ρ by $-\rho$ reverses the sign of the linear coefficient. □

Figure 58.1 shows the two flows. The model is a caricature, but its local instability matches the behavior of the real response, which is unstable near equilibrium unless stabilized by a clamp (Madhav et al. 2013). The proposition predicts that a manipulation which inverted the fish's reading of the sign would convert jamming avoidance into jamming seeking, a qualitative change in the relation the animal maintains produced without touching the policy or the pacemaker.

The prediction runs into an identification problem that the fish makes precise. In the two-patch model of Chapter 21, the reader infers the sign of Δf from the product of the amplitude modulation and the differential phase between a more contaminated and a less contaminated patch. Exchanging which patch is more contaminated reverses the sign of the differential phase and therefore of the inference.

Proposition 58.6 (A mixture change can mimic a reader change)**.** *In the two-patch model, exchanging the contamination ratios of the two patches (an intervention on M) reverses the sign the reader infers, exactly as replacing the reader's rule by its negation (an intervention on D) would. The two interventions produce the same behavior and differ only in the state of the reader's internal variables.*

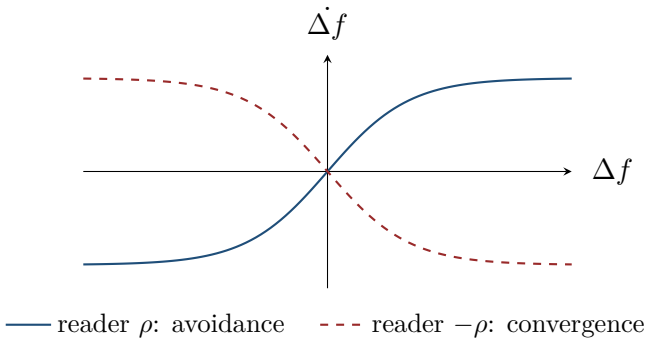


Figure 58.1: Flow of the frequency difference in Model 58.4. With the reader ρ (solid), $\dot{\Delta} f$ has the sign of Δf and the relation moves away from zero; with the inverted reader $-\rho$ (dashed), it moves toward zero. Mixture, policy, and pacemaker are the same in both.

Proof. The inferred sign is $-\operatorname{sgn}\left(\dot{A}\,\delta\right)$, where δ is the phase at the patch the reader treats as more contaminated minus the phase at the other patch. Exchanging m_1 and m_2 exchanges φ_1 and φ_2, reverses δ, and leaves the sign of $\dot{A}$ unchanged; negating the rule reverses the output directly. □

Behavior alone cannot separate the two. A recording from the neurons that combine amplitude and differential phase can: after a change of mixture, they receive reversed inputs and compute correctly; after a change of reader, they receive the usual inputs and compute the opposite. Reader intervention therefore requires a model of the loop and a measurement inside it, not only a manipulation and a behavioral readout.

58.7 Why Reader Intervention Is the Distinctive Operation

Three reasons make reader intervention the operation most specific to this book.

First, it targets the step the book is about. Relational intervention is interventionist causal method applied to relational variables, and the method belongs to the philosophy of causation and to every experimental science (Chapter 57). Reader intervention asks where, and by what, a relation becomes a variable, which is the question of Part IV.

Second, it isolates the factor that no other manipulation isolates. The two-factor claim of the Preface can be tested only by holding each factor fixed while varying the other, and the reader is the factor that experiments usually leave uncontrolled.

Third, in closed loops it changes behavior qualitatively, as Proposition 58.5 shows, which makes it a candidate source of the kind of result Chapter 56 asked for: an effect that the owning fields did not predict.

Hypothesis 58.7 (Reader dependence)**.** For a given system and input distribution, there exist admissible changes of reader (by adaptation, learning, lesion, or pharmacology) that change the effective variable and, in closed loops, the relation the system maintains.

Failure condition. The hypothesis fails for a system if every admissible change of reader leaves the effective variable unchanged, or if every observed effect is traced, by recording, to a change in the mixture or the policy.

The operation has a natural extension to consciousness science. Most tests of theories of consciousness vary the stimulus and record the brain. A reader intervention would hold a neural relation fixed, for instance a phase relation among rhythms maintained by stimulation (Chapter 57), and alter the structures a theory claims read it, then ask whether report changes. Such experiments are harder than stimulus manipulations and much rarer; Chapters 60 and 61 consider how they might be designed.

58.8 Audit

Canonical phenomenon
The same input yields different effective variables when the reader is changed: a single-sideband signal read by envelope and product detectors, a sensory neuron before and after adaptation, and a mormyrid electrosensory lobe before and after learning a negative image.

Minimal model
Reader intervention (Definition 58.1); the assignment of manipulations to operators (Principle 58.2); the reader in the loop (Model 58.4) and its two propositions.

What is relational here?
In the fish, the sign of Δf carried by the joint behavior of amplitude and differential phase; in the mormyrid, the relation between the fish's own discharge and its sensory input; in the radio, the offset of the received frequency from the reference carrier.

What physically reads it?
Product and envelope detectors; an adapting motion-sensitive neuron; the mormyrid electrosensory lobe; the combination sites of the *Eigenmannia* midbrain.

What existing field explains it?
Communication engineering and signal theory for replacement; sensory

neuroscience for adaptation (Brenner et al. 2000; Wark et al. 2007); the physiology of cerebellum-like circuits for learned negative images (C. C. Bell 1981; C. Bell et al. 1997; Requarth and Sawtell 2011; Enikolopov et al. 2018); neuroethology for the jamming avoidance circuit (C. H. Keller et al. 1990; Metzner 1993b).

What remains unexplained?

A clean reader intervention in the jamming avoidance response has not been reported, and the closed-loop prediction of Proposition 58.5 is untested. Crossing reader and relation interventions in one design, as Chapter 56 proposed, is the book's best candidate for a new operation.

A discriminating experiment

Alter the midbrain combination of amplitude and differential phase in *Eigenmannia* while the prepacemaker pathways are intact, and record from the combination sites during the jamming avoidance response. If the response reverses while the neurons receive normal inputs, the reader has been changed; if the neurons receive reversed inputs, the manipulation acted on the mixture (Proposition 58.6).

58.9 Notes and References

Single-sideband reception and product detection are standard in communication engineering; the underlying signal theory is in Oppenheim and Schafer (2010), and practical receivers are described in Horowitz and Hill (2015). Stimulation of the nucleus electrosensorius is reported by C. H. Keller et al. (1990); the two motor pathways by Metzner (1993b); the prepacemaker nucleus by Kawasaki, Maler, et al. (1988); Heiligenberg (1991a) gives the whole circuit. The comparison with *Sternopygus* is G. Rose, C. Keller, et al. (1987); convergent implementation in *Gymnarchus* is reviewed by Kawasaki (1996). Adaptive rescaling is reported by Brenner et al. (2000) and sensory adaptation reviewed by Wark et al. (2007). Learned negative images were discovered by C. C. Bell (1981) and are reviewed by C. Bell et al. (1997) and Requarth and Sawtell (2011); their behavioral function is demonstrated by Enikolopov et al. (2018). Closed-loop dynamics of the jamming avoidance response are analyzed by Madhav et al. (2013).

58.10 Exercises

Exercise 58.1. For $y(t) = \cos\big((\omega_c+\omega_m)t\big) + \cos\big((\omega_c-\omega_m)t\big)$, a double-sideband signal with suppressed carrier, compute the output of an envelope detector and of a product detector with reference $\cos\omega_c t$. Repeat with the reference phase-shifted by $\pi/2$. Which relational variable does each reader make explicit?

Exercise 58.2. Design a double dissociation for the jamming avoidance response: one experiment that varies the mixture at a fixed reader, and one that varies the reader at a fixed mixture. For each, state what you would record and what result would demonstrate reader dependence.

Exercise 58.3. Add a delay to Model 58.4: $\dot{\Delta f}(t) = k\,\rho(\Delta f(t-\tau))$. Show that the equilibrium $\Delta f = 0$ with the inverted reader $-\rho$ becomes oscillatory, and then unstable, as $k\rho'(0)\tau$ grows, and find the threshold. What does this imply for an experiment that inverts the reader in an animal with slow sensorimotor delays?

Exercise 58.4. A reader computes $z = \tanh(g\,y)$ with a gain that adapts to the input's standard deviation, $g = c/\sigma$. Compute the reader's response to a fixed test input y_0 after adaptation to σ_1 and to $\sigma_2 = 3\sigma_1$. Explain why an experiment that compares two conditions with the same test input but different stimulus histories has performed a reader intervention.

Exercise 58.5. Model an electrosensory lobe output as $e_n = s_n + r - w_n$, where r is a fixed reafferent input on each command, s_n an external input uncorrelated with the command, and w_n a learned negative image updated by $w_{n+1} = w_n + \eta\,e_n$. Show that w_n converges to r in expectation, and compute how the signal-to-noise ratio for detecting a small s_n changes as learning proceeds.

Exercise 58.6. Prove Proposition 58.6 using the small-contamination expressions $\varphi_j \approx m_j \sin\psi$ and $A_j \approx A_{\mathrm{s}}(1 + m_j \cos\psi)$. Then specify the recording that distinguishes an exchange of contamination ratios from an inversion of the reader's rule, and state its predicted outcome in each case.

Exercise 58.7. Choose a neural relation that some theory proposes as a correlate of consciousness. Describe a reader intervention for it: how the relation would be held fixed, which structure the theory claims reads it, how that structure would be altered, and what change in report each outcome would imply.

59

Closed-Loop Experiments

A fish that shifts its discharge away from a neighbor's changes the relation it is sensing. An experimenter who wants to measure how the fish reads that relation faces a difficulty that no open-loop measurement removes: the animal's response alters the stimulus, and the altered stimulus alters the response. Madhav, Stamper, Fortune, and Cowan found that the jamming avoidance response of *Eigenmannia virescens* is unstable in exactly this sense. Presented with a jamming stimulus at a fixed frequency, the fish's own frequency diverges from it; the equilibrium at zero frequency difference repels. They stabilized the behavior by tying the instantaneous frequency of the stimulus to the fish's own electrogenic frequency through a feedback law, and with the loop closed around the animal they could measure what the fixed-stimulus experiment could not: a delayed first-order linear system described the behavior near equilibrium, and a nonlinear model of competing motor-return and sensory-escape terms predicted a "snap-through" bifurcation as a function of the frequency difference, which they then confirmed (Madhav et al. 2013).

The thesis of this chapter is that for a system that participates in producing the relation it reads, the unit of experiment is the loop. The decisive manipulations set the loop's gain, open it, close it, or replay its inputs with the loop broken. Control theory already predicts that closing a loop changes a system's dynamics, and reafference theory already predicts that it changes what information is available. The question these experiments can still answer is narrower: whether the reader's own transformation depends on whether the organism can alter the relation it reads.

59.1 Open and Closed Loops

In the notation of Chapter 15, an agent with state s_{A} and an environment with state s_{E} are mixed at a receiving surface, $y = \mathsf{M}(s_{\mathrm{A}}, s_{\mathrm{E}})$; a reader makes a variable explicit, $z = \mathsf{D}(y)$; a policy chooses an action, $a = \pi(z)$; and the action changes the agent, $\dot{s}_{\mathrm{A}} = \mathsf{F}(s_{\mathrm{A}}, a)$. What the experimenter controls is the environment's law of motion.

Definition 59.1 (Loop conditions)**.** Let the environment's state follow $s_{\mathrm{E}}(t) = G(s_{\mathrm{A}}(t), t)$. The experiment is *open-loop* if G does not depend on s_{A}, and *closed-loop* if it does. An *experimenter-closed* loop is a closed loop whose law G is chosen and implemented by the experimenter. If the variable read is a relation $r = \mathcal{R}(s_{\mathrm{A}}, s_{\mathrm{E}})$, its *relational loop gain* is the total sensitivity of the relation to the agent's own state,

$$\lambda = \frac{\mathrm{d}r}{\mathrm{d}s_{\mathrm{A}}} = \frac{\partial \mathcal{R}}{\partial s_{\mathrm{A}}} + \frac{\partial \mathcal{R}}{\partial s_{\mathrm{E}}} \frac{\partial G}{\partial s_{\mathrm{A}}}. \tag{59.1}$$

The definition makes "open" and "closed" relative to a variable, and the fish shows why that matters. Let the relation be the frequency difference, $r = \Delta f = f_{\mathrm{e}} - f_{\mathrm{s}}$, and let the agent's state be the fish's own frequency f_{s}. Four conditions cover the experiments of this chapter.

- *Fixed stimulus*: f_{e} is held constant. The stimulus ignores the fish, and $\lambda = -1$: every hertz the fish moves changes the relation by a hertz.
- *Relation clamp*: $f_{\mathrm{e}}(t) = f_{\mathrm{s}}(t) + \Delta_{\mathrm{c}}$ for a chosen constant Δ_{c}. The stimulus follows the fish, so in absolute terms the loop is closed; in relational terms it is open, since $\lambda = 0$ and nothing the fish does changes the relation it reads.
- *Proportional feedback*: $f_{\mathrm{e}}(t) = f_{\mathrm{ref}} + \kappa\,(f_{\mathrm{s}}(t) - f_{\mathrm{ref}})$ for a reference frequency f_{ref} and gain κ. Then $\Delta f = (\kappa - 1)(f_{\mathrm{s}} - f_{\mathrm{ref}})$ and $\lambda = \kappa - 1$. The fixed stimulus ($\kappa = 0$) and the relation clamp ($\kappa = 1$) are special cases.
- *Natural encounter*: two fish, each performing the response. Each fish's action changes the relation, and so does its neighbor's.

Remark 59.2. A relation clamp keeps the stimulus coupled to the animal and yet removes the animal's influence on what it reads. An experiment described as "closed-loop" because the stimulus tracks the animal can therefore be open-loop in the variable that matters. Reports of such experiments should state the relational loop gain, and not only whether the apparatus responds to the animal.

59.2 A Toy Model of the Jamming Loop

Madhav and colleagues describe their fitted nonlinear model as composed of competing motor-return and sensory-escape terms (Madhav et al. 2013). The model below has that structure. Its specific functional forms are ours, chosen so that each step can be followed by hand, and it should not be read as their fitted model.

Model 59.3 (Return and escape, illustrative)**.** The fish's frequency obeys

$$\frac{\mathrm{d}f_{\mathrm{s}}}{\mathrm{d}t} = -\frac{f_{\mathrm{s}} - f_0}{\tau} - g(\Delta f), \tag{59.2}$$

where f_0 is the fish's resting frequency, $\tau > 0$ is the time constant of the return toward it, and g is an odd, tuned escape function: $g(\Delta f) > 0$ for small positive Δf, $g'(0) = G > 0$, and $g(\Delta f) \to 0$ as $|\Delta f|$ grows. The sign convention follows the jamming avoidance rule (Chapter 20): a neighbor above the fish ($\Delta f > 0$) drives the fish's frequency down.

The first result shows how feedback can hold an animal at an equilibrium that the animal's own behavior repels.

Proposition 59.4 (Stabilization by feedback, in the toy model)**.** *Neglect the return term ($\tau \to \infty$), linearize $g(\Delta f) \approx G\,\Delta f$ near zero, and apply proportional feedback with gain κ. The equilibrium $f_s = f_{ref}$, at which $\Delta f = 0$, is asymptotically stable if $\kappa > 1$, unstable if $\kappa < 1$, and neutral at $\kappa = 1$.*

Proof. With $u = f_s - f_{ref}$, proportional feedback gives $\Delta f = (\kappa - 1)\,u$, so (59.2) becomes $\dot{u} = -G(\kappa - 1)\,u$. The solution decays if and only if $G(\kappa - 1) > 0$, and $G > 0$ by assumption. □

With a fixed stimulus ($\kappa = 0$) the jamming equilibrium repels, which is what the behavior is for: the fish escapes the relation $\Delta f = 0$. With $\kappa > 1$ the stimulus follows each move of the fish and overshoots it, so every attempted escape shrinks $|\Delta f|$, and the animal is held where its local dynamics can be measured. The relation clamp sits at the boundary.

Real sensorimotor loops are delayed, and Madhav and colleagues found that a delayed first-order linear model fitted the behavior near the equilibrium (Madhav et al. 2013). A delay changes the stabilization result in a way that no open-loop measurement anticipates.

Proposition 59.5 (A window of stabilizing gains, in the delayed toy model)**.** *Let the linearized closed loop act with delay $d > 0$, so that $\dot{u}(t) = -a\,u(t - d)$ with $a = G(\kappa - 1)$. For $a > 0$ the equilibrium is asymptotically stable if and only if $ad < \pi/2$. At $ad = \pi/2$ a pair of characteristic roots crosses the imaginary axis with frequency $\omega = a$, and for larger gains the linearized loop oscillates with growing amplitude.*

The stability boundary follows from the characteristic equation $s = -a\,\mathrm{e}^{-sd}$, which has a root $s = \mathrm{i}\omega$ exactly when $\cos(\omega d) = 0$ and $\omega = a\sin(\omega d)$, that is, when $\omega d = \pi/2$ and $\omega = a$ (Exercise 59.2). Delayed negative feedback of this kind is a standard source of oscillation in physiological control (Glass and Mackey 1988). The consequence for experiments is concrete. Too little feedback gain fails to stabilize the animal; too much, combined with the animal's processing delay, drives its frequency into oscillation. In the nonlinear system such oscillations saturate. The closed system has two behaviors, a held equilibrium and a delay-induced oscillation, that belong to none of its parts studied open-loop.

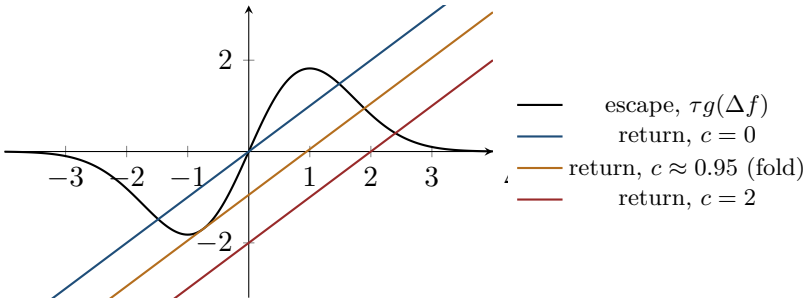


Figure 59.1: Equilibria of the fixed-stimulus toy model, $\tau g(\Delta f) = \Delta f - c$, with illustrative parameters $G\tau = 3$ and unit tuning width. The escape curve (black) meets each return line in one or three points. For $c = 0$ there are three equilibria: the middle one repels and the outer two attract. At $c \approx 0.95$ the line is tangent to the curve and two equilibria merge; beyond it only one remains, and a fish sitting on the vanished branch jumps to the other side of the stimulus.

The return term produces the global nonlinearity. With a fixed stimulus, write $x = \Delta f$ and $c = f_e - f_0$ for the offset of the stimulus from the fish's resting frequency; (59.2) becomes $\dot{x} = (c - x)/\tau + g(x)$. Equilibria satisfy $\tau g(x) = x - c$, and an equilibrium is stable when $g'(x) < 1/\tau$. When escape dominates return near zero, $G\tau > 1$, the equilibrium near $\Delta f = 0$ repels and two stable equilibria flank it; as c varies, pairs of equilibria merge in saddle-node bifurcations (Figure 59.1). A fold of this kind is one generic way in which competing return and escape terms produce hysteresis and sudden jumps; the snap-through that Madhav and colleagues predicted and observed is a bifurcation of the same family, located in their own parameterization of the frequency difference (Madhav et al. 2013). The relation clamp gives the most direct measurement of the escape function: with Δf held at Δ_c, the fish's frequency relaxes to $f_0 - \tau g(\Delta_c)$, so the steady shift traces g point by point.

59.3 Held and Hein's Carousel

The fish experiments set the loop's gain. The oldest closed-loop experiment in perception breaks the loop while keeping the input. Held and Hein reared kittens in darkness and then exposed them in pairs to a lit, striped cylinder, in an apparatus that linked the two animals (Held and Hein 1963). One kitten walked freely and, through a transmission mechanism, moved the second, which rode in a gondola that prevented locomotion with its own legs. The visual stimulation of the two animals was yoked: whatever the active kitten's walking produced on its retina, the passive kitten saw too. The active kittens developed visually guided behavior, tested by visually guided paw placement, avoidance of a visual cliff, and blinking to an approaching object; the passive kittens did not.

In the terms of Definition 59.1, the two kittens received inputs with matched

statistics and different loop status. For the active kitten the visual flow depended on its own movement, and the relation between movement and visual change was available in its input. For the passive kitten the same flow arrived with that relation absent: its movement-to-vision gain was zero. The experiment shows that the variable driving development was the relation between self-produced movement and visual change, since the visual flow itself was matched between the animals.

Two cautions keep the interpretation honest. The yoked design equates visual input and leaves other differences uncontrolled: the passive kitten was restrained, its vestibular stimulation differed, and its arousal may have differed. And the result concerns development, a loop closed over days, while the fish experiments concern a loop closed over seconds. The sensorimotor account of perception generalizes the lesson: to perceive, on that account, is to master the lawful relations between one's movements and the resulting changes in sensory input, relations that exist only in a closed loop (O'Regan and Noë 2001). Reafference theory had already said as much about the nervous system's bookkeeping (Chapter 22).

59.4 Active Sensing Movements in the Fish

Eigenmannia hides in refuges, and when a refuge moves the fish follows it. Stamper, Roth, Cowan, and Fortune varied the sensory information available during this tracking task by changing the illumination and the water's conductivity (Stamper, Roth, et al. 2012). The gain between refuge movement and tracking response was the same in every condition. In the dark, however, tracking error rose, because the fish produced spontaneous whole-body oscillations at 0.1 to 1 Hz; these movements were costly (fish swam more than three times further while tracking in the dark) and grew larger as electrosensory salience was degraded by raising conductivity. Tail bending at 1.5 to 2.35 Hz appeared only in the dark. The movements worsen tracking and reshape sensing: they alter the spatiotemporal pattern of electrosensory feedback.

Biswas and colleagues then closed an experimenter-controlled loop around those movements (Biswas et al. 2018). Their apparatus measured the freely swimming fish's position by video in real time and moved the refuge accordingly, which let them change the gain of reafferent feedback, the amount by which the fish's own movement changed its sensory input. Across a four-fold change of this gain, the fish raised or lowered its fore-aft active sensing movements so as to hold the magnitude of *sensory slip*, the relative motion it sensed, constant; the level of slip it maintained depended on whether visual cues were present. The authors conclude that the fish uses two controllers, one that keeps it in the refuge and one that regulates the acquisition of information through its own movements.

Example 59.6 (A regulated relation). The Biswas experiment dissociates a relation from its constituents. When the experimenter changed the loop gain, the constituent the fish controls, its own movement, changed, and the relation, slip between fish and refuge, did not. A controller that regulated movement amplitude would have held movement constant and let slip vary with gain. The data therefore identify the regulated variable as relational, by the intervention-sensitivity criterion of Chapter 1, and they do so because the experimenter could set the gain of the loop.

The control-theoretic framing of such experiments is explicit in this literature: each behavior is cast as a control problem with sensory input, neural processing, and motor dynamics whose output feeds back to the sensory input, and the closed-loop interaction determines the behavior of the whole (Cowan et al. 2014). The same logic has moved inside the nervous system. In closed-loop neuroscience, stimulation is delivered contingent on the recorded state of the neural tissue itself (Potter et al. 2014), which makes it possible to intervene on neural relations, such as the phase at which a population is stimulated, rather than on neural activity alone. Chapter 60 uses that possibility.

59.5 What the Loop Adds

Closing a loop can make three different kinds of difference, and they have different owners.

Dynamical. Equilibria, their stability, and the system's oscillations change when a loop is closed; Propositions 59.4 and 59.5 are instances. Control theory owns this completely.

Informational. In a closed loop the input carries the consequences of the agent's own actions, so the relation between action and sensation is itself in the data. This is the substance of reafference (Chapter 22) and of the sensorimotor account (Section 59.3). It has a methodological shadow: in a closed loop the input is correlated with the agent's own past output, so estimating the agent's input-output properties from passive observation of the loop is biased unless an external reference signal, uncorrelated with the loop's internal noise, is injected (Exercise 59.5).

Reading. The third possibility is that the reader's transformation D itself depends on loop status: that the same mixture y is read differently when the agent produced it than when it did not. For some systems this is established. In pulse-type mormyrid fish, a corollary discharge from the electromotor command acts on cerebellum-like circuits of the electrosensory lateral line lobe, which cancel predictable reafference with a plastic negative image; identical electrosensory input is processed differently when it follows the fish's own discharge (C. C. Bell 1981; C. Bell et al. 1997; Requarth and Sawtell 2011). For the wave-type jamming avoidance response the literature points the other way: the response is computed from the electrosensory mixture alone, with no

internal reference to the pacemaker command (Heiligenberg 1991a; Kawasaki 1993), and *Gymnarchus* lacks corollary-discharge projections to its sensory areas (Kawasaki 1994).

The contrast yields a test that could fail in either direction.

Hypothesis 59.7 (Loop-invariant reading in the wave-type jamming response)**.** Record the self signal and the neighbor signal during closed-loop jamming behavior, then replay the same time series to the same animal in a preparation in which the fish's discharge is replaced by an electronic mimic, so that the animal's pacemaker no longer affects its input. In *Eigenmannia* and *Gymnarchus*, the responses of sign-selective midbrain neurons to a given beat pattern will not differ between the closed-loop and replay conditions, beyond differences attributable to the neurons' ongoing adaptation. In a mormyrid given the analogous replay of reafferent input, responses in the electrosensory lateral line lobe will differ between the two conditions.

The hypothesis is a proposal, and it states what the book's framework adds and what it does not. Both predictions follow from existing neuroethology, which is exactly the point: the reading loop predicts loop-dependent reading where the anatomy supplies a copy of the motor command and loop-invariant reading where it does not. A difference in the wave-type fish would show that something other than a corollary discharge makes its reading depend on whether it can alter the relation, and that would be a genuinely new result. The chapter's question, whether letting the organism alter the relation creates qualitatively different dynamics, therefore has three answers. For dynamics, yes, and control theory says why. For available information, yes, and reafference theory says why. For the reader's transformation, it depends on the circuit, and the wave-type fish is the place to find out whether anything beyond known circuitry is involved.

59.6 Audit

Canonical phenomenon

The jamming avoidance response under experimenter-controlled feedback, which repels its equilibrium with a fixed stimulus and can be held at it by feedback (Madhav et al. 2013); kittens with yoked visual input and different loop status (Held and Hein 1963); fish that hold sensory slip constant across a four-fold change of reafferent gain (Biswas et al. 2018).

Minimal model

The reading loop with an environmental law $s_{\mathrm{E}} = G(s_{\mathrm{A}}, t)$ and the relational loop gain (59.1); the return-and-escape toy model (Model 59.3), its stabilization window under proportional feedback and delay (Propositions 59.4 and 59.5), and its folds (Figure 59.1).

What is relational here?
The frequency difference Δf and its sensitivity to the fish's own frequency; the relation between movement and visual change in the carousel; sensory slip between fish and refuge.

What physically reads it?
The electrosensory pathway that extracts the sign of Δf (Chapter 21); the developing visuomotor system of the active kitten; the active-sensing controller that measures slip.

What existing field explains it?
Control theory and its application to animal behavior (Cowan et al. 2014), including stabilization, delay-induced oscillation, and closed-loop identification; the neuroethology of the jamming response (Madhav et al. 2013; Heiligenberg 1991a); reafference and corollary discharge (C. C. Bell 1981; Requarth and Sawtell 2011); the sensorimotor account of perception (Held and Hein 1963; O'Regan and Noë 2001); closed-loop neuroscience (Potter et al. 2014).

What remains unexplained?
Whether a reader's transformation depends on loop status in a circuit that carries no copy of the motor command. Everything else in the chapter is explained.

A discriminating experiment
The replay experiment of Hypothesis 59.7: identical beat patterns delivered in closed loop and in replay to wave-type fish, with a mormyrid comparison. A replay difference in the wave-type fish, not attributable to adaptation, would support loop-dependent reading without a corollary discharge; its absence confirms the existing account.

59.7 Notes and References

Madhav et al. (2013) is the model for experiments that close a loop around an animal to measure unstable behavior, and Cowan et al. (2014) surveys the control-theoretic program to which it belongs, with case studies in electric fish, cockroaches, and moths. The active-sensing experiments are Stamper, Roth, et al. (2012), which established the costly whole-body oscillations and tail bends that the fish produce when electrosensory information is degraded, and Biswas et al. (2018), which manipulated reafferent gain in closed loop. The classical experiment on movement-produced stimulation is Held and Hein (1963); O'Regan and Noë (2001) develop the sensorimotor account of perception that generalizes it, and the neuroscience of reafference appears in Chapter 22. Mormyrid negative images and their plasticity are reviewed by C. Bell et al. (1997) and Requarth and Sawtell (2011). On delayed negative feedback as a

source of physiological oscillation, see Glass and Mackey (1988). Closed-loop methods within the nervous system are introduced by Potter et al. (2014). The toy model of Section 59.2 is ours and makes no claim about the functional forms fitted by Madhav and colleagues.

59.8 Exercises

Exercise 59.1. Compute the relational loop gain (59.1) for the four conditions of Section 59.1. For the natural encounter, let both fish obey (59.2) with the same escape function and show that the frequency difference obeys an equation whose linearization at $\Delta f = 0$ has twice the escape rate of the fixed-stimulus case.

Exercise 59.2. For $\dot{u}(t) = -a\,u(t-d)$ with $a, d > 0$, substitute $u = \mathrm{e}^{st}$ to obtain $s = -a\,\mathrm{e}^{-sd}$. Show that a purely imaginary root $s = \mathrm{i}\omega$ requires $\omega d = \pi/2 + k\pi$ and $\omega = a\sin(\omega d)$, and conclude that the first crossing occurs at $ad = \pi/2$. Simulate the equation for $ad = 1.4$ and $ad = 1.7$ and compare.

Exercise 59.3. For $\tau g(x) = G\tau\, x\,\mathrm{e}^{-x^2/2}$, find the condition on $G\tau$ for the equilibrium at $x = 0$ to repel when $c = 0$. Locate the folds by solving $\tau g'(x) = 1$ together with $c = x - \tau g(x)$, and verify that $c \approx \pm 0.95$ when $G\tau = 3$. Sketch the equilibrium x against c and mark the hysteresis loop.

Exercise 59.4. Show that under a relation clamp $\Delta f = \Delta_{\mathrm{c}}$, Model 59.3 relaxes to $f_{\mathrm{s}} = f_0 - \tau g(\Delta_{\mathrm{c}})$. Design a sequence of clamps that estimates g and τ separately, and state which additional measurement is needed if the return is not first-order.

Exercise 59.5. Consider a loop $y = Pu + n$, $u = -Cy + w$, with scalar gains P, C, measurement noise n, and an external reference w independent of n. Show that the least-squares slope of y on u is a biased estimate of P when $n \neq 0$, and that the ratio $\mathrm{Cov}(y, w)/\,\mathrm{Cov}(u, w)$ recovers P. What plays the role of w in the Madhav experiment?

Exercise 59.6. In an idealized version of the Biswas apparatus, suppose the refuge moves by γ times the fish's own fore-aft displacement, so that the relative displacement sensed is $(1 - \gamma)$ times the fish's displacement. If the fish holds the amplitude of relative displacement at s^{*}, how must its movement amplitude depend on γ? What pattern would instead indicate that the fish regulates its own movement amplitude? This idealization is ours; check your conclusions against the published design before using them.

Exercise 59.7. Write the reading loop $(\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$ for the active and the passive kitten of Section 59.3. Which component differs, and which relational loop gain is zero for the passive animal? List two uncontrolled differences between the animals and propose a control condition for each.

Exercise 59.8. Specify the replay experiment of Hypothesis 59.7 in enough detail to preregister it: which signals are recorded, how the self-signal mimic is generated, what must be matched across conditions (beat rate, amplitude, and the spatial pattern of contamination over the body), which neurons are recorded, and which result would count against the hypothesis.

60

Consciousness Experiments

Every quantity that a theory of consciousness cares about changes when consciousness does. Anesthesia, sleep, and brain injury alter power spectra, synchrony, complexity, connectivity, and responsiveness together, so a measure that tracks the state of consciousness has shown only that it belongs to a large, correlated family. Relational measures (phase relations between regions, cross-frequency coupling, intermodulation, bicoherence, perturbational complexity) are no exception, and the bispectral index shows how easily a relational component can ride on a constituent one. When Miller and colleagues compared the bispectral parameter of the commercial monitor with an analogous parameter computed from the ordinary power spectrum during induction of anesthesia in 39 patients, the two correlated at $r = 0.84$, neither was superior in predicting whether a patient was awake or anesthetized, and an amplitude-independent bicoherence version of the bispectral parameter did not change over induction at all (Miller et al. 2004). Most of the change in the bispectral values, they concluded, came from the decrease in the relative high-frequency content of the EEG.

The thesis of this chapter is that a relational neural measure earns a place in consciousness science only when, in a preregistered experiment, it tracks conscious state or content better than a matched constituent measure computed from the same data. Five experiments follow, each built around that rule and each paired with its simpler null model. All five are proposals. Parts of each rest on published findings, which we report as such; none has been run in the form described here.

60.1 The Rule of the Null Model

A relational measure is computed from two or more signals and is a function of none of them alone (Chapter 1). The comparison it needs is with measures that use the same data and treat each signal separately.

Definition 60.1 (Matched constituent measure)**.** Let $R(x_1, \dots, x_k)$ be a relational measure computed from signals $x_1, \dots, x_k$. A *matched constituent*

measure is a predictor C computed from the same recordings and trials, with no more free parameters than R, that combines quantities each of which depends on a single signal: its power spectrum, the offset and exponent of its aperiodic component, its peak frequency, its evoked amplitude and duration, its waveform shape.

Definition 60.2 (Relation-destroying surrogate)**.** A *relation-destroying surrogate* of a data set preserves the constituent properties of each signal and destroys the relation that R measures, for example by randomizing Fourier phases independently across channels or by shifting each channel's trials in time by independent random offsets. The distribution of R over such surrogates is its null distribution (Theiler et al. 1992; Lancaster et al. 2018). Appendix J gives the constructions and their assumptions.

Principle 60.3 (Beat the null)**.** A relational measure is credited with information about consciousness only if both of the following hold, in a comparison whose data, measures, and statistics were fixed before the data were seen:

1. its observed values fall outside its relation-destroying surrogate distribution; and
2. it improves out-of-sample prediction of conscious state or report over its matched constituent measure.

The first condition establishes that there is a relation to measure. The second establishes that the relation carries something the constituents do not. The first condition is the one most often tested; the second is the one the argument of this book needs.

Four confounds make the second condition hard to meet, and each has its own literature. Neural power spectra combine oscillatory peaks with an aperiodic, roughly $1/f$ component, and standard band-power analyses conflate changes in the aperiodic offset and exponent with changes in oscillations (Donoghue et al. 2020). Nonsinusoidal waveforms produce spurious phase-amplitude coupling, because a sharp waveform has harmonics whose amplitudes are locked to its own phase (Cole and Voytek 2017; Aru et al. 2015). Volume conduction mixes a single source into many scalp channels with zero lag, which inflates coherence and phase locking between them (Nolte et al. 2004). And phase-locking statistics are biased upward in small samples (Vinck, Wingerden, et al. 2010). Appendix K discusses each.

Example 60.4 (Clinical use and mechanism)*.* The bispectral index is instructive beyond the Miller result, because it separates clinical usefulness from mechanistic interpretation. In the B-Aware trial, 2,463 adults at high risk of awareness were randomized to anesthesia guided by the index or to routine care; there were 2 reports of awareness in the guided group and 11 in the routine group, a risk reduction of 82% (Myles et al. 2004). In the B-Unaware trial, 2,000 patients were randomized to guidance by the index or by end-tidal anesthetic

gas concentration; definite awareness occurred in two patients in each group, and it occurred even when the monitored values were within their target ranges (Avidan et al. 2008). Whatever the clinical lesson of the two trials, neither bears on whether the bispectral component of the index measures anything the power spectrum does not. A composite index can guide practice while its relational ingredient does no work.

60.2 Proposal 1: Entrainment Tongues Across Sedation

A rhythmic drive entrains an oscillator only when the drive is strong enough relative to the detuning between the drive's frequency and the oscillator's own; plotted against detuning and drive strength, the region of entrainment is an Arnold tongue (Chapter 7, Appendix F). The tongue is a relational object in the sense of this book: it describes how the relation between drive and oscillator, their detuning, governs whether a further relation, phase locking, appears.

Human alpha rhythms show such tongues. Notbohm, Kurths, and Herrmann stimulated participants with rhythmic flicker at several frequencies and intensities and compared phase coupling with that produced by arrhythmic, jittered flicker; phase coupling grew with intensity and with the closeness of the flicker frequency to each participant's intrinsic frequency, only conditions inside the tongue differed from jittered stimulation, and phase locking was intermittent within single steady-state sequences, a nonlinear feature that a linear superposition of evoked responses cannot produce (Notbohm et al. 2016). Rhythmic transcranial magnetic stimulation tuned to an individual's parietal alpha frequency entrains that rhythm, with the three signatures entrainment predicts: alpha is induced, it builds up over the train, and it depends on the phase of the ongoing rhythm before stimulation (Thut et al. 2011). Network models show resonance with an Arnold tongue centered on the network's resonance frequency (Ali et al. 2013), and further mechanisms, including nonlinear acceleration at high stimulation frequencies (Herrmann et al. 2016).

The constituent null for this proposal comes from the same theory that defines the tongue. For a weakly forced limit-cycle oscillator reduced to its phase, the phase difference ϕ between oscillator and drive obeys the Adler equation

$$\dot{\phi} = (\omega_0 - \Omega) - \varepsilon_{\text{eff}} \sin \phi, \tag{60.1}$$

which has a locked solution if and only if $|\omega_0 - \Omega| \leq \varepsilon_{\text{eff}}$ (Pikovsky et al. 2001). For the normal form of an oscillator near a Hopf bifurcation, ε_{eff} is proportional to the drive strength divided by the oscillator's intrinsic amplitude r_0 (Exercise 60.1). A state change that strengthens or weakens the intrinsic alpha rhythm, or shifts its frequency, therefore changes the tongue without changing anything relational about the coupling (Figure 60.1).

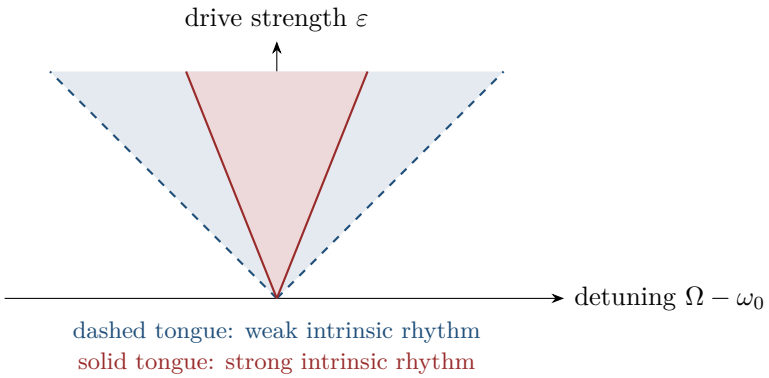


Figure 60.1: Schematic entrainment tongues, not data. In the phase-reduced model of a forced oscillator the half-width of the tongue at drive strength ε is proportional to ε/r_0, where r_0 is the amplitude of the intrinsic rhythm. A stronger intrinsic rhythm therefore narrows the tongue with no change in how the drive couples to the oscillator.

Hypothesis 60.5 (Tongue width tracks experience)**.** Map the alpha entrainment tongue (phase coupling as a function of flicker frequency and intensity, relative to jittered flicker) in each participant during wakefulness, propofol sedation, and ketamine sedation matched for behavioral unresponsiveness. The tongue's width will follow the presence of reported experience, wide in wakefulness and under ketamine and narrow under propofol, beyond what the forced-oscillator null predicts from each participant's intrinsic peak frequency, peak amplitude (measured after removing the aperiodic component), and aperiodic exponent in each state.

The ketamine and propofol contrast is what makes the proposal discriminating. In a study of perturbational complexity during anesthesia, participants reported no conscious experience after emergence from propofol or xenon, and long, vivid dreams unrelated to the environment after ketamine, although all three drugs produced profound unresponsiveness (Sarasso et al. 2015). A measure that follows responsiveness will treat the two drugs alike; a measure that follows reported experience will separate them. The hypothesis fails if tongue width is fully predicted by the constituent null, if it follows responsiveness rather than reports, or if the rhythmic and jittered conditions do not differ under sedation, which would leave no tongue to measure.

60.3 Proposal 2: Intermodulation with Configural Controls

Two inputs tagged at frequencies f_1 and f_2 produce responses at f_1, f_2, and their harmonics in any system; responses at intermodulation frequencies such as $f_1 - f_2$ and $f_1 + f_2$ appear only if some nonlinearity receives both inputs

(M. Regan and D. Regan 1988; Norcia et al. 2015). Chapters 3 and 4 made the same point about beats: a sum of two tones contains no difference frequency until a nonlinear reader puts one there. Intermodulation is therefore a physical trace of a reader that combines two inputs.

Boremanse, Norcia, and Rossion applied the technique to faces (Boremanse et al. 2013). They recorded high-density EEG from 15 participants while the left and right halves of a face flickered at 5.88 Hz and 7.14 Hz. Besides responses at the two fundamental frequencies, which reflected the parts, they found intermodulation components (for example at $7.14 - 5.88 = 1.26$ Hz) over the right occipito-temporal hemisphere. The part-based responses did not depend on the alignment of the halves, their separation, or the orientation of the face. The intermodulation components were almost absent when the halves were misaligned and separated, and inversion of the face reduced them specifically over right occipito-temporal cortex. The configural controls are what license the interpretation: intermodulation that survives scrambling of the configuration would indicate a generic nonlinearity, while intermodulation that depends on the configuration indicates a reader of the whole.

The constituent null is a static nonlinearity downstream of the two part responses. If each half drives a population with amplitude a_i at frequency f_i, and a common stage applies $y = x + \beta x^2$, then the response at $f_1 \pm f_2$ has amplitude $\beta a_1 a_2$ (Exercise 60.2). Under that null, intermodulation carries no information that the product of the fundamental amplitudes does not.

Hypothesis 60.6 (Intermodulation tracks the seen whole)**.** Present aligned face halves, tagged at two frequencies, to one eye and a competing image to the other, so that the physical stimulus stays constant while perception alternates between them. Intermodulation power over right occipito-temporal cortex will be higher in periods the participant reports as face-dominant than in periods reported as suppressed, after regressing out the fundamental amplitudes and their product within the same periods. With misaligned halves presented under the same rivalry, the same analysis will show no such difference.

The first prediction tests whether the intermodulation carries information about the report beyond its constituents; the second tests whether the information is configural. The hypothesis fails if intermodulation adds nothing after the regression, or if it predicts report equally for misaligned halves, in which case it marks a generic nonlinearity rather than a reading of the configuration.

60.4 Proposal 3: States that Separate Responsiveness from Experience

Perturbation separates the brain's capacity for integrated activity from its sensory and motor traffic. Massimini and colleagues stimulated premotor cortex with transcranial magnetic stimulation and recorded high-density EEG:

during quiet wakefulness an initial response of about 15 ms at the stimulation site was followed by a sequence of waves moving to connected cortical areas several centimeters away, while during non-rapid-eye-movement sleep the initial response was stronger but was rapidly extinguished and did not propagate (Massimini et al. 2005). The perturbational complexity index compresses the spatiotemporal pattern of such responses to measure its algorithmic complexity; it discriminated the level of consciousness in single individuals across wakefulness, dreaming, non-rapid-eye-movement sleep, sedation with midazolam, xenon, and propofol, and patients who had emerged from coma (Casali et al. 2013). Under anesthesia, stimulation evoked a local, low-complexity pattern with propofol, a global but stereotyped low-complexity pattern with xenon, and a wakefulness-like complex pattern with ketamine, matching the participants' reports on emergence (Sarasso et al. 2015). Spontaneous EEG carries state markers of its own: loss of consciousness under propofol is marked by increased power below 1 Hz, loss of coherent occipital alpha, and the appearance of coherent frontal alpha, with the phase of the slow oscillation modulating alpha amplitude in two distinct patterns during profound unconsciousness and during the transitions into and out of it (Purdon et al. 2013; Mukamel et al. 2014).

The complexity index is built on relations: its integration component requires that stimulation spread among cortical sites, and its compression depends on the joint spatiotemporal pattern. Whether its discriminating power comes from those relations or from simpler properties of each channel's response is an open question that a relation-destroying surrogate can settle.

Hypothesis 60.7 (Relational structure carries perturbational complexity)**.** Recompute perturbational complexity on surrogates that shift each channel's evoked response by an independent random latency within the response window, preserving each channel's amplitude and time course and scrambling the relative timing across channels. On the original data the index will separate report-positive states (wakefulness, dreaming sleep, ketamine) from report-negative states (non-rapid-eye-movement sleep without reports, propofol, xenon); on the surrogates most of that separation will vanish. Matched constituent measures (per-channel evoked amplitude and duration, number of responsive channels, slow-wave power, aperiodic exponent) will separate ketamine from propofol less well than the original index.

The surrogate leaves every constituent unchanged by construction, so a surviving separation would show that the index's discrimination is carried by what each channel does, and the relational reading of the index would lose its support. The ketamine and propofol pair again carries the weight, because it holds behavioral unresponsiveness fixed while reported experience changes.

60.5 Proposal 4: Active Versus Passive Contingencies

Chapter 59 showed that the relation between an agent's movement and the resulting sensory change is available only in a closed loop, and that this relation drove the development of visually guided behavior in Held and Hein's kittens (Held and Hein 1963). The sensorimotor account holds that perceptual experience consists in the exercise of mastery of such relations (O'Regan and Noë 2001). A rival account of self-generated input already exists: a prediction derived from the motor command attenuates the sensory consequences of one's own actions, as in the reduced response to self-produced touch (Blakemore et al. 1998). The two accounts can be separated, because the motor command and the sensorimotor relation can be separated.

Hypothesis 60.8 (Contingency, not command, changes reportability)**.** Deliver near-threshold stimuli, identical in physical input, under three conditions: contingent on the participant's own active movement; contingent on the same movement imposed passively by a robot, with no motor command; and as a yoked replay with no contingency. Detection rates and reported vividness will be equal in the two contingent conditions and will differ from the replay condition by more than the difference in early evoked sensory responses predicts.

A result in which only the active condition differs, and the difference is accounted for by attenuation of the early response, would favor the motor-prediction account and count against the relational one (Exercise 60.6).

60.6 Proposal 5: Relational Against Univariate, Everywhere

The four proposals share a design that can be applied to any relational measure. Table 60.1 lists the pairings.

The analysis plan follows from Principle 60.3. Fix the data sets, the measures, and the comparisons in advance. Remove the aperiodic component before computing any band-limited measure (Donoghue et al. 2020). For each relational measure, report its position in its surrogate distribution and its incremental out-of-sample predictive value over its matched constituent measure, with an interval estimate, using cross-validation within data sets and prediction across data sets and laboratories.

Hypothesis 60.9 (Relations add information)**.** Across a preregistered battery of data sets spanning wakefulness, sleep, and anesthesia, at least one relational measure in Table 60.1 adds out-of-sample information about reported experience beyond its matched constituent measure, and the addition replicates across laboratories.

Relational measure	Matched constituent measure	Relation-destroying surrogate	Known confound
Phase locking between regions (Lachaux et al. 1999; Mormann et al. 2000)	Band power in each region	Independent phase randomization; trial shuffling	Volume conduction (Nolte et al. 2004); small-sample bias (Vinck, Wingerden, et al. 2010)
Phase-amplitude coupling (Canolty, Edwards, et al. 2006; Tort et al. 2010)	Power at both frequencies; waveform-shape features	Time-shifted amplitude series	Nonsinusoidal waveforms (Cole and Voytek 2017; Aru et al. 2015)
Intermodulation (M. Regan and D. Regan 1988; Boremanse et al. 2013)	Fundamental amplitudes and their product	Misaligned configuration; swapped tags	Generic static nonlinearity
Bicoherence (Sigl and Chamoun 1994)	Power-spectral analog (Miller et al. 2004)	Phase-randomized surrogate	Amplitude dependence of the bispectrum
Perturbational complexity (Casali et al. 2013)	Per-channel evoked amplitude, duration, extent	Independent latency shifts per channel	Number of responsive channels
Entrainment tongue width (Notbohm et al. 2016)	Intrinsic peak frequency and amplitude; aperiodic exponent	Jittered stimulation	Intrinsic amplitude (Adler null)

Table 60.1: Relational measures, their matched constituent measures, relation-destroying surrogates, and the confounds that most often produce false relational findings.

This hypothesis is the programme's central empirical bet. If it fails for every measure in the table, the first falsification condition of Chapter 62 is met.

60.7 What a Positive Result Would Show

Suppose some relational measure beat every null in the table. That would establish a regularity of the kind Chapter 56 asks a new field to supply: relations among neural signals would carry information about conscious state or content that the signals taken one at a time do not. It would also give the reading loop an empirical foothold in consciousness science, because a relation that predicts report beyond its constituents is a candidate for a relation that some neural process reads.

It would not show that the relation constitutes experience. A better correlate is still a correlate, and Chapter 43 explains why no correlate, however refined, closes the explanatory gap by itself. Nor would it favor one theory of

consciousness over another until the theories have committed, in advance, to different predictions about the measure, which is the subject of Chapter 61.

60.8 Audit

Canonical phenomenon

Changes of state (wakefulness, dreaming and dreamless sleep, propofol, xenon, and ketamine anesthesia) that alter many neural measures at once, and a clinical index whose relational ingredient adds little over the power spectrum (Miller et al. 2004).

Minimal model

Matched constituent measures and relation-destroying surrogates (Definitions 60.1 and 60.2); the Adler equation for entrainment tongues (60.1); a static quadratic nonlinearity as the null for intermodulation.

What is relational here?

Detuning and phase locking between drive and rhythm; intermodulation between tagged inputs; the relative timing of evoked responses across cortex; the contingency between movement and sensation.

What physically reads it?

Each measure is computed by the experimenter, and that is the limit of what it shows. Intermodulation is the exception that points inward: it requires a neural nonlinearity that receives both inputs, which is a reader in the physical sense of Chapter 15.

What existing field explains it?

Frequency tagging and steady-state methods (M. Regan and D. Regan 1988; Norcia et al. 2015); entrainment research (Thut et al. 2011; Notbohm et al. 2016); perturbational complexity (Massimini et al. 2005; Casali et al. 2013); the neurophysiology of anesthesia (Purdon et al. 2013; Sarasso et al. 2015); surrogate-data methods (Theiler et al. 1992; Lancaster et al. 2018); spectral parameterization (Donoghue et al. 2020); sensory attenuation (Blakemore et al. 1998).

What remains unexplained?

Whether any relational measure beats its matched constituent null for reported experience. It has not been shown, and the one clinical relational component that was tested in this way did not beat it.

A discriminating experiment

The preregistered battery of Section 60.6, with the ketamine and propofol contrast as its sharpest comparison.

60.9 Notes and References

The comparison of bispectral and power-spectral parameters is Miller et al. (2004); the history and construction of the bispectral index are described by Sigl and Chamoun (1994) and Rampil (1998), and the two large trials of index-guided anesthesia are Myles et al. (2004) and Avidan et al. (2008). The method of surrogate data begins with Theiler et al. (1992) and is reviewed by Lancaster et al. (2018). Spectral parameterization and the aperiodic component are introduced by Donoghue et al. (2020); pitfalls of cross-frequency coupling are discussed by Aru et al. (2015) and Cole and Voytek (2017), and of connectivity measures by Nolte et al. (2004) and Vinck, Wingerden, et al. (2010). On entrainment of human rhythms, see Thut et al. (2011) and Notbohm et al. (2016); the network-level modeling is in Ali et al. (2013) and Herrmann et al. (2016). Frequency-domain analysis of nonlinear biological systems goes back to M. Regan and D. Regan (1988); Norcia et al. (2015) review steady-state methods, and Boremanse et al. (2013) is the configural intermodulation study used here. Perturbational complexity is developed in Massimini et al. (2005), Casali et al. (2013), and Sarasso et al. (2015); the EEG signatures of propofol are in Purdon et al. (2013) and Mukamel et al. (2014). The sensorimotor account of perceptual experience is O'Regan and Noë (2001).

60.10 Exercises

Exercise 60.1. Start from the forced normal form $\dot z = (\mu + \mathrm{i}\omega_0)z - |z|^2 z + \varepsilon\, \mathrm{e}^{\mathrm{i}\Omega t}$ with $\mu > 0$ and small ε. Write $z = r\,\mathrm{e}^{\mathrm{i}\theta}$, assume r stays near $r_0 = \sqrt{\mu}$, and derive (60.1) for $\phi = \theta - \Omega t$. Show that $\varepsilon_{\mathrm{eff}} = \varepsilon / r_0$ and sketch the tongue for two values of r_0.

Exercise 60.2. Pass $x(t) = a_1 \cos(2\pi f_1 t) + a_2 \cos(2\pi f_2 t)$ through $y = x + \beta x^2$. List every frequency in y with its amplitude, and show that the components at $f_1 \pm f_2$ have amplitude $\beta a_1 a_2$. Write the regression that implements the constituent null for Hypothesis 60.6.

Exercise 60.3. Explain why Miller and colleagues computed an amplitude-independent bicoherence version of the bispectral parameter. What does its failure to change over induction imply about the source of the change in the bispectral parameter, and which relational claim would a change in it have supported?

Exercise 60.4. Design the latency-shift surrogate of Hypothesis 60.7 in detail. Which properties of each channel's evoked response does it preserve, which relations does it destroy, and how should the range of shifts be chosen so that the surrogate responses remain physiologically plausible?

Exercise 60.5. Two scalp channels record $x_1 = w_1 s(t) + n_1(t)$ and $x_2 = w_2 s(t) + n_2(t)$ from a single source s with independent noise. Show that their

coherency is real at every frequency and that its imaginary part vanishes. Why does this make the imaginary part of coherency a useful, if conservative, measure of interaction (Nolte et al. 2004)?

Exercise 60.6. For the three conditions of Hypothesis 60.8, write the predicted ordering of detection rates under (a) attenuation by motor prediction alone, (b) the sensorimotor-contingency account alone, and (c) both. Which pairwise comparison discriminates the accounts, and what must be matched across conditions for it to do so?

Exercise 60.7. Write a one-page preregistration for Hypothesis 60.5: primary outcome, secondary outcomes, exclusion criteria, sample-size rationale, the exact constituent null, and the result that would count against the hypothesis.

61

Adversarial Testing

When consciousness fades under general anesthesia, nearly everything that theories of consciousness measure changes with it. The complexity of the cortical response to perturbation falls (Casali et al. 2013), occipital and frontal alpha rhythms reorganize (Purdon et al. 2013), and the coupling between slow and fast rhythms changes its pattern (Mukamel et al. 2014). Each major theory predicts that its preferred quantity will change, and it does. An observation that every theory predicts gives none of them an advantage over the others.

The thesis of this chapter is that theories of consciousness, including the relational hypothesis this book develops, are tested against one another only when their proponents commit in advance to predictions that diverge, to the data that will be collected, and to what each outcome will mean. No theory earns credit because its preferred quantity changes with consciousness. The practices that enforce this (preregistration, adversarial collaboration, theory-neutral laboratories) exist, and a large contest between two theories of consciousness has been run under them (Melloni et al. 2023; Cogitate Consortium et al. 2025). Its result, which challenged both theories, is the best argument for running more.

61.1 Covariation Does Not Discriminate

The reason that state-change results rarely decide between theories is elementary, and stating it exactly shows what a discriminating experiment must do.

Theorem 61.1 (Evidence that every hypothesis predicts)**.** *Let H_1 and H_2 be hypotheses with prior probabilities $\mathbb{P}(H_1), \mathbb{P}(H_2) > 0$, and let E be an observation with $\mathbb{P}(E) > 0$. Then*

$$\frac{\mathbb{P}(H_1 \mid E)}{\mathbb{P}(H_2 \mid E)} = \frac{\mathbb{P}(E \mid H_1)}{\mathbb{P}(E \mid H_2)} \cdot \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_2)}, \tag{61.1}$$

and in particular, if $\mathbb{P}(E \mid H_1) = \mathbb{P}(E \mid H_2)$, observing E leaves the odds between H_1 and H_2 unchanged.

Proof. Apply Bayes's rule to each hypothesis and divide; the factor $\mathbb{P}(E)$ cancels. □

The likelihood ratio in (61.1) is the only channel through which data move belief between hypotheses. Take E to be "measure X differs between wakefulness and propofol anesthesia." A theory whose preferred quantity is X assigns E a high probability. So, in most cases, does every rival theory, because anesthesia changes many quantities at once and the rivals do not deny that X will change; they deny only that X is what matters. The ratio is then close to one, and the result, however robust, is close to silent between the theories.

Example 61.2 (A confirmation that confirms little). Suppose one theory predicts that a measure of cortical integration will fall under propofol with probability 0.9, and a rival that locates consciousness elsewhere expects the same fall with probability 0.8, because propofol suppresses cortical activity broadly. Observing the fall multiplies the odds between them by $0.9/0.8 \approx 1.1$. An experiment in which the first theory predicts a fall with probability 0.9 and the second predicts no change with probability 0.9 would multiply the odds by $0.9/0.1 = 9$ in one direction or the other. The two experiments may cost the same; only the second is a test between the theories.

Discriminating evidence comes from conditions in which the theories' preferred quantities come apart: stimulus manipulations that separate onset from duration, conditions that separate report from experience, drugs that separate responsiveness from reported experience (Chapter 60), and interventions that target one theory's substrate while sparing another's. Designing such conditions requires that the theories say in advance what they predict in each.

61.2 Preregistration

Nosek and colleagues put the distinction on which the rest of this chapter depends plainly: science generates hypotheses from existing observations and tests them against new ones, and treating a postdiction as a prediction reduces the credibility of the result (Nosek et al. 2018). Hindsight bias makes the mistake easy to commit and hard to see. The remedy is to define the research questions and the analysis plan before observing the outcomes, which is what preregistration means; it distinguishes analyses that test predictions from those that generate postdictions, and it can be adapted to cases short of the ideal, such as analyses of data that already exist.

Consciousness science is unusually exposed to the mistake, because its measures are many and its theories flexible. A laboratory that compares wakefulness and sleep on twenty candidate measures, reports the three that differed most, and then explains why those three were the ones its theory cared about has produced a postdiction dressed as a test. Preregistration closes

that path by fixing the measure, the comparison, and the criterion for success before the data exist.

Principle 61.3 (No credit for covariation)**.** A theory of consciousness gains support from an experiment only to the extent that the experiment's outcome was more probable under that theory than under its rivals, as judged by predictions each theory's proponents registered before the data were collected. That the theory's preferred quantity changed with the state of consciousness is not, by itself, such an outcome.

61.3 Adversarial Collaboration

Preregistration fixes one theory's predictions. Adversarial collaboration fixes the predictions of several rival theories in one design, with their proponents' agreement. Mellers, Hertwig, and Kahneman described the method in the course of using it: the disputing parties agree on the empirical tests that would resolve their dispute and conduct them together with the help of an arbiter (Mellers et al. 2001). In their case Hertwig held that frequency formats eliminate the conjunction effects previously reported by Kahneman and Tversky, and that some of those effects arose because participants read "and" as a union; Kahneman disagreed with some of Hertwig's predictions and accepted others; Mellers served as arbiter. Frequency formats by themselves did not eliminate the conjunction effects, while with filler items removed the effects disappeared for the phrasings Hertwig proposed. Each party was right about something and wrong about something.

61.4 The Global Workspace and Integrated Information Contest

An adversarial collaboration between theories of consciousness pitted the global neuronal workspace theory against integrated information theory, two theories that, as the consortium noted, had accrued evidence independently but had not been directly compared (Cogitate Consortium et al. 2025). Their proponents, together with a theory-neutral consortium, devised and preregistered two experiments that tested contrasting predictions about the location and timing of the correlates of visual consciousness, predictions the theories' proponents endorsed (Melloni et al. 2023). The protocol specified in advance which outcomes would support, refute, or challenge each theory. Six theory-impartial laboratories were to carry it out with three complementary methods, functional magnetic resonance imaging, magneto- and electroencephalography, and intracranial recordings, with replications built in between laboratories and within data sets.

The results were published in 2025 (Cogitate Consortium et al. 2025). Participants ($n = 256$) viewed suprathreshold stimuli for variable durations while their brain activity was measured with the three methods. Information about conscious content was found in visual, ventrotemporal, and inferior frontal cortex; sustained responses in occipital and lateral temporal cortex reflected stimulus duration; and content-specific synchronization appeared between frontal and early visual areas. The results agreed with some predictions of each theory and substantially challenged key tenets of both. For integrated information theory, the lack of sustained synchronization within posterior cortex contradicts the claim that network connectivity specifies consciousness. For the global workspace theory, the general lack of ignition at stimulus offset and the limited representation of certain conscious dimensions in prefrontal cortex are the challenges. The authors note that the challenges extend to other theories that share the tested predictions, and they call for a quantitative framework for systematic theory testing.

Three lessons carry over to any contest, including one that involves the relational hypothesis of this book. First, the discriminating manipulation was simple: varying stimulus duration separated theories that tie consciousness to transient events at onset and offset from theories that tie it to activity sustained for as long as the experience lasts. Second, a contest can challenge every contestant at once, and that outcome is informative rather than a failure of design. Third, a challenge does not end a theory. Proponents can revise auxiliary assumptions, and whether such revisions make a research programme progressive or degenerating is the subject of Chapter 62.

61.5 Four Hypotheses and What Separates Them

We now write the contest the book's own hypothesis would have to enter. The statements of the three established theories are brief summaries of their core claims as their proponents state them; in a real contest, as in the protocol above, each theory's row of predictions must be written by its own proponents.

Hypothesis 61.4 (Global neuronal workspace)**.** H_{GNW}: a content is consciously accessed when it enters a fronto-parietal workspace through a nonlinear ignition that makes it globally available to many specialized processes (Dehaene and Changeux 2011; Mashour et al. 2020).

Hypothesis 61.5 (Integrated information)**.** H_{IIT}: consciousness is identical to a maximally integrated cause–effect structure specified by a physical substrate, which its proponents locate chiefly in posterior cortex (Tononi, Boly, et al. 2016; Albantakis et al. 2023; Koch, Massimini, et al. 2016).

Hypothesis 61.6 (Higher-order representation)**.** H_{HOT}: a mental state is conscious when it is the object of a suitable higher-order representation, with

prefrontal cortex implicated in generating it (Lau and D. Rosenthal 2011; R. Brown et al. 2019).

Hypothesis 61.7 (Relational reading)**.** H_{M}: a conscious content corresponds to a relational variable read by an endogenous reader within an agent-centered reading loop. Its testable commitments are three. (M1) Relational measures of the signals carrying a content hold information about its report beyond matched constituent measures (Chapter 60). (M2) That information depends on loop status: it differs between input contingent on the agent's action and the same input replayed (Chapter 59). (M3) Intervening on the relation, for example shifting the phase relation between two regions by closed-loop stimulation while holding each region's power fixed, changes report.

H_{M} is this book's candidate, stated so that it can fail, and it is not an established theory. Its commitments restate for consciousness two candidates that Chapter 56 proposed for Metaxics as a whole: (M1) is the hypothesis of relational measures beyond their relata (Hypothesis 56.5), and (M2) applies the hypothesis of reader relativity in closed loops (Hypothesis 56.4). Table 61.1 places the four hypotheses against six manipulations.

The table sorts manipulations by what they can decide. The state contrast in the first row is predicted by all four hypotheses, and by Theorem 61.1 it moves no odds among them. The duration and prefrontal rows are where the established theories disagree with one another, and the first of them is the row the published contest ran. The last three rows are where H_{M} exposes itself. Because the other theories make no stated commitment there, a confirmation of H_{M}'s predictions would support its distinctive claims without refuting its rivals, and a failure would refute H_{M} directly. That asymmetry is appropriate for a newcomer, and it is the reason the book states H_{M} at all: a hypothesis that entered no row where it could lose would earn nothing from the table.

61.6 Risky Commitments and Noncommittal Rivals

A cell marked "no stated commitment" is easy to misread as a prediction of no effect. It is a refusal to predict, and in the terms of Theorem 61.1 it means that the theory spreads its probability across the possible outcomes of that row. The consequence for a committed newcomer can be computed.

Proposition 61.8 (The price of a risky prediction)**.** *Let H_{M} assign probability p to an outcome O of a binary experiment, and let a rival that makes no commitment assign it probability q, with $p > q$. Observing O multiplies the odds in favor of H_{M} by p/q, and observing its absence multiplies them by $(1-p)/(1-q)$. For $p = 0.9$ and $q = 0.5$, the factors are 1.8 and 0.2.*

The arithmetic follows directly from (61.1). The committed hypothesis stands to lose more than it stands to gain, which is the ordinary price of a

Manipulation	H_{GNW}	H_{IIT}	H_{HOT}	H_{M}
Wakefulness versus propofol	Ignition reduced	Integration reduced	Higher-order signal reduced	Relational signature reduced
Stimulus duration varied	Ignition at onset and offset; content in prefrontal cortex	Posterior representation and synchronization sustained for the duration	No stated commitment	Relational signature sustained for the duration
Prefrontal perturbation, posterior cortex intact	Access degraded	Posterior contents spared	Awareness degraded	No commitment unless the reading loop runs through prefrontal cortex
Relational versus constituent information about content	No stated commitment	Compatible with integration measures	No stated commitment	Relational measure adds information (M1)
Action-contingent versus replayed input	Differs only if access differs	Differs only if cause–effect structure differs	No stated commitment	Differs (M2)
Inter-areal phase shifted at fixed local power	No stated commitment	No stated commitment	No stated commitment	Report changes (M3)

Table 61.1: A schematic contest. Entries for the three established theories summarize their core claims in the cited sources and would have to be replaced by predictions registered by each theory's proponents. The first row discriminates nothing; the second and third discriminate among the established theories; the last three are where H_{M} makes commitments that the others do not.

bold conjecture (Popper 1959). A theory that stays noncommittal can never be refuted by the row, and it can never gain much from it either; if it remains noncommittal across every row of every contest, it has chosen to make no empirical claims about the matters those rows concern.

Two practical rules follow. A newcomer gains the most from rows in which the established theories are committed and could be wrong, since there a single experiment can move the odds among several theories at once; the duration and prefrontal rows of Table 61.1 are such rows, and H_{M} should register predictions for them before it asks for credit anywhere else. And rows in which no theory is committed are exploratory. Their results should be reported as exploration and used to generate registered commitments for the next contest, not counted as tests of anyone.

61.7 Rules for a Fair Contest

The published protocol and its outcome suggest six rules, which apply equally to a contest that includes H_{M}.

1. *Divergence first.* Choose manipulations for which at least two contestants register opposite predictions; a row that every contestant predicts is a control, not a test.
2. *Registered interpretations.* Before data collection, specify for each outcome whether it supports, refutes, or challenges each theory (Melloni et al. 2023).
3. *Neutral execution.* Data are collected and analyzed by laboratories without a stake in the outcome, with several methods and built-in replication.
4. *Simpler nulls.* Every measure is compared with its matched constituent measure and its relation-destroying surrogate (Chapter 60).
5. *Open data.* The data and analysis code are released so that rival analyses can be run.
6. *Quantitative bookkeeping.* Outcomes are expressed as likelihood ratios or equivalent quantities, so that evidence accumulates across contests instead of being rhetorically reallocated after each one (Cogitate Consortium et al. 2025).

The electric fish cannot enter this contest, since it cannot report. It can enter a smaller one of the same form. The replay hypothesis of Chapter 59 sets the established neuroethological account, which predicts loop-invariant reading in wave-type fish, against the relational addition that the reader might depend on loop status without a corollary discharge. The rules above apply without change.

61.8 Audit

Canonical phenomenon

A preregistered adversarial collaboration between two theories of consciousness, run by theory-neutral laboratories with three methods, whose results challenged both (Melloni et al. 2023; Cogitate Consortium et al. 2025).

Minimal model

Posterior odds equal the likelihood ratio times the prior odds (Theorem 61.1); a prediction table with registered commitments per hypothesis (Table 61.1).

What is relational here?

The relation that matters is between theories and data: the likelihood ratio, which is a property of two hypotheses taken together with an observation and of none of them alone.

What physically reads it?
The contest itself: arbiters and neutral laboratories read the data against registered predictions. Within H_{M}, the endogenous reader whose existence rows four to six test.

What existing field explains it?
Bayesian confirmation theory and falsificationism (Popper 1959); preregistration (Nosek et al. 2018); adversarial collaboration (Mellers et al. 2001); the global workspace, integrated information, and higher-order theories (Dehaene and Changeux 2011; Tononi, Boly, et al. 2016; Lau and D. Rosenthal 2011); the comparative survey of theories in Chapter 44.

What remains unexplained?
The rows in which H_{M} makes commitments that no established theory makes. Their outcomes are unknown.

A discriminating experiment
Run the last three rows of Table 61.1 under the rules of Section 61.7, with proponents of the established theories invited to register their own predictions for those rows first.

61.9 Notes and References

The case for preregistration is made by Nosek et al. (2018). Adversarial collaboration is described, and exemplified, by Mellers et al. (2001). The protocol of the contest between the global neuronal workspace and integrated information theories is Melloni et al. (2023), and its results are Cogitate Consortium et al. (2025); both papers should be read in full before designing a similar study, because the details of the registered predictions matter more than any summary. The theories themselves are presented by their proponents in Dehaene and Changeux (2011) and Mashour et al. (2020) (global workspace), Tononi, Boly, et al. (2016) and Albantakis et al. (2023) (integrated information), and Lau and D. Rosenthal (2011) and R. Brown et al. (2019) (higher-order); Seth and Bayne (2022) survey the field, and Chapter 44 compares the theories by asking what additional property each claims. The argument that posterior cortex is the likelier substrate of experience is set out by Koch, Massimini, et al. (2016). The value of risky predictions, and the emptiness of predictions that cannot fail, is the theme of Popper (1959).

61.10 Exercises

Exercise 61.1. Two theories assign probabilities 0.7 and 0.6 to an observation, and their prior odds are $1 : 1$. Compute the posterior odds after one observation, and after ten independent observations of the same kind. How many observations of a kind with likelihoods 0.9 and 0.1 produce the same shift?

Exercise 61.2. Construct a concrete example in which two theories of consciousness both predict that a measure of their choice falls under propofol, and explain why a study reporting that fall supports neither against the other. Then modify the design so that the theories' predictions diverge.

Exercise 61.3. Using the published summary of the contest's results (Cogitate Consortium et al. 2025), list which of the registered predictions of each theory were supported and which were challenged. Propose one follow-up manipulation whose outcome each theory's proponents would, in your judgment, have to predict differently, and justify the judgment from the theories' own statements.

Exercise 61.4. Add a row to Table 61.1 for a no-report paradigm, in which stimuli are consciously seen but task-irrelevant. Fill in only those cells for which you can cite a proponent's stated commitment, and mark the rest. What does the pattern of empty cells tell you about the design of a contest?

Exercise 61.5. Design the intervention of commitment M3: a closed-loop stimulation protocol that shifts the phase relation between two cortical regions while holding each region's band power fixed. State how you would verify that local power was held fixed, what report you would collect, and which result would refute M3.

Exercise 61.6. Using Proposition 61.8, find the probability p that H_{M} must assign to an outcome for a confirmation to multiply its odds against a noncommittal rival ($q = 0.5$) by at least 3. What does a failure then cost it? Repeat for a rival that is committed against the outcome, with $q = 0.1$.

Exercise 61.7. Draft the arbitration rules for a contest between H_{M} and H_{GNW}: who registers which predictions, who collects and analyzes the data, what counts as support, refutation, and challenge for each, and how disagreements about the interpretation of an outcome are to be resolved.

62

What Would Falsify the Programme?

A research programme that cannot say how it could fail has made no claim that the world could contradict. This book has made several. Its central idea is that the step from physical relation to point of view runs through readers that participate in the relations they read, and its method has been to subtract, chapter by chapter, what existing science already explains. The subtraction leaves six claims, some modest and some ambitious, and each can be defeated by an experiment whose outcome is not yet known.

The thesis of this chapter is that each of the programme's six claims has a failure condition that can be stated now, as a concrete result that would count against it, and that stating the conditions in advance is what separates a research programme from a perspective that cannot lose. We write the conditions down, say what would survive each failure, name the moves by which the programme could avoid them dishonestly, and end with the scenario in which the central idea dies.

62.1 Falsification and Research Programmes

Popper made falsifiability the mark of an empirical theory: a hypothesis says something about the world to the extent that it forbids some possible observations, and it is tested by looking for the observations it forbids (Popper 1959). Lakatos observed that scientists do not abandon a programme at its first anomaly and that they are often right not to (Lakatos 1970). A research programme, on his account, has a hard core of claims its proponents treat as irrefutable by decision, surrounded by a protective belt of auxiliary hypotheses that absorb anomalies by being adjusted. The adjustments are what can be judged. A programme is progressive when its adjustments predict novel facts that are then corroborated, and degenerating when they only accommodate facts already known.

Both views apply here, at different levels. Individual hypotheses of the

programme, such as the replay hypothesis of Chapter 59 or the intermodulation hypothesis of Chapter 60, are Popperian: each forbids specific outcomes. The programme as a whole is Lakatosian: it will survive some failed hypotheses, and whether it deserves to depends on what its adjustments predict. What neither view permits is a programme whose proponents decide only after the data arrive which results would have counted against it.

Definition 62.1 (Failure condition)**.** For a claim C of the programme, a *failure condition* is a result R, specified together with the procedure that would produce it before the data are collected, such that obtaining R requires abandoning C or restricting its scope in a way also specified in advance.

Definition 62.2 (Hard core and belt of this programme)**.** The *hard core* of the programme consists of three claims: that the reading loop $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$ is the right unit of analysis for systems that participate in the relations they read; that some relational variables become dynamically explicit only through physical readers; and that the reading of relations within agent-centered loops is part of the explanation of perspective. The *belt* consists of everything the book has proposed to test those claims: specific measures (Chapter 60), specific models (such as the toy jamming loop of Chapter 59), the commitments M1 to M3 of the relational hypothesis (Chapter 61), and specific experiments.

The six claims below translate the hard core into claims that can fail. The first four concern relations and readers in any system; the fifth concerns perspective and experience; the sixth concerns whether the programme adds anything to existing science.

62.2 Six Claims and How Each Could Fail

62.2.1 F1: Relational measures outperform constituent measures

Claim. In at least one domain, relational measures of a system's signals predict its state, behavior, or reports better than matched constituent measures computed from the same data. The primary domain proposed in this book is the neural basis of reported experience; a secondary domain is the behavior of active-sensing animals.

Test. The preregistered battery of Chapter 60, in which each relational measure is compared with its relation-destroying surrogate and with its matched constituent measure, using out-of-sample prediction across data sets and laboratories (Theiler et al. 1992; Lancaster et al. 2018).

Result that would count. For every measure in the battery, the incremental predictive value over the matched constituent measure has an interval estimate that excludes every effect larger than a small margin fixed in advance, in adequately powered and replicated data sets. A precedent exists: when the

bispectral parameter of an anesthesia monitor was compared with its power-spectral analog, the relational parameter added nothing that could be shown (Miller et al. 2004).

What would survive. Relational measures as compact descriptions, and the whole physics of Parts I to III, which never depended on this claim.

62.2.2 F2: Some relational variables do not reduce to known ones

Claim. Some variables that the book treats as relational and reader-dependent, such as the sign of a frequency difference carried by the quadrature relation between two modulations (Chapter 21), do explanatory work that the established variables of the relevant fields (order parameters, phase differences, mutual information, Volterra kernels, the states of a control loop) do not already do.

Test. A translation audit: for each variable the book claims as distinctive, a specialist in the nearest existing field attempts to express it, and every prediction that involves it, in that field's vocabulary, including predictions about interventions.

Result that would count. Every such variable is translated without loss of any prediction. This is the failure condition the book most expects to be met; the audits at the end of the chapters have already met it for most of the book's material, as Chapter 55 records.

What would survive. The organization of known results around the reading loop, which serves teaching and shows connections while adding no new facts.

62.2.3 F3: Analyses transfer across substrates

Claim. A regularity found in one reading loop, stated in terms of its organization, predicts a regularity in reading loops built from unrelated substrates. The example the book offers is the regulation of a relation. When the gain of reafferent feedback was changed four-fold, electric fish changed their active-sensing movements so as to hold constant the sensory slip they sensed, which identifies slip, a relation, as the regulated variable (Biswas et al. 2018). The transfer claim is that other systems that sense actively will, under the same manipulation, regulate a relation rather than their own movements.

Test. Preregistered gain manipulations in several unrelated active-sensing systems, with the regulated variable identified by the same dissociation between movement and slip.

Result that would count. The predicted regulation fails in most systems tested, or holds only where it follows from shared ancestry or shared mechanics rather than from the organization of the loop.

Example 62.3 (A transfer that tests nothing). Some regularities hold in every reading loop because mathematics requires them. A single real modulation cannot carry the sign of the frequency difference that produced it, since

its spectrum is symmetric (Chapter 1); every system that reads that sign must therefore combine at least two modulations in quadrature, either the amplitude and the phase of one signal measured against a reference, or, with no reference available, amplitude and differential phase across two differently mixed channels, as the fish does with regions of its body. Such a regularity will transfer everywhere, and its transfer tests nothing empirical. Only regularities of organization that could have been otherwise count for or against F3.

What would survive. The fish's neuroethology, and the reading loop as a local description of particular systems.

62.2.4 F4: Modeling the reader adds explanatory power

Claim. Specifying the reader, which relation is read and by which physical operation, predicts the effects of interventions that models of the input alone cannot predict.

Test. Reader interventions that change D while holding the input fixed, by lesion, pharmacology, adaptation, or learning (Chapter 58), and the loop-replay experiments of Chapter 59.

Result that would count. In every intervention of the preregistered set, the behavioral consequences are predicted as well by a model that treats the intervention as a generic change of gain or noise at the input as by a model that specifies the relation read. The replay experiment contributes to this condition directly: if reading is loop-invariant wherever no corollary discharge exists, and loop-dependent only where one does, the reader adds nothing beyond known circuitry.

What would survive. The reading loop as a control-theoretic diagram, with readers as components whose internal organization does not matter to the behavior of the whole.

62.2.5 F5: Agent-centered reading bears on phenomenality

Claim. The agent-centered measurement frame, variables of the form "world relative to my action" (Chapters 22 and 34), is part of the explanation of why experience has a point of view. This is the most speculative claim of the programme.

A restriction must be stated before any new data arrive, because a known fact already bears on the claim. Experience does not require an ongoing loop with the external world: after ketamine anesthesia, participants reported long, vivid dreams unrelated to the environment, during a state of profound unresponsiveness (Sarasso et al. 2015), and dreaming sleep is a familiar state of the same kind. The claim can therefore concern only agent-centered variables that are internally modeled (Chapter 31), and not the external sensorimotor loop. Stated after a failure, this restriction would be an ad hoc rescue; stated now, it is part of the claim.

Result that would count. Either of two outcomes. First, manipulations that abolish the difference between action-contingent and replayed input, and every other available marker of agent-centered processing, leave reported experience unchanged across all tested conditions. Second, agent-centered processing is found intact, by the same markers, in states that lack experience by every accepted criterion, such as propofol anesthesia from which participants report nothing.

What would survive. Part VIII as a theory of here, now, and self as coordinates, with no claim about experience.

62.2.6 F6: The programme predicts something existing frameworks do not

Claim. At least one prediction, experimental operation, or regularity of the programme is not expressible within, or was not derivable from, existing frameworks: predictive processing and active inference, integrated information theory, global workspace theory, control theory, synergetics, and information theory. This is the claim that would make Metaxics a field (Chapter 56).

Result that would count. For each confirmed prediction of the programme, a proponent of an existing framework shows, from sources dated before the prediction was registered, that the framework already made it or implies it by a short derivation.

What would survive. Metaxics as a synthetic perspective connecting existing sciences, which is what the Preface and Chapter 56 say it currently is. This failure is compatible with the book being useful.

62.3 How the Programme Could Cheat

Lakatos's distinction between progressive and degenerating programmes names the danger precisely: a programme can survive every failure by adjusting its belt, and the adjustments reveal whether it is still learning anything (Lakatos 1970). This programme has five tempting adjustments, and each is a sign of degeneration.

1. *Redefining "relational" after the fact*, so that whichever measure predicted best is declared to have been the relational one.
2. *Positing readers on demand*: explaining any result by a reader of the relevant relation, without independent evidence of the physical operation that does the reading.
3. *Choosing measures after the data*, the forking-path problem that preregistration exists to prevent (Nosek et al. 2018).
4. *Retreating to ubiquity*: answering failures with the observation that relations are everywhere, the move the Preface rejected at the outset.
5. *Perpetual challenge*: reclassifying every refutation as a "challenge" that calls for further work, without any result ever counting as a loss.

Claim	Test	Result that would count	What survives
F1 Relational beats constituent	Preregistered battery with surrogates and matched measures	No incremental predictive value beyond a preset margin, replicated	Relational descriptions as compressions
F2 Irreducible relational variables	Translation audit by specialists	Every variable translated without loss	The book as an organization of known results
F3 Transfer across substrates	Gain manipulations in unrelated active-sensing systems	Predicted regulation fails, or holds only by ancestry or mechanics	Local descriptions of particular systems
F4 Readers add explanatory power	Reader interventions; loop replay	Input-level models predict every intervention equally well	The loop as a control diagram
F5 Agent-centered reading bears on phenomenality	Contingency manipulations; states without experience	Report unaffected by agent-centered markers, or markers intact without experience	A theory of here, now, and self
F6 Predictions beyond existing frameworks	Priority check against earlier sources	Every confirmed prediction already implied	Metaxics as a synthetic perspective

Table 62.1: The programme's six claims and their failure conditions.

Principle 62.4 (Advance commitment)**.** If the failure conditions of F1 and F4 are both met in preregistered and replicated tests, the programme is to be judged degenerating as a scientific programme, whatever adjustments of its belt are proposed afterward.

The principle picks F1 and F4 because they are the cheapest to test and the most decisive. F1 asks whether relations carry information that constituents do not; F4 asks whether the physical reader of a relation matters. A programme about relations and readers that fails both has nothing left to be about.

62.4 How the Central Idea Could Die

Suppose the tests are run and go badly. The preregistered battery finds no relational measure that beats its matched constituent measure. Replay experiments show loop-invariant reading wherever no corollary discharge exists and loop-dependent reading only where one does, exactly as neuroethology already predicts. Reader interventions are predicted as well by models of the input. The gain manipulations transfer only where shared mechanics explain them. And every confirmed regularity turns out to have been implied by

control theory, predictive processing, or one of the established theories of consciousness.

Then the central idea dies as a scientific claim. The fish remains explained by neuroethology and control theory, as it already is. Consciousness science continues with its theories and its adversarial contests, none of which needs the notion of a reader. What remains of this book is a map of how physics, neuroethology, theoretical biology, and philosophy of mind bear on one question, which is what Part XII concludes Metaxics currently is. The question itself would survive the death of the programme, stated more sharply than it was at the start: what physical organization makes there be something it is like for relations to appear from here (Chapter 64)?

That death would be an acceptable outcome. Writing the failure conditions down is the programme's first result. If the programme survives them, it will have earned its name. If it does not, the reader will know exactly why.

62.5 Audit

Canonical phenomenon

The programme itself: six claims derived from the book's hard core, and the experiments that could defeat them.

Minimal model

Failure conditions (Definition 62.1); the hard core and belt of the programme (Definition 62.2); likelihood reasoning from Chapter 61.

What is relational here?

The relation between a claim and a registered outcome. Each of the six claims is itself relational: between relational and constituent measures, between variables and their translations, between loops in different substrates, between readers and inputs, between agent-centered processing and report.

What physically reads it?

The community of investigators who read results against conditions registered in advance. The chapter's rule is that this reader must be fixed before the data exist.

What existing field explains it?

Falsificationism (Popper 1959); the methodology of research programmes (Lakatos 1970); preregistration (Nosek et al. 2018); surrogate-data testing (Theiler et al. 1992; Lancaster et al. 2018).

What remains unexplained?

None in the method. The residue is the unknown outcome of each test.

A discriminating experiment

F1 and F4 first: the preregistered battery of Chapter 60 and the replay experiment of Chapter 59, under the rules of Chapter 61.

62.6 Notes and References

Popper (1959) is the source of falsifiability as a demarcation criterion and of the preference for bold conjectures. Lakatos (1970) develops the methodology of scientific research programmes, including the hard core, the protective belt, and the distinction between progressive and degenerating problem shifts. The practical case for committing to analyses before seeing data is made by Nosek et al. (2018). The method of surrogate data, on which the first failure condition depends, is due to Theiler et al. (1992) and reviewed by Lancaster et al. (2018). The anesthesia-monitor precedent for F1 is Miller et al. (2004), the active-sensing experiment behind F3 is Biswas et al. (2018), and the reports of dreaming under ketamine that restrict F5 are in Sarasso et al. (2015).

62.7 Exercises

Exercise 62.1. Write a failure condition, in the form of Definition 62.1, for one claim of Part VIII. State the procedure, the result that would count, and the restriction of scope it would force.

Exercise 62.2. For each hypothetical result, say which of F1 to F6 it bears on and whether it supports, challenges, or meets the failure condition: (a) intermodulation predicts face reports beyond its fundamentals in one laboratory and not in two replications; (b) wave-type fish show identical midbrain responses in closed loop and replay; (c) a control theorist derives the fish's regulation of sensory slip from an optimal-control model published before the experiment; (d) perturbational complexity loses its discrimination on latency-shifted surrogates.

Exercise 62.3. Prove that a real signal $\cos(\Delta\omega\, t + \psi_0)$ observed alone cannot reveal the sign of $\Delta\omega$. Explain why this fact, although it holds in every reading loop, cannot count as a successful cross-domain transfer under F3.

Exercise 62.4. Perform a translation audit for the beat envelope of Chapter 3: express every prediction the book makes about it in the vocabulary of signal processing and nonlinear systems theory. Does anything fail to translate?

Exercise 62.5. Identify the hard core and two elements of the protective belt of one established theory of consciousness. Using the published results of the adversarial contest discussed in Chapter 61, describe one adjustment of the belt that would count as progressive and one that would count as degenerating.

Exercise 62.6. F1 requires a margin below which incremental predictive value counts as absent. Choose a margin for an out-of-sample classification of

wakefulness against propofol anesthesia, justify it, and describe how you would estimate the number of participants needed to exclude effects above it with high confidence.

Intervene on the relation while holding the relata fixed; intervene on the reader while holding the input fixed; close and open the loop. Every experiment carries a simpler null model, and every relational measure must beat the constituent measures it is built from. The programme can die, and the conditions of its death are written down. Part XIV returns to the question the book began with.

Part XIV

The Remaining Question

Which arrow have we actually explained?

63

From Reading to Seeing

The book began with a question about moiré patterns and consciousness, and replaced it with a sharper one: how does a physical system become the reader of relations generated partly by its own activity, and why is such a reader ever accompanied by a point of view? Sixty-two chapters later, the first half of that question has a detailed answer and the second half has a precise location.

This chapter lays the progression out end to end and asks, of each step, whether it has been explained. The thesis is that the progression consists of nine arrows: the first three are explained by established science, the next four are explained in outline by fields with competing detailed accounts, and the last two are not explained at all. The book's contribution is the grading itself: the ability to say which arrow a proposed theory of consciousness claims to cross, what evidence bears on each, and where the electric fish stands.

63.1 The Progression

Figure 63.1 collects the stages the book has climbed. Each stage adds a capability to the one below it.

The stages are not the only way to cut the territory, and several could be split or merged. They are the cut the book has used, each corresponding to a part in which the capability was defined and its best-understood instances examined. The question for each arrow is the same: given a system at the lower stage, what additional organization produces the upper stage, and has anyone shown that it does?

63.2 Grading the Arrows

Arrow 1: relation to physical readout. A relation implicit in a mixture becomes an explicit variable through a physical operation: square-law detection, rectification, multiplication, thresholding, phase comparison (Chapters 3 and 4). Which relational variables become explicit depends jointly on the mixture and on the reader (Chapters 15 and 16). Signal theory and physics explain

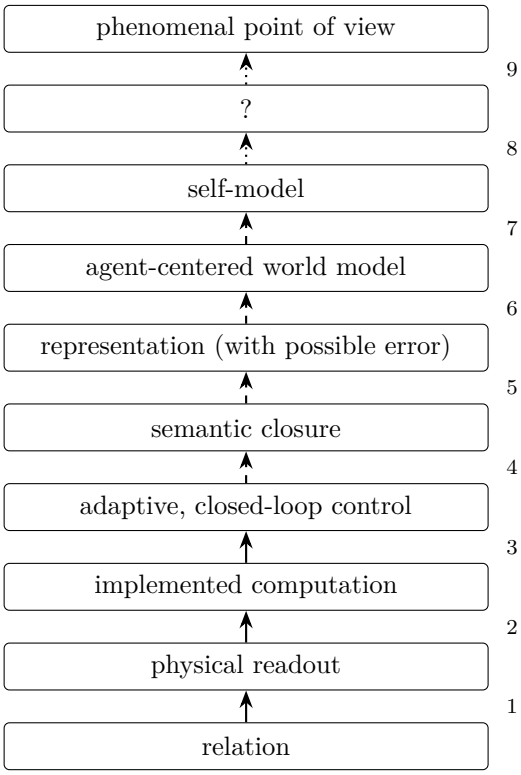


Figure 63.1: The staircase. Solid arrows are explained by established science; dashed arrows are explained in outline, with competing detailed theories; dotted arrows are unexplained. The grading is argued in Section 63.2.

this completely. In the fish, tuberous receptors are the readers: P-units make amplitude explicit and T-units make timing explicit (Chapters 19 and 21). *Status: explained.*

Arrow 2: readout to implemented computation. A readout becomes part of a computation when the system's causal structure supports the counterfactuals the computation requires, which is what blocks the triviality arguments of Putnam and Searle (Chapter 17). The philosophical debate about the right criteria continues, and it is a debate about criteria rather than about any mechanism. In the fish, the counterfactuals are satisfied and experimentally verified: reverse the sign of Δf and the neural and behavioral trajectories reverse (Chapter 21). The algorithm is not observer-imposed; two independent lineages share it (Chapter 24). *Status: explained.*

Arrow 3: computation to adaptive, closed-loop control. A computation becomes control when its output acts on the system so as to change its future input. Control theory explains this; the fish's jamming avoidance loop has been opened, closed, clamped, and modeled (Madhav et al. 2013) (Chapters 20 and 59). *Status: explained.*

Arrow 4: control to semantic closure. Semantic closure adds that the reader,

and the loop it belongs to, are built and maintained by the system whose persistence the loop serves (Chapters 25 to 27). Theoretical biology explains how this is possible and has formal frameworks for it (closure of constraints, organizational accounts of function), though the frameworks compete and their formal details are unsettled. The fish grows its own emitter, receptors, decoder, and effectors, and maintains them as part of maintaining itself; the radio's reader exists because an engineer built it (Chapters 48 and 49). *Status: explained in outline; competing formal accounts.*

Arrow 5: semantic closure to representation. Representation adds correctness conditions: a reading that can be wrong in a way that matters to the system. Phantom jamming makes error operational in the fish (Chapter 23), but what fixes the content of the fish's reading, and so whether the phantom response is misrepresentation, depends on which theory of content is right: teleosemantic, informational, structural, or enactive (Chapters 28 to 30). The disjunction problem is unresolved in general. *Status: explained in outline; the account of content is contested.*

Arrow 6: representation to agent-centered world model. This arrow bundles two capacities. Detachment lets a system carry relations that are not present in its current input: memory, prediction, emulation, counterfactual simulation (Chapters 31 to 33). Agent-centering gives the model's variables an origin at the system: here, now, and the indexical "I" (Chapters 34, 35, 37 and 38). Both are well studied in neuroscience and philosophy. The fish reaches agent-centering by a shortcut: its carrier-based measurement frame produces variables of the form "the world relative to my activity" without any detached model (Chapter 22). *Status: explained in outline.*

Arrow 7: agent-centered world model to self-model. A self-model represents the system itself as an object in its world: its body, its states, its agency, its own reliability (Chapters 36, 39, 41 and 42). Interoception, body ownership, metacognitive confidence, and attention schemas have neural and computational accounts in mammals; the accounts disagree about what a self-model must include. Nothing in the fish's jamming avoidance system models the fish as a source: the computation accepts a mimic as "self" (Chapter 22). *Status: explained in outline in mammals; not present in the fish's jamming system.*

Arrows 8 and 9: self-model to phenomenal point of view. Here the progression breaks. A functional perspective is a set of agent-centered variables that a system reads and uses, including variables about itself. A phenomenal perspective is there being something it is like to be that system. Every major theory of consciousness names a property that is supposed to take a system from the first to the second: global broadcast, integrated information, recurrent processing, higher-order representation, an attention schema, a transparent self-model, a projective geometry of experience, egocentric integration in the midbrain (Chapter 44). Illusionism denies that there is a further step to take, and holds that what needs explaining is why systems like us represent themselves as having one (Frankish 2016). None of these proposals has been shown

to entail phenomenality, and the adversarial experiments designed to test them discriminate between their predictions about neural correlates, not between their claims about entailment (Melloni et al. 2023; Cogitate Consortium et al. 2025) (Chapter 61). The question mark in Figure 63.1 marks the place where some such property would sit. *Status: unexplained.*

Arrow	From → to	Explained by	Status
1	relation → readout	signal theory, physics of transduction	explained
2	readout → computation	counterfactual and mechanistic accounts of implementation	explained
3	computation → control	control theory, neuroethology	explained
4	control → closure	theoretical biology (autopoiesis, closure of constraints)	in outline
5	closure → representation	theories of content (teleosemantics and rivals)	in outline, contested
6	representation → agent-centered model	neuroscience of internal models and spatial frames; logic of indexicals	in outline
7	agent-centered model → self-model	interoception, metacognition, self-model theories	in outline
8–9	self-model → phenomenal view	theories of consciousness (none demonstrated)	unexplained

Table 63.1: The grading of Section 63.2.

63.3 Where the Fish Stands

The electric fish is the smallest system in which the lower staircase has been mapped in one animal. Arrows 1 to 3 are established in its jamming avoidance response with experimental detail at every stage. Arrow 4 holds for the fish as for any organism. Arrow 5 is where the fish first demands philosophy, because phantom jamming makes error real without settling content. Arrow 6 is reached by a shortcut, the carrier-based frame, and arrow 7 is absent from the jamming system.

That position gives the electric fish question of Chapter 50 its precise form. If there is something it is like to be *Eigenmannia*, which piece of its architecture matters? If there is nothing, what does it lack that some other animal has? Each theory of consciousness answers differently. A theory that locates phenomenality in agent-centered sensorimotor integration of the kind Merker (2007) describes may count the fish in; a theory that requires a global workspace or higher-order representation of the kind found in mammalian cortex may count it out; integrated information theory would compute a quantity

from the fish's causal structure. The fish converts the abstract question into a question about a known circuit.

63.4 What Would Count as Progress

An unexplained arrow is not an unapproachable one. The book's research programme (Part XIII) proposes four kinds of work that bear on the upper arrows without presupposing their answer.

Relational measures that beat constituent measures. If some relational property of neural activity, such as a frequency relation, a phase structure, or an intermodulation response, tracks the state of consciousness better than matched measures of the constituent signals, it earns a place in the theory; if not, it does not (Chapters 46 and 60).

Reader interventions. Changing the decoder while holding the input fixed tests whether a variable's role depends on how it is read, the question at the center of Part IV (Chapter 58).

Closed and open loops. Comparing a system's dynamics when it can and cannot act on the relation it reads tests whether agent-centered measurement changes anything beyond input statistics (Chapter 59).

Adversarial prediction. Theories that place arrow 9 in different places make different predictions about which manipulations abolish consciousness; preregistered tests can eliminate some of them (Chapter 61).

None of these would cross the last arrow by itself. Each would narrow the set of properties that could sit at the question mark.

63.5 The Moiré, Revisited

The original question can now be answered. Consciousness is not a moiré effect in any sense that explains it. Moiré-like relational structures appear at several stages of the staircase: in the beats the fish reads, in the frequency relations of neural activity, and in the interference models once proposed for grid cells and orientation maps (Chapter 47). At each stage the structure is real, and at each stage an existing science explains it. The moiré intuition was right that the variables that matter are often relations, and right that a relation can be real without being located in any component. It was wrong to suppose that relationality addresses arrow 9. The analogy's failure is instructive: it names a property that every stage of the staircase has, and so it cannot distinguish the stages.

63.6 Audit

Canonical phenomenon

The progression from a physical relation to a point of view, as instantiated

in part by the electric fish and in full, if at all, by animals like us.

Minimal model
The staircase of Figure 63.1 and the grading of Table 63.1.

What is relational here?
Every stage is relational in the sense of Chapter 1; the stages differ in what reads the relations and what the readings do.

What physically reads it?
At the lower stages, physical transducers and neural circuits; at the middle stages, the organism's self-maintaining loops and models; at the upper stages, whatever systems read their own readings.

What existing field explains it?
Every explained arrow belongs to an existing field, as mapped in Chapter 55.

What remains unexplained?
Arrows 8 and 9: why any system with a self-model, or any other organization, is accompanied by a phenomenal point of view.

A discriminating experiment
The experiments of Part XIII, each designed so that at least one proposed location of the question mark predicts an outcome that another forbids.

63.7 Notes and References

Nagel (1974) gave the phrase "something it is like" its modern use; Levine (1983) named the explanatory gap and Chalmers (1995) the hard problem. Seth and Bayne (2022) survey current theories and their commitments, and Melloni et al. (2023) and Cogitate Consortium et al. (2025) report the first large adversarial collaboration between two of them. Frankish (2016) and Dennett (2016) state the illusionist alternative. The comparative question of which animals are conscious is treated by Birch et al. (2020) and Feinberg and Mallatt (2016); Merker (2007) argues for a midbrain basis that would widen the class considerably.

63.8 Exercises

Exercise 63.1. Grade the nine arrows for a robot vacuum cleaner with bump sensors, an internal map, and a battery-seeking behavior. Which arrows does it cross, which does it cross only by courtesy of its designers, and which not at all?

Exercise 63.2. Grade the arrows for an octopus, using what you can find about its sensing, learning, and behavior. Where is the evidence thinnest?

Exercise 63.3. Choose one arrow that Figure 63.1 treats as a single step and argue that it should be split in two. State the intermediate stage and a system that has it.

Exercise 63.4. Choose one theory of consciousness from Chapter 44. State exactly which arrow it claims to explain, what property it places at the question mark, and one observation that would show the property is present without phenomenality.

Exercise 63.5. Write one paragraph defending the moiré analogy against Section 63.5, and one paragraph answering your defense.

64

The First View from Somewhere

Thomas Nagel opened *The View from Nowhere* by describing its single problem as how to combine the perspective of a particular person inside the world with an objective view of that same world (Nagel 1986). This book has approached the problem from the other end. It began with the objective view (relations between physical systems, described from outside) and asked how far one can climb toward a perspective using only physical organization. The climb was long, and much of it is explained; Chapter 63 graded it arrow by arrow.

This last chapter states what the climb leaves behind. The thesis is a single question, made as precise as the book can make it. Six distinctions lead up to it. Each separates two things that the moiré intuition of the Preface ran together, and each is illustrated by a system that has the first without the second. Here the distinctions are the argument, and the contrasts are stated as contrasts on purpose.

64.1 A Detector Is Not Yet a Reader

A photodiode behind a window produces a current that varies with the light. Its output depends systematically on its input, and in that sense it detects. If nothing uses the current, no variable has been made available to any further process; the relation between light and current is available only to an engineer with a meter.

A reader, in the sense of Chapter 15, is a physical operation that makes a variable of its input dynamically explicit, so that further dynamics can depend on it. The square-law detector that turns a beat into a difference-frequency signal and feeds it to a filter is a reader; the same detector with its output unconnected is only a detector. The difference lies outside the device, in whether its output enters a causal process that depends on it.

64.2 A Reader Is Not Necessarily an Agent

A lock-in amplifier reads the phase relation between a signal and a reference with exquisite precision, and its output can drive a display, a recorder, or a feedback loop. It does not act on itself to change what it reads. The electric fish's reader does: the sign of Δf drives a change in the fish's own frequency, and that change alters the next mixture on its skin (Chapters 20 and 21).

An agent, in the minimal sense this book has used, is a reader embedded in a loop that acts on the system's own state so as to change its future input. The lock-in in a phase-locked loop comes close; it acts on an oscillator that is part of the same apparatus. What it lacks, and what the fish has, is a loop whose parts the system itself builds and maintains (Chapter 26).

64.3 An Agent Is Not Necessarily a Representer

A thermostat acts on its world to keep a variable near a set point. A bacterium swimming up a chemical gradient compares the concentration now with the concentration a few seconds ago and changes its tumbling rate accordingly (Chapters 28 and 49). Both are agents in the minimal sense; the bacterium is also self-maintaining. Whether either represents anything is exactly what theories of content dispute (Chapter 29).

The fish's jamming reader is agent-like and self-maintained, and phantom jamming shows that it can respond to a situation that does not obtain (Chapter 23). That makes error operational. It does not settle representation, because what the reading is about, and so whether the phantom response is an error, depends on a standard that the circuit itself does not supply.

64.4 A Representer Is Not Necessarily Self-Aware

Grant that some animals represent their surroundings: they carry information about objects that are no longer present, predict where prey will be, and act on the predictions. Such a system represents the world without representing its own representing. A rat that wagers more on decisions it is likely to get right has something further: a variable that tracks the reliability of its own first-order states (Kepecs et al. 2008). Metacognition, higher-order representation, and self-models are the family of capacities that add this reflexive layer (Chapters 39 to 41).

The jamming reader has none of them. It does not represent itself as the source of its carrier, and it accepts an electronic mimic as its own discharge (Chapter 22). Its variables are centered on the fish by physics and anatomy, not by any representation of the fish.

64.5 A Self-Model Is Not Obviously a Subject

A robot can carry a model of its own body, update it from proprioceptive feedback, and use it to predict the consequences of its movements. It can carry a model of its own uncertainty and of where its attention is directed. Self-model theories hold that when such a model is transparent (when the system cannot represent it as a model), the system experiences itself as a self (Metzinger 2003); attention schema theory holds that the model of one's own attention is what the system reports as awareness (Graziano 2013).

These theories may be right. Their claims, however, are claims about arrow 9 of Chapter 63, and nothing in the functional description of a self-model shows that there is a subject for whom the model is transparent. The model and its owner are distinct questions (Chapter 41).

64.6 A Functional Perspective Is Not Obviously a Phenomenal One

A *functional perspective* is a set of agent-centered variables that a system reads and uses: where things are relative to its body, when events occur relative to its now, how its own states stand relative to its goals. Everything in the preceding five sections can be described functionally, and every capacity on the list has been or could be built into a machine.

A *phenomenal perspective* is there being something it is like to be the system (Nagel 1974). The book has not found, and no theory has demonstrated, a step that takes a system from the first to the second (Chapters 43 and 44). Illusionism denies that the second is anything over and above a system's representation of itself as having it (Frankish 2016); if illusionism is right, this final distinction marks a representational fact about certain self-models, and arrow 9 collapses into arrow 7. If it is wrong, the distinction marks the one place in the staircase where existing science has nothing to say.

64.7 The Question

The six distinctions leave one question standing:

> *What physical organization makes there be something it is like for relations to appear from here?*

Each phrase carries weight. *Physical organization*, because the answer, if there is one, must name a property that physical systems can have or lack, and that experiments could in principle detect. *Something it is like*, because the question is about phenomenality, not about report, access, or function. *Relations*, because what appears in experience is structured: things beside,

before, and after each other, larger and smaller, louder and softer. *Appear from here*, because the structure has an origin at the system, the first view from somewhere.

The book has established some constraints on any answer. It cannot be an input–output function alone, because an adaptive radio shares the jamming avoidance response's function and nobody holds that the radio experiences its channel (Chapter 48). If it is a property of the algorithmic level, independently evolved systems that share an algorithm, as *Eigenmannia* and *Gymnarchus* do, must share it too (Chapter 24). It must say which arrow of Figure 63.1 it crosses and what it adds there. And it must predict something that a rival answer forbids, so that the experiments of Part XIII can tell them apart.

Metaxics does not yet answer this. It has clarified the staircase leading to it: which steps are physics, which are biology, which are philosophy of content, and which one is still open.

64.8 What Metaxics Is, at the End

The Preface promised that the book would not assume Metaxics to be a science, and Part XII found that it is not one yet. What remains is a research perspective with three components. A question: how does a physical relation become a variable for a system that participates in it, and why is any such reader ever accompanied by a point of view? A method: subtract the established explanations before naming the residue, and coin nothing that existing vocabulary can express. And a programme: intervene on relations while holding their relata fixed, intervene on readers while holding their inputs fixed, open and close loops, and require every relational measure to beat the constituent measures it is built from.

The word itself returns, at the end, to its oldest sense. For Aristotle the *metaxu* was the medium of perception, the transparent something between the object and the eye without which nothing is seen (Aristotle 1984, II.7, 419a). For the electric fish the medium is the water, and the fish fills it with its own field before it can sense anything in it. Every relation the fish reads is carried by a medium that separates it from its neighbor and connects them. Weil's prisoners knew the same thing about their wall (Weil 2002, p. 145). Whether there is, for the fish, anything it is like to read that relation, and what in its organization would make it so, is the question this book leaves where it found it: open, and now precisely placed.

64.9 Audit

Canonical phenomenon

The six distinctions of this chapter, each instantiated by a system that has the first term without the second.

Minimal model
The staircase of Chapter 63 and the constraints of Section 64.7 on any answer to the final question.

What is relational here?
Every stage of the staircase; the final question concerns relations as they appear from a point of view.

What physically reads it?
Detectors, readers, agents, representers, self-models, and, if the final arrow is ever crossed, subjects.

What existing field explains it?
Every distinction belongs to an existing literature, from signal theory to the philosophy of mind; the final question is the hard problem in the form given by Nagel (1974) and Chalmers (1995).

What remains unexplained?
The final question.

A discriminating experiment
None yet for the question itself. For each proposed answer, the test is the one stated in Section 64.7: name the arrow, name what is added, and predict an outcome a rival answer forbids.

64.10 Notes and References

Nagel (1974) and Nagel (1986) frame the problem of the subjective point of view; Chalmers (1995) and Levine (1983) give the hard problem and the explanatory gap their standard forms. Metzinger (2003) and Graziano (2013) develop self-model and attention-schema theories; Frankish (2016) states illusionism. Aristotle's account of the medium of perception is in *De Anima* II.7 (Aristotle 1984); Weil's *metaxu* is in Weil (2002).

64.11 Exercises

Exercise 64.1. For each of the six distinctions, give a system of your own that has the first term without the second, and a system that has both. Which distinctions were hardest to illustrate, and why?

Exercise 64.2. State a proposed answer to the final question (your own or one from Chapter 44) and check it against each constraint in Section 64.7. Which constraint does it most easily fail?

Exercise 64.3. Suppose illusionism is correct. Rewrite Section 64.6 and Section 64.7 so that the final question becomes a question about self-models. What experiments from Part XIII would then bear on it directly?

Exercise 64.4. Write the strongest case you can that there is something it is like to be *Eigenmannia*, and the strongest case that there is not. Identify the single empirical finding that would most change your assessment.

Exercise 64.5. Using the method of the book, audit the book: list its claims of novelty, subtract the prior art for each, and state what residue, if any, survives.

Appendices

Which arrow have we actually explained?

A

Mathematical Prerequisites

This appendix collects the mathematics that the chapters use without derivation: linear algebra for stability and coordinates, complex numbers and phasors for oscillations and their relations, the trigonometric identities behind beats and demodulation, the calculus of periodic signals, probability, and the elementary quantities of information theory. It is a reference, not a course. Each section ends with pointers to fuller treatments in the Notes.

A.1 Linear Algebra

Vectors are written in bold lowercase ($\boldsymbol{x} \in \mathbb{R}^n$) and matrices in bold uppercase ($\boldsymbol{A} \in \mathbb{R}^{n\times n}$). A nonzero vector $\boldsymbol{v}$ is an *eigenvector* of $\boldsymbol{A}$ with *eigenvalue* λ if $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$; the eigenvalues are the roots of the characteristic polynomial $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$ and may be complex.

The linear system

$$\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} \tag{A.1}$$

has the solution $\boldsymbol{x}(t) = \mathrm{e}^{\boldsymbol{A}t}\boldsymbol{x}(0)$. When $\boldsymbol{A}$ has n linearly independent eigenvectors $\boldsymbol{v}_k$ with eigenvalues λ_k, every solution is a superposition

$$\boldsymbol{x}(t) = \sum_{k=1}^{n} c_k\, \mathrm{e}^{\lambda_k t}\, \boldsymbol{v}_k, \tag{A.2}$$

with coefficients c_k fixed by the initial condition. Each term is a *mode*. A mode decays if $\operatorname{Re}\lambda_k < 0$, grows if $\operatorname{Re}\lambda_k > 0$, and oscillates at angular frequency $\operatorname{Im}\lambda_k$ when the eigenvalue is complex. The origin is asymptotically stable exactly when every eigenvalue has negative real part.

For $n = 2$ the eigenvalues are determined by the trace $\tau = \operatorname{tr}\boldsymbol{A}$ and the determinant $\Delta = \det\boldsymbol{A}$:

$$\lambda_{1,2} = \tfrac{1}{2}\bigl(\tau \pm \sqrt{\tau^2 - 4\Delta}\bigr). \tag{A.3}$$

The fixed point is a saddle if $\Delta < 0$; a node if $\Delta > 0$ and $\tau^2 > 4\Delta$; a spiral (focus) if $\Delta > 0$ and $\tau^2 < 4\Delta$; stable if, in addition, $\tau < 0$. Linearizing a

nonlinear system about a fixed point and applying this classification is the first step of every stability analysis in Parts I to III; Appendix C continues it.

A change of coordinates $\boldsymbol{x} = \boldsymbol{P}\boldsymbol{y}$ with invertible $\boldsymbol{P}$ transforms (A.1) into $\dot{\boldsymbol{y}} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}\,\boldsymbol{y}$. The relational coordinates of Chapter 1, such as the center of mass and relative position of two bodies, are changes of coordinates of this kind in which the new variables separate.

A.2 Complex Numbers and Phasors

A complex number $z = a + \mathrm{i}b$ has modulus $|z| = \sqrt{a^2 + b^2}$ and argument $\arg z$, the angle it makes with the positive real axis. Euler's formula,

$$\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta, \tag{A.4}$$

writes every complex number as $z = |z|\mathrm{e}^{\mathrm{i}\arg z}$. Multiplication multiplies moduli and adds arguments.

A sinusoid can be written as the real part of a rotating complex number, its *phasor*:

$$A\cos(\omega t + \phi) = \mathrm{Re}\{A\,\mathrm{e}^{\mathrm{i}\phi}\,\mathrm{e}^{\mathrm{i}\omega t}\}. \tag{A.5}$$

Sinusoids of the same frequency add by adding their complex amplitudes $A\mathrm{e}^{\mathrm{i}\phi}$. Sinusoids of different frequencies can be added by factoring out one of them. For two components,

$$\begin{gathered}\cos\omega_1 t + m\cos(\omega_2 t + \psi_0) = \mathrm{Re}\Big\{\mathrm{e}^{\mathrm{i}\omega_1 t}\big(1 + m\,\mathrm{e}^{\mathrm{i}\psi(t)}\big)\Big\},\\ \psi(t) = (\omega_2 - \omega_1)t + \psi_0.\end{gathered} \tag{A.6}$$

The sum is a carrier at ω_1 whose amplitude and phase are the modulus and argument of the slowly rotating complex number $1 + m\mathrm{e}^{\mathrm{i}\psi}$. This is the form in which Part V analyzes the field on the skin of an electric fish, with $\omega_1 = \omega_\mathrm{s}$, $\omega_2 = \omega_\mathrm{e}$, and m the local contamination ratio.

The modulus and argument are

$$\begin{aligned}|1 + m\mathrm{e}^{\mathrm{i}\psi}| &= \sqrt{1 + m^2 + 2m\cos\psi},\\ \arg(1 + m\mathrm{e}^{\mathrm{i}\psi}) &= \mathrm{atan2}(m\sin\psi,\ 1 + m\cos\psi),\end{aligned} \tag{A.7}$$

and for $m \ll 1$ they reduce to

$$|1 + m\mathrm{e}^{\mathrm{i}\psi}| \approx 1 + m\cos\psi, \qquad \arg(1 + m\mathrm{e}^{\mathrm{i}\psi}) \approx m\sin\psi. \tag{A.8}$$

Amplitude and phase are then modulated in quadrature: one follows $\cos\psi$ and the other $\sin\psi$. Reversing the sign of $\omega_2 - \omega_1$ reverses the direction in which $1 + m\mathrm{e}^{\mathrm{i}\psi}$ circles the point 1, and the direction of circling is the only place in (A.6) where that sign appears.

A.3 Trigonometric Identities Used in the Book

The product-to-sum identities turn products of sinusoids into sums at the sum and difference frequencies:

$$\cos a \cos b = \tfrac{1}{2}\big[\cos(a-b) + \cos(a+b)\big], \tag{A.9}$$

$$\sin a \sin b = \tfrac{1}{2}\big[\cos(a-b) - \cos(a+b)\big], \tag{A.10}$$

$$\sin a \cos b = \tfrac{1}{2}\big[\sin(a+b) + \sin(a-b)\big]. \tag{A.11}$$

The sum-to-product identity turns a sum of two sinusoids into a product of an envelope and a carrier:

$$\cos a + \cos b = 2\cos\Big(\frac{a-b}{2}\Big)\cos\Big(\frac{a+b}{2}\Big). \tag{A.12}$$

The double-angle identity gives the square of a sinusoid:

$$\cos^2 a = \tfrac{1}{2}(1 + \cos 2a). \tag{A.13}$$

Together they explain the central example of Chapters 3 and 4. With $a = \omega_1 t$ and $b = \omega_2 t$, (A.12) writes the sum of two tones as an envelope at $(\omega_1 - \omega_2)/2$ times a carrier at $(\omega_1 + \omega_2)/2$, although the sum contains only the frequencies ω_1 and ω_2. Squaring the sum and applying (A.13) and (A.9) gives

$$\begin{aligned}(\cos\omega_1 t + \cos\omega_2 t)^2 = 1 &+ \tfrac{1}{2}\cos 2\omega_1 t + \tfrac{1}{2}\cos 2\omega_2 t \\ &+ \cos(\omega_1 - \omega_2)t + \cos(\omega_1 + \omega_2)t,\end{aligned} \tag{A.14}$$

which contains a constant term and components at $2\omega_1$, $2\omega_2$, $\omega_1 - \omega_2$, and $\omega_1 + \omega_2$. A square-law reader makes the difference frequency a component of its output; the linear sum had none.

A.4 Calculus of Oscillations

The harmonic oscillator $\ddot{x} + \omega_0^2 x = 0$ has solutions $x(t) = A\cos(\omega_0 t + \phi)$. With damping and a periodic drive,

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x = F\cos\omega t, \tag{A.15}$$

every solution approaches the steady state $x(t) = X\cos(\omega t - \chi)$ with

$$X = \frac{F}{\sqrt{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}}, \qquad \tan\chi = \frac{\gamma\omega}{\omega_0^2 - \omega^2}, \tag{A.16}$$

obtained by substituting the phasor $X e^{-i\chi} e^{i\omega t}$ into (A.15). Chapter 6 develops the resonance curve and the quality factor $Q = \omega_0/\gamma$.

Time averages over a period are written $\langle\cdot\rangle$. For sinusoids,

$$\begin{gathered}\langle\cos^2\omega t\rangle = \tfrac{1}{2}, \qquad \langle\cos\omega t\,\sin\omega t\rangle = 0,\\ \langle\cos\omega t\,\cos\omega' t\rangle = 0 \quad (\omega \neq \omega'),\end{gathered} \tag{A.17}$$

where the last average is taken over a time long compared with $2\pi/|\omega - \omega'|$. These *orthogonality* relations are the principle of lock-in detection and synchronous demodulation: multiplying a signal by $\cos\omega_{\mathrm{r}} t$ and averaging retains only the component at the reference frequency ω_{r}, and multiplying by $\sin\omega_{\mathrm{r}} t$ as well recovers its phase (Chapters 4 and 18). A low-pass filter implements the average physically, with a time constant that sets how long the averaging window is.

A.5 Probability

A random variable X has a distribution $\mathbb{P}(X = x)$ (or a density $p(x)$), an expectation $\mathbb{E}[X] = \sum_x x\,\mathbb{P}(X = x)$, and a variance $\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$. The covariance of two variables is $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. Two variables are *independent* when $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$ for all x, y; independence implies zero covariance, and the converse fails.

Conditional probability and Bayes' rule,

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(Y = y \mid X = x)\,\mathbb{P}(X = x)}{\mathbb{P}(Y = y)}, \tag{A.18}$$

underlie the generative models of Chapter 33.

Phases require their own statistics, because an angle and the same angle plus 2π are the same. For phases $\theta_1, \ldots, \theta_N$, the *mean resultant*

$$R\,\mathrm{e}^{\mathrm{i}\bar{\theta}} = \frac{1}{N}\sum_{j=1}^{N} \mathrm{e}^{\mathrm{i}\theta_j} \tag{A.19}$$

has length $R \in [0, 1]$, equal to 1 when all phases coincide and near 0 when they are spread uniformly, and mean direction $\bar{\theta}$. The Kuramoto order parameter of Chapter 9 is this quantity computed across oscillators at one time; the phase-locking value of Appendix K is the same quantity computed across trials or time for a phase difference.

A.6 Entropy and Mutual Information

For a discrete random variable X, the *entropy* is

$$H(X) = -\sum_x \mathbb{P}(X = x)\log_2 \mathbb{P}(X = x), \tag{A.20}$$

measured in bits; a fair coin has $H = 1$ bit. The conditional entropy $H(X \mid Y)$ is the average entropy of X given each value of Y, and the *mutual information*

$$I(X;Y) = H(X) - H(X \mid Y) = H(X) + H(Y) - H(X,Y) \tag{A.21}$$

measures how much knowing one variable reduces uncertainty about the other. Mutual information is symmetric, nonnegative, and zero exactly when the variables are independent. It obeys the chain rule $I(X,Y;Z) = I(X;Z) + I(Y;Z \mid X)$ and the *data processing inequality*: if Z depends on X only through Y, then $I(X;Z) \le I(X;Y)$. The inequality bounds what any reader can extract from a signal (Chapter 18).

The exclusive or shows that dependence need not be pairwise. With X and Y independent fair bits and $Z = X \oplus Y$,

$$I(X;Z) = I(Y;Z) = 0, \qquad I(X,Y;Z) = 1 \text{ bit}, \tag{A.22}$$

because each of X and Y alone leaves Z a fair coin, while together they determine it. Chapter 1 uses this example to define an irreducibly three-way relation, and Chapter 17 shows that the electric fish's sign computation has the same form. Appendix H develops the decompositions of multivariate information that treat such synergy systematically.

The natural logarithm gives information in nats; one nat is $1/\ln 2 \approx 1.44$ bits.

A.7 Notes and References

Strogatz (2015) covers linear and nonlinear dynamics at the level assumed here, including the trace–determinant classification. The driven damped oscillator and time averages are standard in Landau and Lifshitz (1976). Phasors, Fourier methods, and demodulation are treated in Bracewell (2000) and Oppenheim and Schafer (2010); Appendix B continues this appendix. Circular statistics and the order parameter appear throughout Pikovsky et al. (2001). Cover and Thomas (2006) is the standard reference for entropy, mutual information, and the data processing inequality.

A.8 Exercises

Exercise A.1. Classify the fixed point of $\dot{x} = -x + 2y$, $\dot{y} = -2x - y$ using (A.3). Write the general solution in the form (A.2).

Exercise A.2. Derive (A.7) from (A.6), then expand to first order in m to obtain (A.8). Show that the point $1 + me^{i\psi}$ circles 1 counterclockwise when $\omega_2 > \omega_1$ and clockwise when $\omega_2 < \omega_1$.

Exercise A.3. Verify (A.14). Then cube the sum $\cos\omega_1 t + \cos\omega_2 t$ and list every frequency in the result. Which of them lie close to ω_1 when ω_1 and ω_2 are close?

Exercise A.4. For $Z = X \oplus Y$ with X, Y independent fair bits, compute $H(X, Y, Z)$ and verify (A.22) using (A.21). Then compute $I(X; Y \mid Z)$ and explain why conditioning on Z creates dependence between variables that were independent.

B

Fourier and Spectral Methods

This appendix collects the Fourier-analytic facts used in Parts I and II and in the chapters on neural rhythms. It fixes conventions, states the rules that turn sums and products of signals into statements about spectra, and explains three things the main text leans on repeatedly: why products create difference frequencies and sums do not, why a finite observation window trades frequency resolution against time resolution, and how the envelope and phase of a signal are defined through the analytic signal. The last sections introduce cross-spectra and the bispectrum, the higher-order spectrum that detects the phase relations produced by quadratic nonlinearities.

B.1 Conventions

For a signal $x(t)$ we use

$$\hat{x}(\omega) = \int_{-\infty}^{\infty} x(t)\, \mathrm{e}^{-\mathrm{i}\omega t}\, \mathrm{d}t, \qquad x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{x}(\omega)\, \mathrm{e}^{\mathrm{i}\omega t}\, \mathrm{d}\omega, \tag{B.1}$$

with angular frequency $\omega = 2\pi f$. Energy is conserved in the form $\int |x(t)|^2 \mathrm{d}t = \frac{1}{2\pi} \int |\hat{x}(\omega)|^2 \mathrm{d}\omega$ (Parseval's theorem). A signal of period T has a Fourier series

$$x(t) = \sum_{n=-\infty}^{\infty} c_n\, \mathrm{e}^{\mathrm{i}n\omega_T t}, \qquad c_n = \frac{1}{T} \int_0^T x(t)\, \mathrm{e}^{-\mathrm{i}n\omega_T t}\, \mathrm{d}t, \qquad \omega_T = \frac{2\pi}{T}, \tag{B.2}$$

and its transform is a train of impulses, $\hat{x}(\omega) = 2\pi \sum_n c_n\, \delta(\omega - n\omega_T)$. For signals on a plane, such as the gratings of Chapter 2, the transform uses a wavevector: $\hat{x}(\boldsymbol{k}) = \int x(\boldsymbol{r})\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\, \mathrm{d}^2 r$.

B.2 Rules

The following pairs are used throughout; each follows directly from (B.1) (Bracewell 2000; Oppenheim and Schafer 2010).

Signal	Transform
$a\,x(t) + b\,y(t)$	$a\,\hat{x}(\omega) + b\,\hat{y}(\omega)$
$x(t-\tau)$	$\mathrm{e}^{-\mathrm{i}\omega\tau}\,\hat{x}(\omega)$
$\mathrm{e}^{\mathrm{i}\omega_0 t}\,x(t)$	$\hat{x}(\omega-\omega_0)$
$(x * y)(t) = \int x(\tau)\,y(t-\tau)\,\mathrm{d}\tau$	$\hat{x}(\omega)\,\hat{y}(\omega)$
$x(t)\,y(t)$	$\frac{1}{2\pi}(\hat{x} * \hat{y})(\omega)$
$\delta(t)$	1
1	$2\pi\,\delta(\omega)$
$\cos\omega_0 t$	$\pi\big[\delta(\omega-\omega_0) + \delta(\omega+\omega_0)\big]$

Two of these rules carry most of the weight in the main text. The convolution theorem says that a linear time-invariant system multiplies the input spectrum by its transfer function, which is why such a system cannot create frequencies (Chapter 4). The product rule says that multiplying two signals convolves their spectra, which is why products create sums and differences of frequencies.

For a real signal, $\hat{x}(-\omega) = \overline{\hat{x}(\omega)}$, so the magnitude spectrum is symmetric.

Proposition B.1 (Real signals do not carry the sign of a frequency)**.** *If x is real, $|\hat{x}(-\omega)| = |\hat{x}(\omega)|$ for every ω. In particular,* $\cos(\Omega t+\psi_0)$ *and* $\cos(-\Omega t-\psi_0)$ *are the same function, and no operation on a single real sinusoidal modulation can determine whether its frequency is $+\Omega$ or $-\Omega$.*

The sign of a frequency is meaningful only for a complex signal, such as $\mathrm{e}^{\mathrm{i}\Omega t}$, or equivalently for a pair of real signals in quadrature, $\cos\Omega t$ and $\sin\Omega t$, whose relative phase ($+90°$ or $-90°$) carries the sign. The proposition underlies the image problem of heterodyne receivers (Chapter 4) and the sign problem of the electric fish, which solves it by combining two modulations that are in quadrature (Chapter 21).

B.3 Sums and Products of Sinusoids

The sum $\cos\omega_1 t + \cos\omega_2 t$ has four impulses in its spectrum, at $\pm\omega_1$ and $\pm\omega_2$. The product has four impulses at other places:

$$\cos\omega_1 t\,\cos\omega_2 t \;\longleftrightarrow\; \frac{\pi}{2}\big[\,\delta(\omega-\omega_1-\omega_2) + \delta(\omega+\omega_1+\omega_2) + \delta(\omega-\omega_1+\omega_2) + \delta(\omega+\omega_1-\omega_2)\big], \tag{B.3}$$

obtained by convolving the two pairs of impulses and dividing by 2π. Each impulse of one factor is copied to the location of each impulse of the other: frequencies add under multiplication.

The same holds in two dimensions. A grating $\cos(\boldsymbol{k}_0\cdot\boldsymbol{r})$ has impulses at $\pm\boldsymbol{k}_0$, and the product of two gratings has impulses at $\pm\boldsymbol{k}_1 \pm \boldsymbol{k}_2$. When the gratings nearly match, the pair $\pm(\boldsymbol{k}_1 - \boldsymbol{k}_2)$ lies near the origin of the frequency

plane, and a reader that passes only low spatial frequencies sees those two impulses and nothing else: they are the moiré (Amidror 2009). Square-wave gratings have harmonics at $\pm n\boldsymbol{k}_i$, and their products have impulses at every $n_1\boldsymbol{k}_1 + n_2\boldsymbol{k}_2$; the ones near the origin are the higher-order moirés.

B.4 Sampling and Aliasing

Sampling a signal every T_s seconds multiplies it by a train of impulses. By the product rule, the spectrum of the sampled signal is a sum of shifted copies of the original spectrum,

$$x(t)\sum_n \delta(t - nT_s) \;\longleftrightarrow\; \frac{1}{T_s}\sum_m \hat{x}(\omega - m\omega_s), \qquad \omega_s = \frac{2\pi}{T_s}. \tag{B.4}$$

If $\hat{x}$ vanishes for $|\omega| \geq \omega_s/2$, the copies do not overlap and x can be recovered exactly from its samples (the sampling theorem). If not, a component at frequency ω reappears at $\omega - m\omega_s$ for every integer m, and one of those copies lands in the recorded band $|\omega| < \omega_s/2$. This is aliasing (Oppenheim and Schafer 2010). A pixel grid is a spatial sampler, and aliasing between a fine pattern in a scene and the grid is the moiré seen in photographs of fabric and screens (Chapter 2).

For a record of N samples, the discrete Fourier transform

$$X_k = \sum_{n=0}^{N-1} x_n \, \mathrm{e}^{-2\pi \mathrm{i} kn/N}, \qquad k = 0, \ldots, N-1, \tag{B.5}$$

evaluates the spectrum at frequencies $f_k = k/(NT_s)$, spaced by $1/T$, where $T = NT_s$ is the duration of the record. The fast Fourier transform computes all N values in a number of operations proportional to $N \log N$.

B.5 Finite Windows, Leakage, and Resolution

Every real measurement observes a signal for a finite time, which amounts to multiplying it by a window $w(t)$ that is nonzero only over the record. By the product rule, each spectral line of the signal is replaced by a copy of the window's transform. For the rectangular window of duration T,

$$\hat{w}(\omega) = \frac{2\sin(\omega T/2)}{\omega}, \tag{B.6}$$

a peak of width about $2\pi/T$ with zeros at $\omega = 2\pi k/T$ and slowly decaying sidelobes.

Two consequences follow. First, *resolution*: two lines of equal strength separated by $\Delta\omega = 2\pi/T$, or $\Delta f = 1/T$, can just be told apart, since each peak then falls on the other's first zero; closer lines merge into one peak. Second,

leakage: the sidelobes spread energy from strong lines into distant frequencies, where it can hide weak lines. Tapered windows, which fall smoothly to zero at the ends of the record, reduce sidelobes at the cost of a wider central peak (Percival and Walden 1993).

The resolution limit is the formal content of an observation made in Chapter 3. Two tones separated by Δf observed for a time much shorter than $1/\Delta f$ cannot be separated, and the record shows a single tone whose amplitude changes: a segment of the beat. Observed for much longer than $1/\Delta f$, they appear as two steady lines, and the beat is no longer visible as such. What an observer reports depends on its integration time, and no choice of window escapes the trade-off: narrowing a signal's extent in time broadens its extent in frequency, and conversely. The same trade-off fixes the product of a resonator's bandwidth and its response time (Chapter 6).

B.6 Power Spectra

For a stationary random signal, the power spectral density describes how its variance is distributed over frequency. The simplest estimate is the *periodogram*, the squared magnitude of the discrete Fourier transform of the record, suitably normalized. Its drawback is that its variance at each frequency does not decrease as the record grows longer: more data buys finer frequency spacing, not a steadier estimate.

Two standard remedies trade resolution for stability. Welch (1967) divided the record into overlapping segments, tapered each, computed a periodogram for each, and averaged them; averaging K roughly independent segments reduces the variance by a factor of about K, at the cost of a resolution set by the segment length rather than by the whole record. Thomson (1982) instead applied several orthogonal tapers, designed to concentrate each estimate as tightly as possible within a chosen bandwidth, to the whole record and averaged the resulting spectra; the multitaper method controls leakage and variance together and is treated at length by Percival and Walden (1993). Appendix K discusses the choices that matter for electrophysiological data, including the separation of rhythmic peaks from the broadband background.

B.7 The Analytic Signal

The envelope and phase of an oscillation are defined through the *analytic signal*

$$z(t) = x(t) + \mathrm{i}\,\mathcal{H}[x](t), \qquad \mathcal{H}[x](t) = \frac{1}{\pi}\,\mathrm{p.v.}\int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau}\,\mathrm{d}\tau, \tag{B.7}$$

where $\mathcal{H}$ is the Hilbert transform. In the frequency domain the construction is simple: $\hat{z}(\omega) = 2\hat{x}(\omega)$ for $\omega > 0$, $\hat{z}(0) = \hat{x}(0)$, and $\hat{z}(\omega) = 0$ for $\omega < 0$. The analytic signal discards the redundant negative-frequency half of a real signal's

spectrum and doubles the positive half. For $\omega_0 > 0$, $\mathcal{H}[\cos \omega_0 t] = \sin \omega_0 t$, so that the analytic signal of $\cos \omega_0 t$ is $\mathrm{e}^{\mathrm{i}\omega_0 t}$.

Definition B.2 (Envelope, instantaneous phase, instantaneous frequency)**.** For a real signal x with analytic signal z, the *envelope* is $A(t) = |z(t)|$, the *instantaneous phase* is $\phi(t) = \arg z(t)$, and the *instantaneous frequency* is $\dot{\phi}(t)$.

The Hilbert transform is linear. The modulus and the argument are not. The envelope is therefore a nonlinear functional of the signal, which is why it does not appear among the signal's Fourier components (Chapter 3).

For a sum of two tones with positive frequencies the construction is exact without any approximation:

$$\begin{aligned} z(t) &= a_1 \mathrm{e}^{\mathrm{i}\omega_1 t} + a_2 \mathrm{e}^{\mathrm{i}(\omega_2 t + \psi_0)}, \\ A(t) &= \sqrt{a_1^2 + a_2^2 + 2a_1 a_2 \cos\bigl((\omega_2 - \omega_1)t + \psi_0\bigr)}. \end{aligned} \tag{B.8}$$

For a general modulated carrier $x(t) = a(t)\cos\bigl(\omega_c t + \theta(t)\bigr)$, the envelope equals $|a(t)|$ and the phase equals $\omega_c t + \theta(t)$ when the spectrum of the complex modulation $a(t)\mathrm{e}^{\mathrm{i}\theta(t)}$ lies entirely within $|\omega| < \omega_c$, that is, when the modulation is slow compared with the carrier. When the modulation is not slow, or when the signal contains several unrelated components, the envelope and instantaneous frequency are still defined but may not correspond to anything a physical system tracks; in practice one band-pass filters around a frequency of interest before forming the analytic signal (M. X. Cohen 2014).

B.8 Cross-Spectra and Coherence

For two stationary signals x and y, the cross-spectrum $S_{xy}(\omega)$ is the expected product of the transform of one with the complex conjugate of the transform of the other, normalized by the record length. Its magnitude measures how strongly the two signals covary at frequency ω, and its phase gives the average phase lead of one over the other at that frequency. The *coherence*

$$C_{xy}(\omega) = \frac{|S_{xy}(\omega)|^2}{S_{xx}(\omega)\, S_{yy}(\omega)} \in [0, 1] \tag{B.9}$$

measures how well y can be predicted from x by a linear time-invariant filter, frequency by frequency. Coherence is a relational measure in the sense of Chapter 1, defined on a pair of signals and on neither alone. It is also a purely linear one: two signals related by a strong nonlinearity can have low coherence. Its estimation and its pitfalls in brain recordings, where a single source can appear in many sensors, are taken up in Appendix K.

B.9 Higher-Order Spectra

The power spectrum of a stationary signal depends only on its second-order statistics, and it is blind to phase: shifting the phases of a signal's components leaves the power spectrum unchanged. Relations among the phases of different components are visible only in higher-order statistics. The first of these is the bispectrum.

Definition B.3 (Bispectrum and bicoherence)**.** For a zero-mean stationary signal with third-order moment $c_3(\tau_1, \tau_2) = \mathbb{E}[x(t)\,x(t+\tau_1)\,x(t+\tau_2)]$, the *bispectrum* is its two-dimensional Fourier transform,

$$B(\omega_1, \omega_2) = \iint c_3(\tau_1, \tau_2)\, \mathrm{e}^{-\mathrm{i}(\omega_1 \tau_1 + \omega_2 \tau_2)}\, \mathrm{d}\tau_1\, \mathrm{d}\tau_2 .$$

From segments of a record with transforms $X(\omega)$, it is estimated as the average $\hat{B}(\omega_1, \omega_2) = \langle X(\omega_1)\, X(\omega_2)\, \overline{X(\omega_1 + \omega_2)} \rangle$. The *bicoherence* normalizes it to lie between zero and one; a common choice is

$$b^2(\omega_1, \omega_2) = \frac{|\hat{B}(\omega_1, \omega_2)|^2}{\langle |X(\omega_1) X(\omega_2)|^2 \rangle \, \langle |X(\omega_1 + \omega_2)|^2 \rangle} .$$

Each term of the average has the phase $\phi(\omega_1) + \phi(\omega_2) - \phi(\omega_1 + \omega_2)$. If the phases of the three components are unrelated from segment to segment, the terms point in random directions and the average tends to zero. If the phase at $\omega_1 + \omega_2$ equals the sum of the phases at ω_1 and ω_2 in every segment, the terms align and the bicoherence approaches one. That condition, *quadratic phase coupling*, is exactly what a quadratic nonlinearity produces: squaring a signal containing components at ω_1 and ω_2 creates a component at $\omega_1 + \omega_2$ whose phase is the sum of theirs (Chapter 4). For a Gaussian signal, and for a linear system driven by one, the bispectrum vanishes, since all statistics beyond the second order of a Gaussian process vanish (Nikias and Mendel 1993). The bispectrum is therefore a test for the kind of nonlinearity that makes difference and sum frequencies physical.

Bispectral measures have been applied to the electroencephalogram, most visibly in anesthesia monitoring (Sigl and Chamoun 1994). Whether they add information beyond ordinary spectral measures in a given application is an empirical question, and for anesthesia it has been disputed (Miller et al. 2004). Chapter 46 returns to the dispute, and to the general rule it illustrates: a relational measure earns its place only where it outperforms the simpler measures built from its constituents.

B.10 Notes and References

Bracewell (2000) is a classic treatment of the Fourier transform with many worked pairs and physical interpretations; Oppenheim and Schafer (2010) covers

sampling, the discrete Fourier transform, windows, and the Hilbert transform from the signal-processing side. Spectral estimation, including the multitaper method, is treated thoroughly by Percival and Walden (1993), building on Welch (1967) and Thomson (1982). Nikias and Mendel (1993) is a standard tutorial on higher-order spectra and quadratic phase coupling. Amidror (2009) develops the spectral approach to moiré patterns in full. For analysis of neural time series in practice, see M. X. Cohen (2014) and Appendix K.

C

Nonlinear Systems

This appendix collects the facts about nonlinear dynamical systems that the book uses, with the notation of the main text. Two kinds of system appear. *Autonomous* systems, $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$, describe oscillators, coupled populations, and the collective variables of Part III. *Input–output* systems map an input signal to an output signal and describe the readers of Part I and Part IV: detectors, mixers, and demodulators. Proofs are omitted; the Notes at the end give sources.

C.1 Fixed Points and Their Stability

Consider a system $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ with $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{f}$ smooth. A *fixed point* $\boldsymbol{x}^*$ satisfies $\boldsymbol{f}(\boldsymbol{x}^*) = 0$. Small deviations $\boldsymbol{\xi} = \boldsymbol{x} - \boldsymbol{x}^*$ obey, to first order, $\dot{\boldsymbol{\xi}} = \boldsymbol{J}\boldsymbol{\xi}$, with $\boldsymbol{J} = D\boldsymbol{f}(\boldsymbol{x}^*)$ the Jacobian matrix.

Theorem C.1 (Linear stability)**.** *If every eigenvalue of $\boldsymbol{J}$ has negative real part, $\boldsymbol{x}^*$ is asymptotically stable. If some eigenvalue has positive real part, $\boldsymbol{x}^*$ is unstable. If no eigenvalue has zero real part (the fixed point is* hyperbolic*), the flow near $\boldsymbol{x}^*$ is topologically equivalent to the flow of its linearization (Hartman–Grobman).*

In the plane the eigenvalues are fixed by the trace τ and determinant Δ of $\boldsymbol{J}$. If $\Delta < 0$ the fixed point is a saddle. If $\Delta > 0$ it is stable when $\tau < 0$ and unstable when $\tau > 0$, and it is a spiral when $\tau^2 < 4\Delta$ and a node otherwise. Non-hyperbolic fixed points ($\tau = 0$ with $\Delta > 0$, or $\Delta = 0$) are where bifurcations happen, and there the linearization does not decide the question.

C.2 Limit Cycles

A *limit cycle* is an isolated closed orbit. Nearby trajectories spiral toward it (a stable cycle) or away from it (an unstable one). Limit cycles are the mathematical form of self-sustained oscillation: a pendulum clock, a pacemaker neuron, the electric organ discharge of a wave-type electric fish, a laser above

threshold. Their amplitude and period are set by the system and not by initial conditions, which is what distinguishes them from the closed orbits of a frictionless pendulum.

In the plane the Poincaré–Bendixson theorem gives a practical criterion: a trajectory that stays in a bounded region containing no fixed point approaches a closed orbit. The standard example is the normal form of the Hopf bifurcation, written for a complex variable $z = x + \mathrm{i}y$:

$$\dot{z} = (\mu + \mathrm{i}\omega)\, z - |z|^2 z. \tag{C.1}$$

In polar coordinates $z = R\mathrm{e}^{\mathrm{i}\phi}$ it separates into $\dot{R} = \mu R - R^3$ and $\dot{\phi} = \omega$. For $\mu > 0$ every trajectory other than the origin approaches the circle $R = \sqrt{\mu}$, on which the phase advances at rate ω. Stability of a cycle in higher dimensions is determined by its Floquet multipliers, the eigenvalues of the linearized return map on a surface that the cycle crosses once per period; the cycle is stable when all multipliers other than the trivial one lie inside the unit circle. Appendix D describes how the phase of a stable cycle can be extended to its whole basin of attraction.

C.3 Elementary Bifurcations

A *bifurcation* is a qualitative change in the set of attractors as a parameter μ passes a critical value. Near a bifurcation the dynamics reduces, by the center manifold theorem of Section C.4, to a low-dimensional *normal form*. The following normal forms cover every bifurcation in the book.

- *Saddle-node*: $\dot{x} = \mu - x^2$. For $\mu > 0$ there are fixed points at $\pm\sqrt{\mu}$, the positive one stable; they collide and disappear at $\mu = 0$.
- *Transcritical*: $\dot{x} = \mu x - x^2$. Two fixed points exchange stability as they pass through each other.
- *Supercritical pitchfork*: $\dot{x} = \mu x - x^3$. The origin loses stability at $\mu = 0$ and two stable fixed points $\pm\sqrt{\mu}$ appear. This is the normal form reached by adiabatic elimination in Chapter 13.
- *Subcritical pitchfork*: $\dot{x} = \mu x + x^3 - x^5$. The branches bend backward, the transition is discontinuous, and there is a range of μ with coexisting stable states and hysteresis.
- *Hopf*: $\dot{z} = (\mu + \mathrm{i}\omega)z - (1 + \mathrm{i}c)|z|^2 z$. A limit cycle of radius $\sqrt{\mu}$ and frequency $\omega - c\mu$ is born from the origin. The parameter c makes the frequency depend on the amplitude (*non-isochronicity*), which becomes a phase lag in the coupling of such oscillators (Appendix D).
- *Saddle-node on an invariant circle*: $\dot{\theta} = \nu - K\sin\theta$ on the circle. For $|\nu| < K$ there is a stable and an unstable fixed point; at $|\nu| = K$ they merge, and for $|\nu| > K$ the phase drifts around the circle with period $2\pi/\sqrt{\nu^2 - K^2}$.

The last normal form deserves a name of its own. With θ read as the phase difference between an oscillator and a periodic drive, ν as their frequency difference, and K as the coupling, it is Adler's equation for the locking of an oscillator to an injected signal (R. Adler 1946). Locking occurs when $|\nu| \le K$; outside that range the oscillator drifts relative to the drive, and the time-averaged drift rate grows as $\sqrt{\nu^2 - K^2}$ from zero at the edge of locking. The same equation governs a phase-locked loop near lock (Chapter 15), two weakly coupled oscillators (Appendix D), and the boundaries of the 1:1 Arnold tongue (Appendix F).

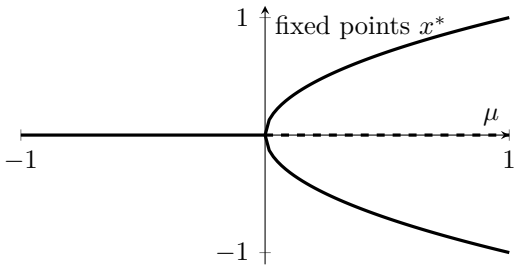


Figure C.1: Supercritical pitchfork, $\dot{x} = \mu x - x^3$. Solid: stable fixed points. Dashed: unstable. The two new stable branches are $x^* = \pm\sqrt{\mu}$.

C.4 Center Manifolds and Slow Manifolds

At a bifurcation some eigenvalues of the Jacobian have zero real part. The *center manifold theorem* states that there is then a locally invariant manifold, tangent at the fixed point to the eigenspace of those critical eigenvalues, that contains all the dynamics that do not decay exponentially. The stable directions are functions of the critical ones on this manifold, and the long-time behavior near the fixed point is determined by the reduced dynamics on it. Adding the bifurcation parameter as a trivial variable, $\dot{\mu} = 0$, extends the result to a neighborhood of the bifurcation.

When the critical eigenvalues are small but not zero, and the others are large and negative, the invariant manifold persists as a *slow manifold* that attracts nearby trajectories quickly and carries slow dynamics. The adiabatic elimination of Chapter 13 is the leading-order computation of such a manifold, and Haken's slaving principle is a physically oriented statement of its existence. The manifold's shape follows from the invariance condition, as in Chapter 13: if the fast variables satisfy $\boldsymbol{f} = \boldsymbol{h}(\boldsymbol{s})$ on the manifold, then the time derivative of $\boldsymbol{h}(\boldsymbol{s})$ along the slow flow must equal the fast dynamics evaluated on the manifold.

C.5 Nonlinear Input–Output Systems

A reader is an input–output system, and its nonlinearity decides which relations in its input become explicit in its output (Chapters 4 and 16).

Memoryless polynomial nonlinearities. A system with output $z(t) = \sum_{k\ge 1} c_k\, y(t)^k$ acts instantaneously on its input. For a two-tone input $y = \cos\omega_1 t + \cos\omega_2 t$, the term of order k produces frequencies $|p\,\omega_1 \pm q\,\omega_2|$ with nonnegative integers $p+q \le k$ and $p+q \equiv k \pmod 2$. The quadratic term produces a constant, $2\omega_1$, $2\omega_2$, and the sum and difference frequencies $\omega_1 \pm \omega_2$; the cubic term produces $3\omega_1$, $3\omega_2$, $2\omega_1 \pm \omega_2$, and $2\omega_2 \pm \omega_1$, as well as contributions at ω_1 and ω_2. The components at $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$ lie close to the input frequencies when the tones are close, and they are the cubic distortion products measured in the ear (Goldstein 1967; Kemp 1978; Robles and Ruggero 2001).

Volterra and Wiener series. A nonlinear system with memory can often be represented by a Volterra series,

$$z(t) = h_0 + \sum_{n=1}^{\infty} \int\!\cdots\!\int h_n(\tau_1,\dots,\tau_n) \prod_{i=1}^{n} y(t-\tau_i)\,\mathrm{d}\tau_i, \tag{C.2}$$

a generalization of the impulse response of a linear system to kernels of every order. The first-order kernel is the ordinary impulse response; the second-order kernel describes how pairs of past input values combine, and it is the term responsible for sum and difference frequencies in a two-tone input. Wiener reorganized the series into terms that are mutually orthogonal for Gaussian white-noise inputs, which makes the kernels estimable from input–output data by cross-correlation (Schetzen 1980).

Harmonic balance. For a nonlinear oscillator driven periodically, the method of harmonic balance assumes that the response is dominated by the driving frequency and matches the coefficients of that harmonic. For the Duffing oscillator,

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x + \beta x^3 = F\cos\omega t, \tag{C.3}$$

the ansatz $x \approx a\cos(\omega t - \varphi)$ and the identity $\cos^3\theta = \tfrac{3}{4}\cos\theta + \tfrac{1}{4}\cos 3\theta$ give, after discarding the third harmonic,

$$a^2\Big[\big(\omega_0^2 - \omega^2 + \tfrac{3}{4}\beta a^2\big)^2 + \gamma^2\omega^2\Big] = F^2. \tag{C.4}$$

The *backbone curve* $\omega^2 = \omega_0^2 + \tfrac{3}{4}\beta a^2$ shows the resonance frequency moving with amplitude, upward for a hardening spring ($\beta > 0$) and downward for a softening one. For large enough drive, (C.4) has three solutions for a over a range of ω, two of them stable, and the response jumps between branches as

the drive frequency is swept, with hysteresis (Nayfeh and Mook 1979; Landau and Lifshitz 1976). Chapter 6 uses the result to separate nonlinear resonance from linear resonance.

C.6 Notes and References

Strogatz (2015) is the most accessible introduction to fixed points, limit cycles, and bifurcations. Guckenheimer and Holmes (1983) give the rigorous treatment, including the Hartman–Grobman and center manifold theorems, and Kuznetsov (2004) is the standard reference for normal forms and for numerical bifurcation analysis. Nonlinear oscillations, the Duffing equation, and the method of harmonic balance are treated in Nayfeh and Mook (1979); Landau and Lifshitz (1976) give a compact account of nonlinear and parametric resonance. Adler's locking equation is R. Adler (1946). Schetzen (1980) covers Volterra and Wiener series and their identification. Cochlear distortion products are discussed in Goldstein (1967), Kemp (1978), and Robles and Ruggero (2001).

C.7 Exercises

Exercise C.1. For the Hopf normal form with non-isochronicity, $\dot z = (\mu + \mathrm{i}\omega)z - (1 + \mathrm{i}c)|z|^2 z$, derive the equations for R and ϕ, find the limit cycle for $\mu > 0$, and show that its frequency is $\omega - c\mu$.

Exercise C.2. For Adler's equation $\dot\theta = \nu - K\sin\theta$ with $|\nu| > K$, compute the period of drift by integrating $\mathrm{d}\theta/(\nu - K\sin\theta)$ over one turn, and confirm that the mean drift rate is $\sqrt{\nu^2 - K^2}$. Sketch the mean drift rate against ν for fixed K and identify the locking range.

Exercise C.3. List all frequencies produced by $z = y + c_2 y^2 + c_3 y^3$ for $y = \cos\omega_1 t + \cos\omega_2 t$, with their amplitudes. Which of them would a low-pass filter with cutoff well below ω_1 and ω_2 pass, and what relational variable does each surviving term make explicit?

Exercise C.4. Derive (C.4). For $\omega_0 = 1$, $\gamma = 0.1$, $\beta = 0.2$, and $F = 0.5$, plot a against ω, identify the range of ω with three solutions, and describe what happens when the drive frequency is swept slowly upward and then downward.

D

Phase Reduction

Many chapters treat an oscillator as nothing but a phase: the Kuramoto model (Chapter 9), the relative-phase dynamics of coordination (Chapter 10), chimeras (Chapter 12), the phase-locked loop (Chapter 15), and the electric organ discharge of the fish (Part V). This appendix explains when that is legitimate. A weakly perturbed stable limit cycle can be described by a single phase variable, and weakly coupled limit-cycle oscillators can be described by equations for their phase differences alone. The reduction also shows where the phase lags and coupling functions used in the main text come from.

D.1 Asymptotic Phase and Isochrons

Let $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ have a stable limit cycle $\boldsymbol{x}_0(t)$ of period T, and write $\omega = 2\pi/T$. On the cycle the phase is defined by the time elapsed since a reference point: $\theta(\boldsymbol{x}_0(t)) = \omega t \bmod 2\pi$, so that $\dot{\theta} = \omega$.

The definition extends to every point of the cycle's basin of attraction. A point $\boldsymbol{x}$ off the cycle is attracted toward it, and its trajectory becomes asymptotically indistinguishable from that of exactly one point $\boldsymbol{y}$ on the cycle. The *asymptotic phase* of $\boldsymbol{x}$ is the phase of that point, $\theta(\boldsymbol{x}) = \theta(\boldsymbol{y})$. The level sets of θ are the *isochrons*: all points on one isochron end up in step with each other (Winfree 2001; Guckenheimer 1975). By construction the phase advances uniformly along every trajectory in the basin, which gives the identity

$$\nabla\theta(\boldsymbol{x}) \cdot \boldsymbol{f}(\boldsymbol{x}) = \omega. \tag{D.1}$$

D.2 The Phase Response Curve

Perturb the oscillator weakly, $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) + \varepsilon\, \boldsymbol{p}(t)$ with ε small. By the chain rule and (D.1), $\dot{\theta} = \omega + \varepsilon\, \nabla\theta(\boldsymbol{x}) \cdot \boldsymbol{p}(t)$. Because the perturbation is weak, the state stays close to the cycle, $\boldsymbol{x} \approx \boldsymbol{x}_0(\theta)$, and to first order in ε

$$\dot{\theta} = \omega + \varepsilon\, \boldsymbol{Z}(\theta) \cdot \boldsymbol{p}(t), \qquad \boldsymbol{Z}(\theta) = \nabla\theta(\boldsymbol{x}_0(\theta)). \tag{D.2}$$

The vector function $\boldsymbol{Z}(\theta)$ is the *phase sensitivity function*, also called the infinitesimal phase response curve. It says how much a small push in each direction, delivered at phase θ, advances or delays the oscillator. It satisfies the normalization $\boldsymbol{Z}(\theta)\cdot\boldsymbol{f}(\boldsymbol{x}_0(\theta))=\omega$, and it can be computed as the periodic solution of the adjoint equation $\dot{\boldsymbol{Z}}=-\boldsymbol{J}(\boldsymbol{x}_0(t))^{\mathsf{T}}\boldsymbol{Z}$, where $\boldsymbol{J}$ is the Jacobian of $\boldsymbol{f}$ (Ermentrout and Terman 2010).

Experimentally, the phase response curve is measured by delivering brief perturbations at different phases and recording the resulting shift in the timing of later cycles. For small perturbations the shift is $\boldsymbol{Z}(\theta)\cdot\Delta\boldsymbol{x}$. Two broad shapes recur in neuroscience. *Type I* curves are nonnegative, so a small excitatory push can only advance the oscillator; they arise when oscillation begins through a saddle-node bifurcation on an invariant circle. *Type II* curves change sign over the cycle, so the same push advances the oscillator at some phases and delays it at others; they arise when oscillation begins through a Hopf bifurcation (Ermentrout and Terman 2010; Izhikevich 2007). The shape of the curve is one of the main determinants of whether and how coupled oscillators synchronize.

D.3 A Worked Example

The Stuart–Landau oscillator is the normal form of the Hopf bifurcation (Appendix C), scaled to put the limit cycle on the unit circle:

$$\dot{z}=(1+\mathrm{i}\omega)\,z-(1+\mathrm{i}c)\,|z|^2 z. \tag{D.3}$$

In polar coordinates $z=R\mathrm{e}^{\mathrm{i}\phi}$ it becomes $\dot{R}=R-R^3$ and $\dot{\phi}=\omega-cR^2$. The limit cycle is $R=1$, on which the frequency is $\omega-c$. The parameter c makes the frequency depend on the amplitude: an oscillator knocked outward runs slower (for $c>0$) until it relaxes back.

The asymptotic phase is

$$\theta=\phi-c\ln R, \tag{D.4}$$

as a direct check confirms: $\dot{\theta}=\omega-cR^2-c(1-R^2)=\omega-c$ everywhere, not only on the cycle. The isochrons $\phi-c\ln R=\text{const}$ are logarithmic spirals when $c\neq 0$ and radial lines when $c=0$ (Figure D.1). Differentiating (D.4) with respect to $x=R\cos\phi$ and $y=R\sin\phi$ and evaluating on the cycle gives the phase sensitivity function

$$Z_x(\theta)=-\sin\theta-c\cos\theta,\qquad Z_y(\theta)=\cos\theta-c\sin\theta, \tag{D.5}$$

a sinusoidal, Type II response. The normalization holds: on the cycle $\boldsymbol{f}=(\omega-c)(-\sin\theta,\cos\theta)$, and $\boldsymbol{Z}\cdot\boldsymbol{f}=\omega-c$.

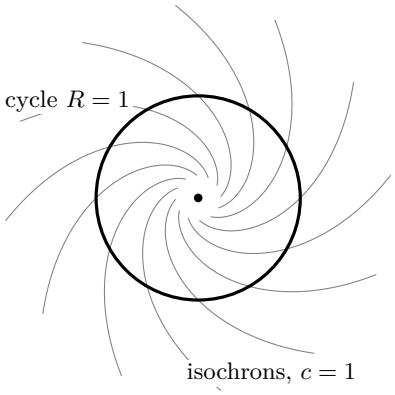


Figure D.1: Isochrons of the Stuart–Landau oscillator (D.3) with $c = 1$, drawn every 30° of asymptotic phase. Points on one spiral fall into step with the point where that spiral crosses the limit cycle (thick circle). With $c = 0$ the isochrons would be straight rays.

D.4 Weak Coupling and Averaging

Now couple two oscillators with the same limit cycle, or with nearly equal frequencies ω_1 and ω_2:

$$\dot{\boldsymbol{x}}_i = \boldsymbol{f}_i(\boldsymbol{x}_i) + \varepsilon\, \boldsymbol{P}(\boldsymbol{x}_i, \boldsymbol{x}_j), \qquad i \neq j \in \{1, 2\}. \tag{D.6}$$

By (D.2), $\dot\theta_i = \omega_i + \varepsilon\, \boldsymbol{Z}(\theta_i) \cdot \boldsymbol{P}\big(\boldsymbol{x}_0(\theta_i), \boldsymbol{x}_0(\theta_j)\big)$. Write $\theta_i = \bar\omega t + \vartheta_i$ with $\bar\omega$ a common reference frequency. When ε and $|\omega_1 - \omega_2|$ are small, the ϑ_i change little over one cycle, and averaging the right-hand side over a cycle at fixed ϑ_i gives

$$\dot\vartheta_i = (\omega_i - \bar\omega) + \varepsilon\, \Gamma(\vartheta_j - \vartheta_i), \tag{D.7}$$

$$\Gamma(\varphi) = \frac{1}{2\pi} \int_0^{2\pi} \boldsymbol{Z}(\psi) \cdot \boldsymbol{P}\big(\boldsymbol{x}_0(\psi), \boldsymbol{x}_0(\psi + \varphi)\big)\, \mathrm{d}\psi. \tag{D.8}$$

The coupling now depends only on the phase difference. This is the step that turns coupled limit-cycle oscillators into the phase models of Part III (Kuramoto 1984; Nakao 2016).

For the relative phase $\varphi = \theta_1 - \theta_2$ and the frequency difference $\Delta\omega = \omega_1 - \omega_2$,

$$\dot\varphi = \Delta\omega + \varepsilon\big[\Gamma(-\varphi) - \Gamma(\varphi)\big] = \Delta\omega - 2\varepsilon\, \Gamma_{\mathrm{odd}}(\varphi), \tag{D.9}$$

where $\Gamma_{\mathrm{odd}}(\varphi) = \tfrac{1}{2}\big[\Gamma(\varphi) - \Gamma(-\varphi)\big]$. Only the odd part of the coupling function matters for locking. A locked state is a fixed point φ^* with $\Delta\omega = 2\varepsilon\, \Gamma_{\mathrm{odd}}(\varphi^*)$, stable when $\Gamma'_{\mathrm{odd}}(\varphi^*) > 0$, and locking is possible exactly when $|\Delta\omega| \le 2\varepsilon \max|\Gamma_{\mathrm{odd}}|$.

Example D.1 (Coupled Stuart–Landau oscillators). Couple two Stuart–Landau oscillators diffusively through their complex amplitudes, $\boldsymbol{P}(z_i, z_j) = z_j - z_i$.

On the cycle, $z_i = \mathrm{e}^{\mathrm{i}\theta_i}$, and (D.4) gives $\mathrm{d}\theta = \mathrm{Im}(\mathrm{d}z/z) - c\,\mathrm{Re}(\mathrm{d}z/z)$. With $\mathrm{d}z = \varepsilon(z_j - z_i)\,\mathrm{d}t$ and $\varphi = \theta_j - \theta_i$,

$$\Gamma(\varphi) = \sin\varphi + c\,(1 - \cos\varphi) = c + \sqrt{1 + c^2}\,\sin(\varphi - \alpha), \qquad \tan\alpha = c. \tag{D.10}$$

Here the reduction is exact to first order in ε, without averaging, because the coupling respects the rotational symmetry of the oscillator. The odd part is $\Gamma_{\mathrm{odd}}(\varphi) = \sin\varphi$, so (D.9) becomes $\dot\varphi = \Delta\omega - 2\varepsilon\sin\varphi$: Adler's locking equation (R. Adler 1946), with locking for $|\Delta\omega| \le 2\varepsilon$. In a population of such oscillators, the reduced model is the Kuramoto model with a phase lag α, the model of Sakaguchi and Kuramoto (1986), together with a frequency shift proportional to c. The dependence of frequency on amplitude in the individual oscillators is therefore one physical origin of the phase lag that frustrates synchronization and makes the chimeras of Chapter 12 possible.

D.5 Periodic Forcing

A single oscillator driven by a weak periodic force $\varepsilon\,\boldsymbol{p}(\Omega t)$, with $\boldsymbol{p}$ of period 2π and Ω close to ω, reduces in the same way. With $\psi = \theta - \Omega t$ the phase relative to the drive,

$$\dot\psi = (\omega - \Omega) + \varepsilon\,H(\psi), \qquad H(\psi) = \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{Z}(\psi + s)\cdot\boldsymbol{p}(s)\,\mathrm{d}s. \tag{D.11}$$

The oscillator locks to the drive when $\omega - \Omega$ lies between $-\varepsilon\max H$ and $-\varepsilon\min H$. In the plane of drive frequency and drive strength, the locking region is a wedge whose width grows linearly with ε near its tip: the 1:1 Arnold tongue. Locking at other ratios p:q appears at higher orders in ε and gives narrower tongues, whose organization is the subject of Appendix F and Chapter 7 (Pikovsky et al. 2001).

D.6 When the Reduction Fails

Phase reduction assumes that the limit cycle attracts perturbed states quickly compared with the rate at which the coupling or forcing moves the phase. It fails in three situations that the main text meets. Near the bifurcation that creates the oscillation, the cycle's attraction is weak, amplitude and phase evolve on comparable timescales, and the amplitude must be kept. With strong coupling, the oscillators leave their cycles altogether, and phenomena without a phase description appear, such as the mutual suppression of oscillation known as amplitude death (Pikovsky et al. 2001). And with noise, the phase diffuses, so that a phase description holds only on average and locking becomes a statement about the distribution of phase differences. In each case the relational variable of interest, the phase difference, is still defined, but its dynamics are no longer closed.

D.7 Notes and References

Winfree introduced the phase description of biological oscillators and the notion of phase resetting (Winfree 1967; Winfree 2001); Guckenheimer (1975) established the properties of isochrons and of the phaseless sets where they converge. Kuramoto derived phase equations for weakly coupled oscillators, including the reduction from the complex Ginzburg–Landau equation that yields a phase lag (Kuramoto 1984). Nakao (2016) is a clear modern review of phase reduction, with the Stuart–Landau example; Ermentrout and Terman (2010) and Izhikevich (2007) treat phase response curves and their use in neuroscience; Pikovsky et al. (2001) cover forcing, locking, and the limits of the phase description. The phase-lag model is Sakaguchi and Kuramoto (1986); Adler's equation is R. Adler (1946).

D.8 Exercises

Exercise D.1. Verify that (D.4) satisfies $\dot\theta = \omega - c$ throughout the plane minus the origin. Derive (D.5) and check the normalization $\boldsymbol{Z}\cdot\boldsymbol{f} = \omega - c$.

Exercise D.2. Simulate the Stuart–Landau oscillator with $\omega = 2$ and $c = 0.5$. At each of 24 phases, apply an instantaneous kick of size 0.05 in the x direction, integrate until the state has returned to the cycle, and record the phase shift. Plot the measured shifts divided by 0.05 against phase and compare with Z_x from (D.5).

Exercise D.3. Derive (D.10). For two coupled oscillators with frequency difference $\Delta\omega$, find the locked phase difference and the common frequency of the locked pair, and show that the common frequency depends on c.

Exercise D.4. For the Stuart–Landau oscillator forced by $\boldsymbol{p}(s) = (\cos s, 0)$, compute $H(\psi)$ in (D.11) and show that the 1:1 tongue is $|\omega - c - \Omega| \le \tfrac{1}{2}\varepsilon\sqrt{1+c^2}$, where $\omega - c$ is the unforced frequency.

E

The Kuramoto Model

This appendix collects the derivations behind Chapter 9: the continuum limit of the Kuramoto model, its stationary states, the self-consistency condition and the threshold, the expansion near threshold, the exact low-dimensional reduction for Lorentzian frequencies, finite-size effects, and a numerical recipe. The notation follows Chapter 9.

E.1 The Model and Its Mean Field

N phase oscillators with natural frequencies ω_i drawn from a density $g(\omega)$ obey

$$\dot{\theta}_i = \omega_i + \frac{K}{N}\sum_{j=1}^{N}\sin(\theta_j - \theta_i). \tag{E.1}$$

With the complex order parameter $z = r\,\mathrm{e}^{\mathrm{i}\psi} = N^{-1}\sum_j \mathrm{e}^{\mathrm{i}\theta_j}$, the identity $\mathrm{Im}(z\,\mathrm{e}^{-\mathrm{i}\theta_i}) = N^{-1}\sum_j \sin(\theta_j - \theta_i)$ gives the exact mean-field form

$$\dot{\theta}_i = \omega_i + Kr\sin(\psi - \theta_i) = \omega_i + K\ \mathrm{Im}\big(z\,\mathrm{e}^{-\mathrm{i}\theta_i}\big). \tag{E.2}$$

The model is invariant under a common shift of all phases, and under a shift of all frequencies by Ω combined with the change to a frame rotating at Ω. We assume throughout that g is symmetric about a center, and we choose the rotating frame in which the center is 0.

E.2 The Continuum Limit

As $N \to \infty$, the state of the population is described by a density $\rho(\theta, \omega, t)$: the fraction of oscillators with natural frequency ω that have phase near θ at time t, normalized so that $\int_0^{2\pi} \rho\,\mathrm{d}\theta = 1$ for each ω. Oscillators are conserved, so ρ obeys the continuity equation

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial \theta}(\rho\, v) = 0, \qquad v(\theta, \omega, t) = \omega + K\ \mathrm{Im}\big(z\,\mathrm{e}^{-\mathrm{i}\theta}\big), \tag{E.3}$$

with the order parameter now given self-consistently by

$$z(t) = \int_{-\infty}^{\infty} \int_{0}^{2\pi} \mathrm{e}^{\mathrm{i}\theta}\, \rho(\theta, \omega, t)\, g(\omega)\, \mathrm{d}\theta\, \mathrm{d}\omega. \tag{E.4}$$

The incoherent state $\rho = 1/2\pi$, $z = 0$, is a stationary solution for every K.

E.3 Stationary States

Look for stationary states with constant $r > 0$ and, by the choice of frame, $\psi = 0$. The velocity is $v = \omega - Kr \sin\theta$.

Locked oscillators. If $|\omega| \le Kr$, the velocity vanishes at $\sin\theta = \omega/Kr$, and the stable zero lies at $\theta^*(\omega) = \arcsin(\omega/Kr) \in [-\pi/2, \pi/2]$. Locked oscillators sit at this phase: $\rho = \delta(\theta - \theta^*(\omega))$.

Drifting oscillators. If $|\omega| > Kr$, the velocity never vanishes, and a stationary density must make the flux ρv constant in θ:

$$\rho(\theta, \omega) = \frac{\sqrt{\omega^2 - K^2 r^2}}{2\pi\, |\omega - Kr \sin\theta|}, \tag{E.5}$$

where the prefactor fixes the normalization. The density is largest where the oscillators move slowest.

Contribution of the drifting group. Because (E.5) is unchanged under $\theta \mapsto \pi - \theta$, which reverses the sign of $\cos\theta$, the drifting oscillators at each ω contribute nothing to $\operatorname{Re} z$. Their contribution to $\operatorname{Im} z$ at ω is canceled by that at $-\omega$ when g is symmetric, since $\rho(\theta+\pi, -\omega) = \rho(\theta, \omega)$ and $\sin(\theta+\pi) = -\sin\theta$. The drifting group therefore contributes nothing to the order parameter.

E.4 The Self-Consistency Condition and the Threshold

Only the locked oscillators contribute, and the imaginary parts of their contributions cancel by symmetry. With the substitution $\omega = Kr \sin\theta$,

$$r = \int_{-Kr}^{Kr} \cos\theta^*(\omega)\, g(\omega)\, \mathrm{d}\omega = Kr \int_{-\pi/2}^{\pi/2} \cos^2\theta\; g(Kr \sin\theta)\, \mathrm{d}\theta. \tag{E.6}$$

Dividing by r and letting $r \to 0^+$ gives $1 = Kg(0) \int_{-\pi/2}^{\pi/2} \cos^2\theta\, \mathrm{d}\theta = Kg(0)\pi/2$, so the partially locked branch leaves the incoherent state at

$$K_c = \frac{2}{\pi g(0)}. \tag{E.7}$$

Behavior near threshold. Expand $g(x) = g(0) + \frac{1}{2}g''(0)x^2 + O(x^4)$ in (E.6). Using $\int_{-\pi/2}^{\pi/2} \cos^2\theta \sin^2\theta \, d\theta = \pi/8$,

$$1 = \frac{K}{K_c} + \frac{\pi g''(0)}{16} K^3 r^2 + O(r^4),$$

so that, for $g''(0) < 0$,

$$r^2 = \frac{16}{\pi K^3 |g''(0)|}\Big(\frac{K}{K_c} - 1\Big) \approx \frac{16\,(K - K_c)}{\pi K_c^4 |g''(0)|}. \tag{E.8}$$

The coherence grows as $(K - K_c)^{1/2}$: the transition is continuous (a supercritical bifurcation) for every smooth, symmetric, unimodal g with $g''(0) < 0$.

Examples.

- *Lorentzian*, $g(\omega) = \gamma/[\pi(\gamma^2 + \omega^2)]$: $g(0) = 1/\pi\gamma$, so $K_c = 2\gamma$. The integral in (E.6) can be done exactly, giving $r = \sqrt{1 - 2\gamma/K}$ for $K > 2\gamma$.
- *Gaussian* of standard deviation σ: $g(0) = 1/(\sigma\sqrt{2\pi})$, so $K_c = \sigma\sqrt{8/\pi} \approx 1.60\,\sigma$.
- *Uniform* on $[-\gamma, \gamma]$: $g(0) = 1/2\gamma$, so $K_c = 4\gamma/\pi$. Here $g''(0) = 0$ and the expansion fails; at $K = K_c$ every $r \le \pi/4$ satisfies (E.6), and the coherence jumps from 0 to $\pi/4$.

Worked example. Take Lorentzian frequencies with half-width γ and coupling $K = 4\gamma$, twice the threshold. The coherence is $r = \sqrt{1 - 2\gamma/K} = 1/\sqrt{2} \approx 0.707$, so the effective coupling felt by each oscillator is $Kr \approx 2.83\gamma$. The locked oscillators are those with $|\omega| \le Kr$, and their fraction is

$$\int_{-Kr}^{Kr} g(\omega)\, d\omega = \frac{2}{\pi} \arctan\frac{Kr}{\gamma} = \frac{2}{\pi}\arctan(2.83) \approx 0.78.$$

About four in five oscillators run at the common frequency. A drifting oscillator with natural frequency $\omega = 4\gamma$ obeys an Adler equation with coupling Kr (Chapter 7), so its time-averaged frequency relative to the center is $\sqrt{\omega^2 - K^2 r^2} \approx 2.83\gamma$, pulled down from 4γ by the population it fails to join. The numbers illustrate how a single relational variable, r, fixes both how many members lock and how strongly the rest are pulled.

Stability. The self-consistency argument finds stationary states; it does not show that they are stable. The linear stability of the incoherent state and of the partially locked branch turned out to be delicate, because the linearized operator has a continuous spectrum, and its resolution occupied a quarter century of work (Strogatz 2000). For the Lorentzian case the reduction of the next section settles the matter directly.

E.5 The Ott–Antonsen Reduction

Ott and Antonsen (2008) showed that for large N the dynamics of certain globally coupled phase oscillators collapses onto a low-dimensional invariant manifold. Expand the density in Fourier modes,

$$\rho(\theta,\omega,t)=\frac{1}{2\pi}\Big[1+\sum_{n=1}^{\infty}\big(\alpha(\omega,t)^{n}\mathrm{e}^{\mathrm{i}n\theta}+\text{c.c.}\big)\Big], \tag{E.9}$$

in which the nth coefficient is the nth power of a single function $\alpha(\omega,t)$ with $|\alpha|<1$. Substituting into (E.3) shows that this family is invariant, provided α obeys

$$\frac{\partial\alpha}{\partial t}+\mathrm{i}\omega\alpha+\frac{K}{2}\big(z\,\alpha^{2}-\bar{z}\big)=0, \qquad \bar{z}=\int\alpha(\omega,t)\,g(\omega)\,\mathrm{d}\omega. \tag{E.10}$$

For the Lorentzian density the integral can be evaluated by closing the contour in the lower half of the complex ω plane, where g has its pole at $\omega=-\mathrm{i}\gamma$; this gives $\bar{z}=\alpha(-\mathrm{i}\gamma,t)$. Evaluating (E.10) at that point yields a closed equation for the order parameter:

$$\dot{z}=-\gamma z+\frac{K}{2}\,z\big(1-|z|^{2}\big), \qquad \dot{r}=-\gamma r+\frac{K}{2}\,r\big(1-r^{2}\big). \tag{E.11}$$

The incoherent state $r=0$ has growth rate $K/2-\gamma$ and loses stability at $K_c=2\gamma$; above threshold the stable fixed point is $r^{*}=\sqrt{1-2\gamma/K}$, which recovers the self-consistent result and proves its stability on the invariant manifold. The same reduction applies to several extensions of the model, including time-delayed coupling (Ott and Antonsen 2008).

E.6 Finite Populations

For finite N the order parameter fluctuates. Below threshold the phases are nearly independent and uniformly distributed, so z is a sum of N unit vectors with random directions divided by N. Its squared modulus has mean $1/N$, and for large N the modulus r follows approximately a Rayleigh distribution with mean

$$\mathbb{E}[r]\approx\sqrt{\frac{\pi}{4N}}\approx\frac{0.886}{\sqrt{N}}. \tag{E.12}$$

For $N=2000$ this is about 0.020, the level of the residual coherence below threshold in Figure E.1. Near threshold, finite-size fluctuations grow and the transition is rounded; how the fluctuations scale with N at criticality is itself a subject of study (Acebrón et al. 2005). Any empirical estimate of coherence from a finite number of units, including the brain-network indices of Chapter 11, has to be compared with the level (E.12) before it is interpreted.

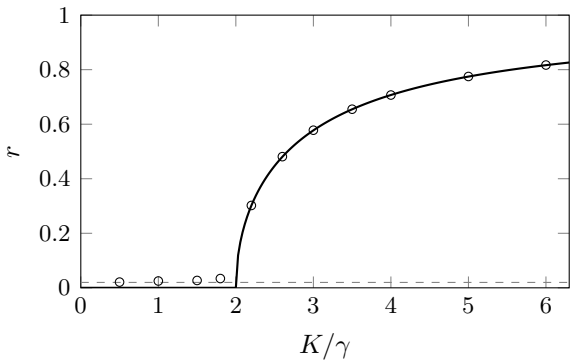


Figure E.1: Coherence against coupling for Lorentzian frequencies: the stable branch of (E.11) (solid), simulations of $N = 2000$ oscillators (circles; the same runs as in Chapter 9), and the finite-size level (E.12) (dashed).

E.7 A Numerical Recipe

The simulations in Chapter 9 used the following procedure, which reproduces the threshold and the branch (E.11) to about three decimal places for $N = 2000$.

1. *Frequencies.* Draw ω_i from g, or, to reduce sampling noise, set them at the quantiles $\omega_i = G^{-1}((i - \frac{1}{2})/N)$ of the cumulative distribution G. Lorentzian tails produce a few very fast oscillators; clipping $|\omega_i|$ at a large value (we used 60γ) keeps the time step manageable and has a negligible effect on $g(0)$.
2. *Integration.* Use the mean-field form (E.2), which costs $O(N)$ per step instead of $O(N^2)$, with a fourth-order Runge–Kutta scheme and a step small compared with the inverse of the largest frequency (we used $0.02/\gamma$ with the clipping above).
3. *Initial conditions.* Draw initial phases uniformly on the circle.
4. *Averaging.* Discard a transient of several relaxation times, which grows near threshold, and average $r(t)$ over the remainder (we discarded the first half of runs of length $150/\gamma$).
5. *Checks.* Confirm that r below threshold matches (E.12), that the estimate of K_c does not move when N is doubled, and that above threshold the fraction of oscillators with $|\omega_i| \le Kr$ matches the fraction whose time-averaged frequency equals the center frequency.

E.8 Fitting the Model to Data

Experiments rarely give the natural frequencies and the coupling directly. Three strategies recover them, and each has a characteristic failure.

The first measures natural frequencies by uncoupling the units: isolating oscillators, cutting the shared medium, or blocking the pathway that couples

them. It fails when uncoupling changes the units themselves, as when a cell's intrinsic rhythm depends on the network it sits in.

The second varies the coupling and fits the curve of coherence against coupling. The threshold gives $g(0)$ through (E.7), and the shape above threshold constrains the rest of g. It fails when the manipulation that varies the coupling also shifts the frequencies, and it needs enough units that finite-size rounding (Section E.6) does not hide the threshold.

The third fits the phase equations directly to recorded phases, estimating each unit's frequency and the coupling function from how phase velocities depend on phase differences. It fails when the units share a common input that is not in the model, because a common drive produces phase relations that look like coupling. The safeguard is an intervention: perturb one unit and check that the others respond as the fitted coupling predicts. The same rule appears in Chapter 1: a relational variable earns a causal role only when changing it, with the relata held as fixed as possible, changes what follows.

E.9 Beyond the Basic Model

The model has been extended in many directions, each of which appears somewhere in this book. On a network the coupling becomes $N^{-1}\sum_j A_{ij}\sin(\theta_j-\theta_i)$, and the symmetries of A constrain which clusters can synchronize (Pecora et al. 2014). A phase lag in the coupling function, or coupling that decays with distance, produces the chimera states of Chapter 12. Time delays and other extensions are surveyed by Acebrón et al. (2005), and the Ott–Antonsen reduction covers several of them (Ott and Antonsen 2008). The reduction from general limit-cycle oscillators to phase equations, which justifies the model's form, is the subject of Appendix D.

E.10 What the Model Contains

The derivations above use three ingredients. Each oscillator is reduced to a phase, so that only its position on its cycle can be influenced. The coupling depends only on phase differences, so that only relations enter the dynamics. And every oscillator responds to the same sum, so that the population's relation to each member is carried by a single complex number. The order parameter is therefore at once a description of the population and the physical input to each member, which is why the self-consistency argument can close. Nothing in the model represents, stores, or evaluates r; the reading that takes place is the transduction of the mean field by each oscillator's phase response, which is the thin sense of reading analyzed in Chapter 15. Systems in which a unit builds its own means of measuring a relation, and acts on the measurement to change the relation, lie outside this model, and Part V begins with one.

Notes and References

Kuramoto (1975) introduced the model and Kuramoto (1984) gave the self-consistent analysis; Winfree (1967) is the precursor. Strogatz (2000) tells the story of the stability problem, and Acebrón et al. (2005) review the model and its extensions. The low-dimensional reduction is due to Ott and Antonsen (2008). Pikovsky et al. (2001) place the model in the wider theory of synchronization.

F

Arnold Tongues

This appendix develops the theory of circle maps used in Chapter 7: lifts and rotation numbers, the conditions that bound the Arnold tongues, their widths for weak coupling, their ordering, the critical line at which they fill the parameter space, what happens beyond it, and a numerical recipe that computes tongues reliably. Phases are measured in cycles, so that the circle is the interval $[0, 1)$ with its ends identified.

F.1 Circle Maps and Their Lifts

A continuous map f of the circle to itself can be described by a *lift*: a continuous function $F : \mathbb{R} \to \mathbb{R}$ with $f(x \bmod 1) = F(x) \bmod 1$. For the maps of interest here the lift satisfies

$$F(x + 1) = F(x) + 1, \tag{F.1}$$

which says that going once around the circle in the input takes the output once around as well (the map has degree one). Different lifts of the same map differ by integers, and we fix the lift by the formula. The map is an orientation-preserving homeomorphism of the circle when F is continuous and strictly increasing; it is a diffeomorphism when, in addition, F is differentiable with $F' > 0$.

Model F.1 (Sine circle map)**.** The lift

$$F_{\Omega,K}(x) = x + \Omega - \frac{K}{2\pi} \sin 2\pi x \tag{F.2}$$

defines the sine circle map. Its derivative is $F' = 1 - K \cos 2\pi x$. For $0 \le K < 1$ the map is a diffeomorphism; at $K = 1$ it is a homeomorphism whose derivative vanishes at $x = 0$; for $K > 1$ it is not invertible (Figure F.1).

Where circle maps come from. An oscillator of natural period T_0, kicked every T seconds, with a phase response curve $\Delta(\theta)$ giving the phase shift produced by a kick at phase θ, obeys $\theta_{n+1} = \theta_n + \Omega + \Delta(\theta_n) \bmod 1$ with

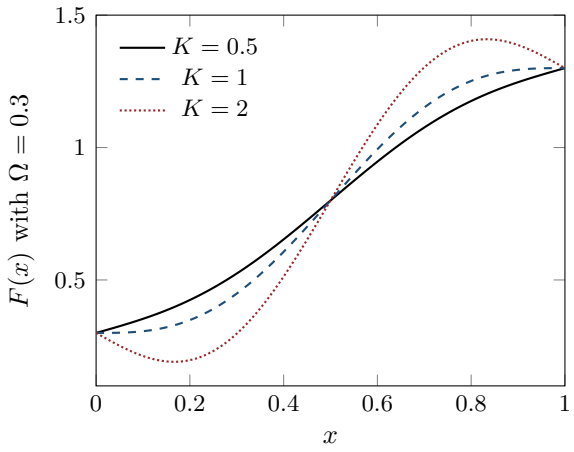


Figure F.1: Lifts of the sine circle map (F.2). For $K < 1$ the lift is strictly increasing; at $K = 1$ it has a single point of zero slope; for $K = 2$ it decreases on part of the interval, so the circle map folds and is no longer invertible.

$\Omega = T/T_0$ (Glass and Mackey 1988). The sine circle map is the case $\Delta(\theta) = -(K/2\pi)\sin 2\pi\theta$. Maps built from measured phase response curves of cardiac cells predicted the locking and period-doubling seen under periodic stimulation (Guevara et al. 1981).

F.2 The Rotation Number

Definition F.2 (Rotation number)**.** For a lift F satisfying (F.1), the rotation number is

$$\rho(F) = \lim_{n\to\infty} \frac{F^n(x) - x}{n}, \tag{F.3}$$

where F^n is the n-fold composition, whenever the limit exists.

Theorem F.3 (Poincaré)**.** *If F lifts an orientation-preserving homeomorphism of the circle, the limit* (F.3) *exists and does not depend on x. The rotation number is rational, $\rho = p/q$ in lowest terms, if and only if there is a point x with $F^q(x) = x + p$, that is, a periodic orbit that goes around the circle p times in q steps.*

The key step in the proof is that for a homeomorphism the quantities $F^n(x) - x$ for different x differ by less than 1, because F^n is increasing and commutes with integer translation; the limit is therefore uniform in x (Arnold 1983). Three properties of ρ as a function of parameters follow for families such as (F.2) with $0 \le K \le 1$:

1. ρ is continuous in (Ω, K);
2. ρ is nondecreasing in Ω, because F increases with Ω for every x;

3. for $K > 0$, each rational value is attained on a closed interval of Ω of positive length, and each irrational value at a single Ω.

The third property is what makes locking robust: a periodic orbit that is transversally attracting survives small changes of Ω, while an irrational rotation number requires the parameter to be tuned exactly.

F.3 Tongue Boundaries

A periodic orbit of type p/q exists exactly when the function $G(x) = F^q(x) - x - p$ has a zero. Because F^q increases with Ω, the set of Ω at which G has a zero is an interval, and its ends are the values at which the zero is born and dies. At an end, G touches zero tangentially:

$$F^q(x) - x - p = 0, \qquad (F^q)'(x) = 1. \tag{F.4}$$

These are the conditions for a saddle-node bifurcation of the periodic orbit, in which a stable and an unstable orbit of type p/q are created or destroyed together. The boundaries of the Arnold tongue are the curves in the (Ω, K) plane on which (F.4) holds.

The $0/1$ tongue. A fixed point of the circle map ($q = 1$, $p = 0$) requires $\Omega = (K/2\pi)\sin 2\pi x$ for some x, which is possible if and only if $|\Omega| \le K/2\pi$. The tongue is the wedge between the lines $\Omega = \pm K/2\pi$, with width K/π, exactly. It is the discrete-time analogue of the Adler locking range $|\nu| \le K$ of Chapter 7.

F.4 Widths for Weak Coupling

For small K, write $\varepsilon = K/2\pi$ and $\Omega = \frac{1}{2} + \delta$ with δ small. Composing the map twice and expanding to second order in ε,

$$F^2(x) - x - 1 \approx 2\delta + 2\pi\varepsilon\delta\cos 2\pi x - \pi\varepsilon^2 \sin 4\pi x.$$

To leading order a zero exists when $2\delta = \pi\varepsilon^2 \sin 4\pi x$ for some x, that is, when $|\delta| \le \pi\varepsilon^2/2$. The $1/2$ tongue therefore has width

$$\Delta\Omega_{1/2} \approx \pi\varepsilon^2 = \frac{K^2}{4\pi} \tag{F.5}$$

for small K. The same calculation for a general p/q requires q compositions, and the first term that can produce a zero is of order ε^q; the width of the p/q tongue scales as K^q as $K \to 0$. Tongues with large denominators are therefore extremely thin at weak coupling, and locking at such ratios is fragile in practice even though it exists in principle.

Example F.4 (Checking the estimate numerically). With the recipe of Section F.9, the 1/2 tongue at $K = 0.52$ has width 0.0211, against $K^2/4\pi = 0.0215$ from (F.5) (our computation). At the same coupling the 1/3 tongue has width 0.0048, the 1/4 tongue 0.0014, and the 1/5 tongue 0.0005, each step up in the denominator dividing the width by a factor between about three and four and a half.

F.5 Farey Ordering

Two fractions $p/q < p'/q'$ are *Farey neighbors* if $p'q - pq' = 1$. Between Farey neighbors, the fraction with the smallest denominator is their mediant $(p+p')/(q+q')$, and every fraction between them can be generated by repeatedly taking mediants. For circle maps the tongues inherit this order: between the tongues of two Farey neighbors, the widest tongue belongs to the mediant (Pikovsky et al. 2001). Starting from 0/1 and 1/1, the hierarchy of tongues is the Farey tree: 1/2 first, then 1/3 and 2/3, then 1/4, 2/5, 3/5, 3/4, and so on. The order makes definite predictions for experiments that scan a drive frequency: the largest locking plateaus appear at the simplest ratios, and between any two observed plateaus the next largest is found at the mediant.

F.6 The Critical Line and the Complete Staircase

As K increases toward 1 the tongues widen, and at $K = 1$ they fill the Ω interval up to a set of measure zero. The rotation number as a function of Ω is then a complete devil's staircase: constant on the countably many locking intervals, which together have full length, and still continuous and increasing from 0 to 1. Jensen et al. (1983) computed the set that the plateaus leave uncovered and found a Cantor set of fractal dimension $D \approx 0.87$, and found the same exponent for a large class of maps, a universality that Bak (1986) describes for a general audience. Two consequences are worth stating. At the critical coupling almost every natural frequency ratio locks to some rational, so a system poised at $K = 1$ is almost never quasiperiodic. And the irrational rotation numbers, though still each realized at one value of Ω, occupy no length at all.

F.7 Beyond the Critical Line

For $K > 1$ the lift is no longer monotone, the circle map folds, and Theorem F.3 no longer applies: the limit (F.3) may depend on the initial point, or fail to exist. Tongues overlap, so that a given (Ω, K) can support two stable periodic orbits of different types, with hysteresis between them as Ω is swept. Inside the tongues, periodic orbits can undergo period-doubling bifurcations, and cascades of them lead to chaotic dynamics (Glass and Mackey 1988). The

periodically stimulated heart-cell aggregates of Guevara et al. (1981) showed period doubling of exactly this kind, which is why a one-dimensional map sufficed to organize their behavior.

F.8 Measuring Tongues in Experiments

In an experiment the tongue is measured, not computed, and the measurement has a standard form. Choose a set of drive frequencies around a candidate ratio and a set of drive strengths. For each combination, estimate the phase of the driven rhythm relative to the drive over many cycles and summarize its concentration by a phase-locking measure (Chapter 9), compared against a null built from drives with jittered timing or from surrogate data (Appendix J). The tongue is the region of the frequency–strength plane in which locking exceeds the null. Notbohm et al. (2016) carried out this design for human alpha rhythms driven by flickering light: locking grew with intensity and was strongest near each participant's intrinsic frequency, and only within that region did rhythmic stimulation differ from jittered stimulation.

A measured tongue can then be compared with the tongue predicted from the rhythm's phase response curve, measured separately with single perturbations, as in the heart-cell work (Guevara et al. 1981). Agreement means that the circle-map description captures the driven dynamics; systematic disagreement means that something beyond a one-dimensional phase is involved, such as changes in the amplitude of the rhythm or slow adaptation. The width of a tongue at fixed drive strength is a direct measure of how strongly a rhythm can be captured, which is why Chapter 60 proposes tracking it across states of consciousness.

F.9 A Numerical Recipe

Rotation numbers. Iterate the lift (F.2) for a transient of a few hundred steps, then for n further steps, and estimate $\rho \approx (F^n(x) - x)/n$. For $K \le 1$ the error is less than $1/n$, because $F^n(x) - x$ differs from $n\rho$ by less than 1. Resolving a plateau of width w requires sampling Ω more finely than w and choosing n large enough that $1/n$ is smaller than the gap between the rational of interest and its neighbors of comparable denominator.

Tongue edges. Rather than detect plateaus from estimated rotation numbers, use the periodic-orbit criterion of Theorem F.3. For a grid of points x_k on $[0,1)$, compute $G_{\max}(\Omega) = \max_k [F^q(x_k) - x_k - p]$ and $G_{\min}(\Omega) = \min_k [F^q(x_k) - x_k - p]$. The left edge of the p/q tongue is the smallest Ω with $G_{\max} \ge 0$, and the right edge the largest Ω with $G_{\min} \le 0$; both functions increase with Ω, so each edge can be found by bisection.

A pitfall. Bisection needs a bracket in which the criterion changes only once. It is tempting to bracket the p/q tongue by a small interval around $\Omega = p/q$, but tongues bend as K grows and their centers leave p/q; at $K = 1$ the tongues with denominators three to five have moved far enough that a narrow bracket misses them and returns nonsense widths. The criterion is monotone in Ω over the whole line, so the safe bracket is wide, for instance $[p/q - \tfrac{1}{2}, p/q + \tfrac{1}{2}]$. The tongues in Chapter 7 were computed this way, with a grid of 4001 points in x and fifty bisection steps per edge.

Notes and References

Arnold (1983) is the classical source for circle maps, rotation numbers, and the tongues named after him. Glass and Mackey (1988) develop circle maps as models of biological rhythms, with the heart-cell experiments of Guevara et al. (1981) as a central example. Pikovsky et al. (2001) treat locking, Farey ordering, and the devil's staircase within the general theory of synchronization. The complete devil's staircase and its universal dimension are due to Jensen et al. (1983); Bak (1982) and Bak (1986) discuss staircases in condensed matter. The continuous-time analogue of the 0/1 tongue is the locking range of R. Adler (1946).

G

Hypergraphs and Higher-Order Interactions

A graph records relations between pairs. Many systems contain interactions among three or more components at once, and some statistical dependencies are irreducibly higher-order: present in a triple and absent from every pair. This appendix collects the structures used to represent such relations (hypergraphs and simplicial complexes), the higher-order Kuramoto model with its mean-field reduction, the reasons higher-order interactions appear in effective descriptions of systems whose physical couplings are pairwise, and the settings in which pairwise models demonstrably fail. Chapter 1 introduced the distinction; Battiston et al. (2020) and Bick et al. (2023) are the standard surveys.

G.1 Graphs and Their Limits

A graph $G = (V, E)$ has vertices $V = \{1, \dots, N\}$ and edges E, each a pair of vertices. Its adjacency matrix has $A_{ij} = 1$ when $\{i, j\} \in E$, the degree of vertex i is $d_i = \sum_j A_{ij}$, and its Laplacian is $L = D - A$ with $D = \mathrm{diag}(d_1, \dots, d_N)$ (Newman 2010). Dynamics on a graph usually take the form

$$\dot{x}_i = f(x_i) + \sum_j A_{ij}\, g(x_i, x_j), \tag{G.1}$$

in which every interaction term depends on the states of two components. Whatever such a model can represent is built from pairwise relations; what it cannot represent directly is a term whose effect on x_i depends jointly on x_j and x_k in a way that no sum of pairwise terms reproduces.

G.2 Hypergraphs

Definition G.1 (Hypergraph)**.** A *hypergraph* $H = (V, E)$ consists of a vertex set V and a collection E of hyperedges, each a subset of V with at least two

elements. H is *k-uniform* if every hyperedge has exactly k elements; a graph is a 2-uniform hypergraph.

The incidence matrix $B \in \{0,1\}^{N\times|E|}$ has $B_{ie} = 1$ when vertex i belongs to hyperedge e. Vertex degrees are row sums, $d_i = \sum_e B_{ie}$, and hyperedge sizes are column sums. A 3-uniform hypergraph can equally be represented by an adjacency tensor with $A_{ijk} = 1$ when $\{i, j, k\} \in E$.

Hyperedges are often projected onto graphs, by joining every pair of vertices that share a hyperedge (the clique expansion), with weights $W = BB^\top - D$ counting shared hyperedges. The projection loses information.

Proposition G.2 (Projection is not injective)**.** *The hypergraph with the single hyperedge* $\{1, 2, 3\}$ *and the graph with edges* $\{1, 2\}$*,* $\{2, 3\}$*,* $\{1, 3\}$ *have the same clique expansion.*

Proof. Both produce $W_{12} = W_{23} = W_{13} = 1$ and no other nonzero weights. □

A three-way interaction and three pairwise interactions are different physical claims, and a representation that cannot tell them apart cannot be used to test which one holds.

G.3 Simplicial Complexes

Definition G.3 (Simplicial complex)**.** A *simplicial complex* K on a vertex set V is a collection of nonempty subsets of V (simplices) that is closed under taking nonempty subsets: every face of a simplex in K is also in K. A simplex with $k + 1$ vertices is a k-simplex; vertices, edges, and filled triangles are 0-, 1-, and 2-simplices.

Downward closure is what distinguishes a simplicial complex from a hypergraph: a filled triangle in a complex brings its three edges with it, while a hyperedge $\{1, 2, 3\}$ in a hypergraph asserts nothing about pairs. The clique complex of a graph fills in every complete subgraph, turning a graph into a simplicial complex; it is a common way to build complexes from pairwise data, and it inherits the projection problem above, because it infers triangles from pairs.

Orienting each simplex defines boundary matrices B_k, which map k-simplices to signed sums of their $(k-1)$-dimensional faces and satisfy $B_k B_{k+1} = 0$. The *Hodge Laplacians*

$$L_k = B_k^\top B_k + B_{k+1} B_{k+1}^\top, \qquad B_0 = 0, \tag{G.2}$$

generalize the graph Laplacian, which is $L_0 = B_1 B_1^\top$. The dimension of the kernel of L_k equals the k-th Betti number of the complex, the number of independent k-dimensional holes, and L_k governs diffusion and synchronization of quantities that live on k-simplices rather than on vertices (Lim 2020).

G.4 Higher-Order Kuramoto Dynamics

The Kuramoto model of Appendix E couples phase oscillators in pairs. Adding a term that couples them in triples gives a model with qualitatively different behavior.

Model G.4 (Kuramoto model with triadic coupling)**.** N phase oscillators with natural frequencies ω_i drawn from a Lorentzian distribution of center ω_0 and half-width Δ obey

$$\dot{\theta}_i = \omega_i + \frac{K_1}{N}\sum_j \sin(\theta_j - \theta_i) + \frac{K_2}{N^2}\sum_{j,l}\sin(2\theta_j - \theta_l - \theta_i), \tag{G.3}$$

where K_1 is the pairwise and K_2 the triadic coupling strength.

Write

$$Z_1 = re^{i\psi} = \frac{1}{N}\sum_j e^{i\theta_j}, \qquad Z_2 = \frac{1}{N}\sum_j e^{2i\theta_j}$$

for the usual order parameter and the second moment. The pairwise and triadic terms can then be written

$$K_1\,\mathrm{Im}\left(Z_1 e^{-i\theta_i}\right) = K_1 r\sin(\psi - \theta_i) \qquad \text{and} \qquad K_2\,\mathrm{Im}\left(Z_2\bar{Z}_1 e^{-i\theta_i}\right).$$

In the limit of many oscillators, the Ott–Antonsen reduction restricts the dynamics to an invariant manifold on which, for Lorentzian frequencies, $Z_2 = Z_1^2$ (Ott and Antonsen 2008). The triadic term then becomes $K_2 r^3\sin(\psi - \theta_i)$, and the model is the ordinary Kuramoto model with an effective coupling $K_1 + K_2 r^2$ that grows with the coherence it produces. The reduction for the ordinary model, $\dot{r} = -\Delta r + \frac{1}{2}Kr(1 - r^2)$, therefore becomes

$$\dot{r} = -\Delta r + \frac{K_1}{2}\,r(1 - r^2) + \frac{K_2}{2}\,r^3(1 - r^2). \tag{G.4}$$

The incoherent state $r = 0$ is stable when $K_1 < 2\Delta$, exactly as in the pairwise model. Nonzero fixed points satisfy, with $u = r^2$,

$$K_2 u^2 + (K_1 - K_2)\,u + (2\Delta - K_1) = 0. \tag{G.5}$$

For $K_2 = 0$ this gives the familiar branch $r = \sqrt{1 - 2\Delta/K_1}$, which grows continuously from zero at $K_1 = 2\Delta$. For large enough K_2, (G.5) has two positive roots when $K_1 < 2\Delta$: a stable partially synchronized state coexists with the stable incoherent state, separated by an unstable branch. The system is bistable, synchronizes abruptly when K_1 crosses 2Δ from below, and desynchronizes abruptly, at a lower coupling, when K_1 is decreased (Figure G.1).

Models of this family have been analyzed in detail. Three-body phase interactions produce infinitely many multistable synchronized states above

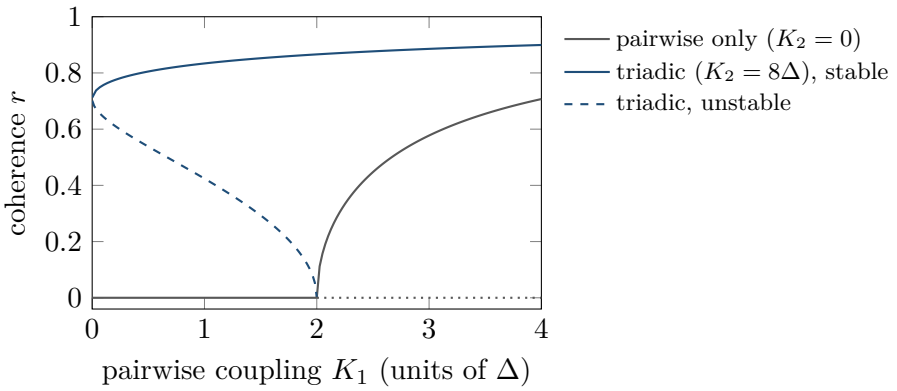


Figure G.1: Fixed points of (G.4) with $\Delta = 1$. With pairwise coupling alone, coherence grows continuously from $K_1 = 2$. With triadic coupling $K_2 = 8$, a synchronized branch (solid blue) coexists with the incoherent state for $0 < K_1 < 2$, separated by an unstable branch (dashed); the incoherent state (gray) loses stability at $K_1 = 2$ (dotted beyond). The resulting hysteresis is absent from the pairwise model.

a critical coupling while the incoherent state remains stable (Tanaka and Aoyagi 2011); globally coupled oscillator simplexes show a continuum of abrupt desynchronization transitions and extensive multistability (Skardal and Arenas 2019); on networks, three- and four-way interactions induce abrupt synchronization through hysteresis and bistability, and can stabilize strong synchrony even when the pairwise coupling is repulsive (Skardal and Arenas 2020). When oscillators live on links and triangles of a simplicial complex, rather than only on vertices, the dynamics are governed by the Hodge Laplacians of (G.2), and adaptive couplings can make the synchronization transition explosive (Millán et al. 2020).

G.5 Where Higher-Order Interactions Come From

Higher-order terms can arise even when every physical coupling is pairwise. Phase reduction, which turns weakly coupled limit-cycle oscillators into phase oscillators (Appendix D), is exact to first order in the coupling strength, and at first order the phase equations are pairwise. At the next order they are not. For identical oscillators near a Hopf bifurcation with full permutation symmetry, the second-order phase equations generically contain terms with up to four interacting phases (Ashwin and Rodrigues 2016), and a second- and third-order reduction of the mean-field complex Ginzburg–Landau equation shows each new order contributing additional nonpairwise interactions, which account for dynamics that the first-order Kuramoto model cannot produce (León and Pazó 2019).

Two consequences follow. First, the distinction between higher-order *structure*, the physical interactions of a system, and higher-order *dynamics*, the

terms in an effective description, matters; a model with triadic terms is not evidence of triadic physical interactions (Bick et al. 2023). Second, eliminating unobserved variables, whether fast modes, hidden nodes, or amplitude degrees of freedom, generally produces effective interactions among the remaining variables that are not pairwise. An apparent three-way interaction among recorded neurons may reflect a pairwise circuit with an unrecorded participant.

G.6 When Pairwise Models Fail

Pairwise models fail demonstrably in at least four settings.

Irreducible statistical dependence. For independent fair bits X, Y and $Z = X \oplus Y$, every pairwise mutual information vanishes while $I(X, Y; Z) = 1$ bit (Chapter 1). No graph weighted by pairwise statistics represents the dependence. McGill's interaction information registers it (McGill 1954); partial information decomposition names it synergy (Williams and Beer 2010); Timme et al. (2014) compare the measures, and synergy-based definitions of emergence build on them (Rosas et al. 2020). Appendix H develops the measures.

Group contagion. When social contagion can occur through groups as well as pairs, modeled on a simplicial complex, the transition to an endemic state becomes discontinuous and a bistable region appears in which healthy and endemic states coexist, which pairwise models do not produce (Iacopini et al. 2019).

Ecological coexistence. Simple models of pairwise competition cannot sustain very large communities of competitors; adding higher-order interactions, in which one species modifies the interaction between two others, stabilizes the dynamics and makes coexistence robust to perturbations of abundances and parameters (Grilli et al. 2017).

Synchronization. As Section G.4 showed, triadic coupling produces bistability and hysteresis that the pairwise Kuramoto model lacks.

Detecting higher-order interactions in data is harder than defining them. A higher-order model will always fit at least as well as the pairwise model it contains, so comparisons need penalties for complexity or held-out data; finite samples inflate estimates of synergy; and, as the previous section showed, hidden variables produce effective higher-order terms. The most direct test is interventional: manipulate one component while holding a second fixed, and ask whether the effect on a third depends on the second's state (Chapter 57).

G.7 A Higher-Order Relation in the Fish

The electric fish's sign computation has exactly the structure of the exclusive-or. In the two-patch model of Chapter 21, with patch 2 the more contaminated, the fish's inference is $\operatorname{sgn} \Delta f = -\operatorname{sgn}(\dot{A}\,\delta)$, where $\dot{A}$ is the rate of change of amplitude at either patch and $\delta = \varphi_2 - \varphi_1$ the differential phase. Over a beat

cycle, $\operatorname{sgn}\dot{A}$ equals $-\operatorname{sgn}\Delta f \cdot \operatorname{sgn}(\sin\psi)$ and $\operatorname{sgn}\delta$ equals $\operatorname{sgn}(\sin\psi)$. If the sign of Δf is equally likely to be positive or negative and the beat phase ψ is uniform, each of $\operatorname{sgn}\dot{A}$ and $\operatorname{sgn}\delta$ is a fair coin independent of $\operatorname{sgn}\Delta f$, yet together they determine it. In the binary encoding of signs, the fish computes an exclusive-or: all pairwise informations between the three signs vanish, and the triple carries one bit.

The observation restates, in information-theoretic terms, the result of Chapter 21 that neither the amplitude nor the phase modulation carries the sign by itself while their joint behavior carries it exactly. It also explains why the reader must combine two sensory streams at a site that receives both, and why the relevant midbrain neurons are combination-sensitive, responding to particular conjunctions of amplitude and phase change (Heiligenberg and G. Rose 1986; Carlson and Kawasaki 2004): a reader that monitored the streams separately, or their pairwise correlations with anything else, would be blind to the variable the fish needs.

G.8 Notes and References

Battiston et al. (2020) is the comprehensive review of networks beyond pairwise interactions; Bick et al. (2023) give a careful account of what "higher-order network" can mean and of the difference between higher-order structure and higher-order dynamics. Newman (2010) covers graphs; Lim (2020) gives an accessible treatment of Hodge Laplacians. The Ott–Antonsen reduction is Ott and Antonsen (2008). Higher-order Kuramoto models are analyzed by Tanaka and Aoyagi (2011), Skardal and Arenas (2019), Skardal and Arenas (2020), and, for oscillators on simplices, Millán et al. (2020). The origin of nonpairwise phase interactions in phase reduction is shown by Ashwin and Rodrigues (2016) and León and Pazó (2019). Higher-order effects in contagion and ecology are reported by Iacopini et al. (2019) and Grilli et al. (2017). The information-theoretic treatment of irreducible dependence runs from McGill (1954) through Williams and Beer (2010) and Timme et al. (2014) to Rosas et al. (2020).

G.9 Exercises

Exercise G.1. For the complex consisting of a filled triangle $\{1,2,3\}$ with its edges and vertices, write the boundary matrices B_1 and B_2 for a chosen orientation, verify $B_1 B_2 = 0$, and compute L_0, L_1, and L_2. Remove the filled triangle, keeping its edges, and show that L_1 acquires a one-dimensional kernel. Interpret the result.

Exercise G.2. Derive (G.4) from (G.3) assuming $Z_2 = Z_1^2$. For $\Delta = 1$, find the smallest K_2 for which bistability exists at some $K_1 < 2$, and the value of K_1 at which the synchronized branch disappears for $K_2 = 8$.

Exercise G.3. Simulate (G.3) with $N = 2000$ oscillators, $\Delta = 1$, and $K_2 = 8$, sweeping K_1 slowly up from 0 to 4 and back down. Plot r against K_1 for both sweeps and compare with Figure G.1.

Exercise G.4. Using the small-contamination expressions of Chapter 21, verify that $\operatorname{sgn} \dot{A}$, $\operatorname{sgn} \delta$, and $\operatorname{sgn} \Delta f$ are pairwise independent when $\operatorname{sgn} \Delta f$ is equally likely to be positive or negative and ψ is uniform, and compute the interaction information of the triple. Does the result change for a fish whose two patches have equal contamination?

H

Information Decomposition

This appendix collects the information-theoretic measures the book uses to quantify relations among variables, and in particular relations that no pair of variables carries on its own. The basic quantities of Section H.1 are settled mathematics. The decompositions of Sections H.4 and H.8 are younger, and their central definitions are still disputed; we mark where. Logarithms are base 2, so information is measured in bits.

H.1 Entropy and Mutual Information

For a discrete random variable X with distribution $p(x)$, the *entropy* is

$$H(X) = -\sum_{x} p(x) \log p(x), \tag{H.1}$$

the average uncertainty about X (Shannon 1948). The conditional entropy $H(X \mid Y) = -\sum_{x,y} p(x,y) \log p(x \mid y)$ is the uncertainty that remains about X once Y is known. The *mutual information*

$$I(X;Y) = H(X) - H(X \mid Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \tag{H.2}$$

is the reduction in uncertainty about either variable that knowing the other provides. It is symmetric, non-negative, and zero exactly when X and Y are independent; it equals the Kullback–Leibler divergence between the joint distribution and the product of the marginals (Cover and Thomas 2006).

Conditional mutual information $I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z)$ obeys the *chain rule*

$$I(X;Y,Z) = I(X;Y) + I(X;Z \mid Y). \tag{H.3}$$

If $W \to X \to Z$ is a Markov chain, the chain rule gives the *data-processing inequality*, $I(Z;W) \le I(Z;X)$: processing cannot create information about a source. Chapter 29 uses the inequality to show that information alone favors proximal over distal content.

H.2 Interaction Information

With three variables, the mutual information between two of them can go up or down when the third is known. McGill introduced a measure of this effect, the interaction information (McGill 1954). We use the symmetric form known as *co-information*,

$$C(X;Y;Z) = I(X;Y) - I(X;Y \mid Z), \tag{H.4}$$

which is unchanged under any permutation of X, Y, Z. Authors differ in sign convention, and the reader should check which convention a paper uses before comparing values (Timme et al. 2014).

Co-information can be positive or negative. A positive value means that conditioning on Z removes some of the dependence between X and Y, as happens when all three share a common source. A negative value means that conditioning on Z creates dependence, as happens when Z is jointly determined by X and Y. The measure has a defect that motivates the next section: a single number cannot distinguish a situation with no shared structure from one with equal amounts of the two opposite kinds.

H.3 A Worked Example: Exclusive Or

Let X and Y be independent fair bits and $Z = X \oplus Y$. The four equally likely outcomes are listed in Table H.1, alongside two contrasting systems on the same three bits: a redundant one, in which Z copies a single bit shared by all three, and a system in which Z copies X alone.

System	$I(X;Z)$	$I(Y;Z)$	$I(X,Y;Z)$	$I(X;Y)$	$I(X;Y \mid Z)$	C
Exclusive or	0	0	1	0	1	-1
Redundant	1	1	1	1	0	$+1$
Unique	1	0	1	0	0	0

Table H.1: Information quantities, in bits, for three systems of fair bits: exclusive or ($Z = X \oplus Y$), redundant ($X = Y = Z$), and unique ($Z = X$ with Y independent); $C = C(X;Y;Z)$. In the exclusive or, no single variable tells anything about Z, and the pair tells everything.

For the exclusive or, every pair of variables is independent, yet any two determine the third. The dependence is irreducibly three-way, and a reader that monitors pairs of variables cannot detect it, which is the point of Chapter 1. The negative co-information registers the effect. In the redundant system the co-information is positive, and in the unique system it is zero.

H.4 Partial Information Decomposition

Williams and Beer proposed to split the information that two *sources* S_1, S_2 carry about a *target* T into four non-negative parts (Williams and Beer 2010):

$$\begin{aligned} I(S_1, S_2; T) &= \mathrm{Red} + \mathrm{Unq}_1 + \mathrm{Unq}_2 + \mathrm{Syn}, \\ I(S_1; T) &= \mathrm{Red} + \mathrm{Unq}_1, \qquad I(S_2; T) = \mathrm{Red} + \mathrm{Unq}_2. \end{aligned} \tag{H.5}$$

Redundancy is information that either source provides; *unique* information is provided by one source and not the other; *synergy* is provided only by the two together. The three equations constrain four unknowns, so one more quantity must be defined. Once redundancy is fixed, the rest follows.

Two consequences hold for any choice of redundancy measure. First, subtracting the three equations shows that co-information is redundancy minus synergy:

$$C(S_1; S_2; T) = I(S_1; T) + I(S_2; T) - I(S_1, S_2; T) = \mathrm{Red} - \mathrm{Syn}, \tag{H.6}$$

using $C(S_1; S_2; T) = I(S_1; T) - I(S_1; T \mid S_2)$ and the chain rule. This is the defect noted above: zero co-information can hide equal redundancy and synergy. Second, for the exclusive or, $I(S_i; T) = 0$ forces $\mathrm{Red} = \mathrm{Unq}_i = 0$, so the whole bit is synergy. That verdict does not depend on how redundancy is measured.

The decomposition generalizes to more sources through a lattice of collections of sources, ordered by which redundancies contain which. Figure H.1 shows the two-source case.

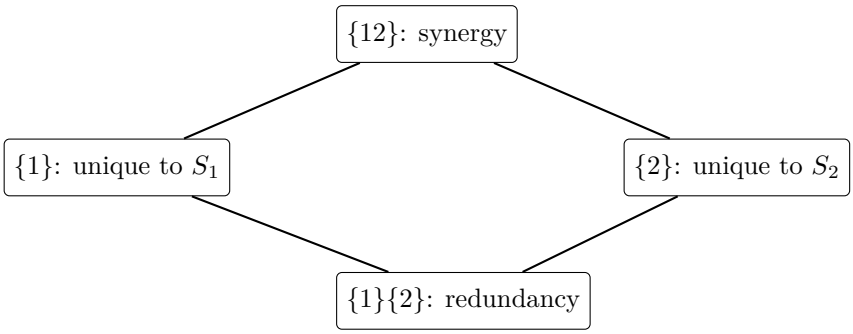


Figure H.1: The redundancy lattice for two sources (Williams and Beer 2010). Each node is an information atom; the atoms at or below a node sum to the information that node's collection of sources provides.

Measuring redundancy. Williams and Beer's own measure, $I_{\min}$, defines redundancy through the *specific information* each source provides about each

outcome of the target,

$$I(T{=}t; S_i) = \sum_s p(s \mid t) \log \frac{p(t \mid s)}{p(t)},$$
$$\mathrm{Red} = \sum_t p(t)\ \min_i I(T{=}t; S_i). \tag{H.7}$$

Consider the *two-bit copy*, $T = (X, Y)$ for independent fair bits X, Y. Each source has specific information of one bit about every outcome, so $I_{\min}$ assigns one bit of redundancy, no unique information, and one bit of synergy. Yet X and Y are independent, and it seems that each contributes one unique bit. Harder, Salge, and Polani traced the problem to the fact that $I_{\min}$ registers when sources provide the same *amount* of information about an outcome rather than the same information, and proposed a different measure (Harder et al. 2013). Several further measures have since been proposed, and they disagree on examples of this kind (Timme et al. 2014); there is no consensus redundancy measure. Any quantitative claim about synergy should state which measure it uses, and claims that hold for every measure (such as the exclusive-or verdict above) are the ones to rely on.

H.5 Synergy and Relations

Synergy is the information-theoretic form of an irreducibly higher-order relation. Timme and colleagues compared the measures of synergy and redundancy in use in neuroscience and showed that they can disagree in sign and size on the same data, so that "synergy" in one paper need not mean what it means in another (Timme et al. 2014). The distinction between pairwise and higher-order relations in Chapter 1 and Appendix G has a statistical counterpart here: a relation that exists only as synergy is invisible to any analysis restricted to pairs, whether the analysis is a matrix of correlations or a reader that combines its inputs two at a time.

H.6 Transfer Entropy and Granger Causality

For time series, the question is often directional: does the past of Y tell us about the future of X beyond what the past of X already tells us? Schreiber's *transfer entropy* answers it with a conditional mutual information (Schreiber 2000),

$$T_{Y \to X} = I(X_{t+1}; Y_t^{(l)} \mid X_t^{(k)}), \tag{H.8}$$

where $X_t^{(k)}$ and $Y_t^{(l)}$ are the last k and l values of each series. Granger's earlier criterion asks the same question with linear prediction: Y Granger-causes X if including the past of Y reduces the error of predicting X (Granger 1969). For

jointly Gaussian variables the two are equivalent up to a factor of two (Barnett et al. 2009).

Both measure directed *predictive* dependence. Neither is a measure of causation in the interventional sense of Appendix I: a common driver, an unobserved variable, or a mismatch of time scales can produce transfer entropy between variables that do not influence each other (Pearl 2009).

H.7 Information in Time: The Anatomy of a Bit

A single observation X_0 of a stationary process can be decomposed according to its relations with the past $X_{:0}$ and the future $X_{1:}$ (R. G. James et al. 2011). Two applications of the chain rule give

$$\begin{aligned} H[X_0] &= \underbrace{I[X_{:0}; X_0]}_{\rho_\mu} + \underbrace{H[X_0 \mid X_{:0}]}_{h_\mu}, \\ h_\mu &= \underbrace{I[X_0; X_{1:} \mid X_{:0}]}_{b_\mu} + \underbrace{H[X_0 \mid X_{:0}, X_{1:}]}_{r_\mu}. \end{aligned} \tag{H.9}$$

The first term, ρ_μ, is the part of the present that the past already predicts. The entropy rate h_μ is the genuinely new part, and it splits into *bound* information b_μ, new information that the process carries into its future, and *ephemeral* information r_μ, new information that is forgotten at once. The decomposition is useful for the book's question about detachment (Chapter 31): it separates what a reader could gain by remembering, which is bounded by ρ_μ and b_μ, from what no reader can use beyond the present moment, which is r_μ.

H.8 Integrated Information Decomposition

Mediano and colleagues combined the partial information decomposition with the analysis of dynamics (Mediano et al. 2021). For a system of two components, the information that the joint past carries about the joint future, $I(X_t, Y_t; X_{t+\tau}, Y_{t+\tau})$, is decomposed into atoms indexed by pairs of partial-information atoms, one for the sources in the past and one for the targets in the future, which gives sixteen atoms for two components. The atoms separate kinds of information dynamics that a single number conflates: information stored in one component, copied from one component to the other, transferred, or carried synergistically by the whole. The framework underlies one information-theoretic criterion for causal emergence (Chapter 14) (Rosas et al. 2020). It inherits the unresolved choice of redundancy measure from Section H.4, and its quantitative results depend on that choice.

H.9 Estimation

All of the quantities above are defined for known distributions and must be estimated from finite data. Plug-in estimates of entropy are biased downward and plug-in estimates of mutual information upward when samples are few relative to the number of states, and the bias grows with the number of variables, which makes higher-order quantities such as synergy the hardest to estimate. Every estimate should therefore be compared against a null distribution obtained from surrogate data that preserve the properties the claim does not concern (Appendix J) (Theiler et al. 1992; Schreiber and Schmitz 2000). A synergy of 0.05 bits means nothing until it is shown to exceed what shuffled data produce.

H.10 Notes and References

Shannon (1948) founded the subject; Cover and Thomas (2006) is the standard textbook. Interaction information is due to McGill (1954). The partial information decomposition was introduced by Williams and Beer (2010); Harder et al. (2013) gave an influential critique of its original redundancy measure and an alternative; Timme et al. (2014) compare the measures of synergy and redundancy used in neuroscience. Transfer entropy is due to Schreiber (2000); Granger causality to Granger (1969); their Gaussian equivalence to Barnett et al. (2009). The decomposition of a single observation is from R. G. James et al. (2011). Integrated information decomposition is developed in Mediano et al. (2021), and its use in defining causal emergence in Rosas et al. (2020). For surrogate-data methods, see Theiler et al. (1992) and Schreiber and Schmitz (2000).

I

Causal Inference and Interventions

Two of the criteria for a useful relational variable in Chapter 1 are causal: a relation earns its place when interventions that change it change outcomes, and when it can be changed while its relata are held as fixed as physically possible. Part XIII is organized around interventions on relations, on readers, and on loops (Chapters 57 to 59). This appendix collects the tools those chapters use: structural causal models and the intervention operator, interventions on relations, the interventionist account of causation, potential outcomes, Granger causality and transfer entropy, and causal discovery in systems with feedback.

I.1 Structural Causal Models

Definition I.1 (Structural causal model)**.** A *structural causal model* consists of endogenous variables $V_1, \dots, V_n$, exogenous variables $U_1, \dots, U_n$ with a joint distribution, and structural assignments

$$V_i := f_i(\mathrm{PA}_i, U_i), \qquad i = 1, \dots, n, \tag{I.1}$$

where the parents PA_i are a subset of the other endogenous variables. The *causal graph* has an arrow $V_j \to V_i$ whenever $V_j \in \mathrm{PA}_i$. When the graph is acyclic and the U_i are independent, the model is *Markovian* and the joint distribution factorizes as $P(v) = \prod_i P(v_i \mid \mathrm{pa}_i)$ (Pearl 2009).

The assignment symbol $:=$ is asymmetric on purpose. An equation can be rearranged; an assignment says which variable is set by which, and an intervention replaces one of them.

Definition I.2 (Intervention)**.** The intervention $\mathrm{do}(X{=}x)$ on a variable X replaces the assignment for X by the constant assignment $X := x$ and leaves all other assignments unchanged. In a Markovian model the resulting distribution

is given by the truncated factorization

$$P(v \mid \mathrm{do}(X{=}x)) = \prod_{i:\, V_i \neq X} P(v_i \mid \mathrm{pa}_i)\Big|_{X=x} \tag{I.2}$$

for every v consistent with $X = x$, and zero otherwise (Pearl 2009).

The difference between seeing and doing is the difference between $P(y \mid x)$ and $P(y \mid \mathrm{do}(x))$, and a confounder is enough to make it large.

Example I.3 (Confounding). Let a binary Z with $\mathbb{P}(Z{=}1) = 0.5$ cause both X and Y, with $\mathbb{P}(X{=}1 \mid Z{=}1) = 0.9$, $\mathbb{P}(X{=}1 \mid Z{=}0) = 0.1$, and $\mathbb{P}(Y{=}1 \mid X, Z) = 0.2 + 0.1X + 0.6Z$. Intervening,

$$\mathbb{P}(Y{=}1 \mid \mathrm{do}(X{=}1)) - \mathbb{P}(Y{=}1 \mid \mathrm{do}(X{=}0)) = 0.1,$$

which is the coefficient of X in the assignment for Y. Observing, $\mathbb{P}(Z{=}1 \mid X{=}1) = 0.9$ and $\mathbb{P}(Z{=}1 \mid X{=}0) = 0.1$, so

$$\mathbb{P}(Y{=}1 \mid X{=}1) - \mathbb{P}(Y{=}1 \mid X{=}0) = 0.84 - 0.26 = 0.58.$$

Most of the observed association is carried by Z.

When the confounders are measured, the interventional distribution can be computed from observational data.

Theorem I.4 (Back-door adjustment)**.** *A set of variables S satisfies the* back-door criterion *relative to (X, Y) if no member of S is a descendant of X and S blocks every path between X and Y that contains an arrow into X. If S satisfies the criterion, then*

$$P(y \mid \mathrm{do}(X{=}x)) = \sum_{s} P(y \mid x, s)\, P(s). \tag{I.3}$$

The proof, and the more general rules of the do-calculus for cases the back-door criterion does not cover, are in Pearl (2009). In Example I.3, $S = \{Z\}$ satisfies the criterion and (I.3) recovers the interventional difference of 0.1.

I.2 Interventions on Relations

A relation $r = \mathcal{R}(x_1, x_2)$ is a function of variables, not a variable with its own structural assignment. No intervention sets r directly: every intervention that changes r changes at least one relatum. The claim that r, and not its relata, is causally relevant must therefore be made relative to a specification of what is held fixed.

Definition I.5 (Relational intervention relative to held variables)**.** Let $(x_1, x_2) \mapsto (r, s)$ be an invertible change of variables, where $r = \mathcal{R}(x_1, x_2)$ is the relation of interest and s collects the constituent variables to be held fixed. An *intervention on r relative to s* sets $r := r_0$ while holding s at its pre-intervention value; the new values of the relata are those that the inverse change of variables assigns to (r_0, s).

The choice of s is part of the causal claim. For two oscillators with phases θ_1, θ_2, the relative phase $r = \theta_1 - \theta_2$ can be shifted while holding θ_2 fixed, while holding θ_1 fixed, or while holding the mean phase fixed, and in general these are three different interventions. Symmetry removes the ambiguity where it applies.

Proposition I.6 (Invariance makes relational interventions well defined)**.** *Suppose an outcome Y depends on the relata only through quantities that are invariant under a common shift of both phases, as in any model with the symmetry of Chapter 1. Then every intervention that sets $\theta_1 - \theta_2 = r_0$ and leaves all other invariant quantities unchanged has the same effect on Y, whichever combination of the phases it holds fixed.*

Proof. Two such interventions produce configurations that differ by a common shift of both phases. By hypothesis Y is a function of shift-invariant quantities only, so it takes the same value on both configurations. □

When no available manipulation changes r without also changing a constituent that affects Y, the intervention is *fat-handed*, and the separate contributions of relation and relata cannot be identified from that manipulation alone. The experimental designs of Part XIII are attempts to avoid fat-handed interventions: the relation clamp of Chapter 59, which fixes the fish's frequency difference while leaving its own frequency free, and phase-specific stimulation at fixed local power (Chapter 61).

I.3 The Interventionist Account

Woodward's account makes interventions the content of causal claims rather than only a way of testing them (Woodward 2003). Relative to a set of variables, X is a cause of Y if some possible intervention on X would change Y, or its probability distribution, when the remaining variables are held fixed at suitable values. An intervention I on X with respect to Y must satisfy four conditions: I causes X; I acts as a switch, so that the other causes of X no longer determine it; any route from I to Y goes through X; and I is statistically independent of every cause of Y that does not lie on the route through X. A generalization is explanatory to the extent that it is *invariant*, that is, continues to hold under some range of interventions.

The account suits this book for two reasons. It makes causal claims relative to a choice of variables, which is exactly the situation of a relation and its relata (Definition I.5). And it ties explanation to invariance under intervention, which is the fifth criterion of Chapter 1 restated: a relational variable explains when its role survives the interventions we can make.

I.4 Potential Outcomes

The potential-outcomes framework states causal effects at the level of units (Imbens and Rubin 2015). Each unit i has an outcome $Y_i(1)$ under treatment and $Y_i(0)$ under control; the unit-level effect $Y_i(1) - Y_i(0)$ is never observed, because each unit receives only one condition. The average treatment effect $\tau = \mathbb{E}[Y(1) - Y(0)]$ can be estimated when assignment is independent of the potential outcomes, as randomization ensures, in which case $\tau = \mathbb{E}[Y \mid T{=}1] - \mathbb{E}[Y \mid T{=}0]$. The estimate also requires the stable unit treatment value assumption: no interference between units and no hidden versions of the treatment.

Yoked designs, such as Held and Hein's carousel (Chapter 59), are matched-pair designs in this framework. The treatment is loop status, and the pairing equates a covariate (visual input) by construction. They also illustrate why the stable unit assumption must be examined in each design: in the carousel the two units are mechanically coupled, since the active kitten's walking produces the passive kitten's ride, so the passive unit's condition is defined through the active unit's behavior. That coupling is intended, and it means the comparison estimates the effect of loop status given matched visual flow, not the effect of treatment in isolated units.

I.5 Granger Causality and Transfer Entropy

For time series, prediction offers a weaker substitute for intervention. Granger proposed that X causes Y in his sense if the past of X improves the prediction of Y beyond what the past of Y, and of the other variables included, already provides (Granger 1969). In a linear autoregressive model the magnitude can be written as

$$F_{X \to Y} = \ln \frac{\sigma^2_{\text{restricted}}}{\sigma^2_{\text{full}}}, \tag{I.4}$$

the log ratio of the prediction-error variance of Y without and with the past of X.

Transfer entropy generalizes the idea beyond linear models (Schreiber 2000):

$$T_{X \to Y} = I\big(Y_{t+1}; X_t^{(l)} \,\big|\, Y_t^{(k)}\big), \tag{I.5}$$

the information that the last l values of X carry about the next value of Y beyond the last k values of Y. Conditioning on the past of Y excludes

information shared merely through common history, which lets the measure distinguish driving from responding elements and detect asymmetric interactions. For jointly Gaussian variables the two measures are equivalent (Barnett et al. 2009): the conditional mutual information in (I.5) then equals half the log ratio of the conditional variances of Y_{t+1} without and with the past of X, so that $F_{X\to Y} = 2\,T_{X\to Y}$ when both are computed with natural logarithms. Appendix H places transfer entropy among the other information-theoretic decompositions.

Both measures are predictive, and their limits follow from that. A common driver omitted from the model produces Granger causality between the variables it drives. When sensors record instantaneous mixtures of sources, as scalp electrodes do (Appendix K), the past of one sensor carries information about sources that also drive another, so directed predictability between sensors does not establish interaction between the underlying sources. Filtering and coarse sampling can create or remove apparent directed influence. And in a closed loop both directions of influence are present by construction; the measures can apportion predictability between the directions, but they do not identify the mechanisms that carry it. None of these limits applies to a well-designed intervention, which is why Part XIII prefers interventions wherever they are possible.

I.6 Causal Discovery and Feedback

Causal discovery infers graphs from data. Under the causal Markov condition (each variable is independent of its non-descendants given its parents) and faithfulness (every independence in the data follows from the graph), the pattern of conditional independences in observational data determines the causal graph up to a class of graphs with the same independences; constraint-based procedures such as the PC algorithm recover that class, and extensions handle unmeasured confounders (Spirtes et al. 2000).

Reading loops are cycles, and acyclic methods seem to exclude them. Time resolves the conflict whenever every influence takes time. Indexing each variable by time turns the loop $s_{\mathrm{A}} \to y \to z \to a \to s_{\mathrm{A}}$ into an acyclic graph over time slices, in which a_t affects $s_{\mathrm{A},t+1}$ and so on; the standard methods then apply to the unrolled graph. Influences faster than the sampling interval break the construction and require either finer sampling or models that allow instantaneous cycles.

Example I.7 (Unrolling the jamming loop). Write the fish's loop in discrete time with variables for the neighbor's frequency $f_{\mathrm{e},t}$, the fish's frequency $f_{\mathrm{s},t}$, the frequency difference $\Delta f_t = f_{\mathrm{e},t} - f_{\mathrm{s},t}$, the sign read by the midbrain z_t, and the pacemaker command a_t. The unrolled graph contains $\Delta f_t \to z_t \to a_t \to f_{\mathrm{s},t+1} \to \Delta f_{t+1}$ and $f_{\mathrm{e},t+1} \to \Delta f_{t+1}$, and it is acyclic. The experimental conditions of Chapter 59 become interventions on this graph. A fixed stimulus

is $\mathrm{do}(f_{\mathrm{e},t} = c)$ for all t. Proportional feedback replaces the assignment for $f_{\mathrm{e},t}$ with one that depends on $f_{\mathrm{s},t}$, adding arrows $f_{\mathrm{s},t} \to f_{\mathrm{e},t}$. A relation clamp sets Δf_t itself for all t, cutting the arrows from both frequencies into it; it is the intervention on the relation relative to the fish's own frequency in the sense of Definition I.5.

Closed loops raise one more difficulty, which the exercises of Chapter 59 work through in a scalar example; here it is stated in general form. In a feedback loop the input to any component is correlated with the noise that entered its own output earlier, so regressing output on input gives a biased estimate of the component's dynamics. An external signal injected into the loop and independent of its internal noise serves as an instrument: the ratio of the output's covariance with the injected signal to the input's covariance with it recovers the component's gain. This is how experiments that close a loop around an animal can still identify the animal's own dynamics.

I.7 Notes and References

Pearl (2009) develops structural causal models, the intervention operator, the back-door criterion, and the do-calculus. Woodward (2003) gives the interventionist account of causation and explanation used throughout Part XIII. The potential-outcomes framework is presented systematically by Imbens and Rubin (2015). Granger causality originates in Granger (1969), transfer entropy in Schreiber (2000), and their equivalence for Gaussian variables is shown by Barnett et al. (2009). Constraint-based causal discovery, including the PC algorithm and its extensions to latent confounders, is developed in Spirtes et al. (2000).

J

Experimental Surrogate Methods

A relational measure computed on data is almost never zero. Two independent noise signals have a nonzero sample phase-locking value; two signals with similar power spectra have nonzero sample coherence; a filtered signal with sharp transients shows phase–amplitude coupling where none exists in the underlying process. To claim that a relation is present, one must compare the measured value with the values the same measure takes on data that share every property of the real data except the relation in question. Such data are called *surrogates*. This appendix sets out the logic of surrogate testing, the standard constructions, and the pairing of each relational measure used in this book with a null that matches it. Part XIII relies on it throughout, and Chapters 60 and 62 make its central rule a condition of the research programme.

J.1 The Logic of a Surrogate Test

A surrogate test specifies a null hypothesis by the properties it preserves. Theiler et al. (1992) introduced the method for testing nonlinearity in time series: generate many synthetic series that share with the data the properties allowed under the null (for instance, the power spectrum and the distribution of values) and are otherwise random, compute a test statistic T on the data and on each surrogate, and reject the null if the data's value is extreme among the surrogates' values.

With N_s surrogates and a one-sided test, a rank-based p-value is

$$p = \frac{1 + \#\{k : T_k \geq T_{\text{data}}\}}{N_s + 1}, \tag{J.1}$$

which is valid whenever the data and the surrogates are exchangeable under the null. The smallest attainable p is $1/(N_s + 1)$, so a test at level α requires at least $N_s \geq 1/\alpha - 1$ surrogates. Estimating the full null distribution, or correcting for many tests, requires many more.

Two errors bracket every choice of surrogate. A surrogate that preserves too little tests the wrong null: shuffling the samples of an autocorrelated signal destroys its autocorrelation, so almost any temporal statistic will reject the shuffle null, and the rejection says nothing about the relation of interest. A surrogate that preserves too much has no power: if it preserves the relation, nothing can reject it. The art is to preserve exactly the properties that could produce the measured value without the relation, and to destroy the relation.

Principle J.1 (A matched null for every relational measure)**.** A relational measure is reported with (i) its value on the data, (ii) the distribution of its values on surrogates that preserve every property of the constituent signals that could produce a nonzero value without the relation under test, and (iii) the performance of the constituent measures it is built from, so that a relational measure is credited only with what the constituents do not already explain.

Item (iii) is the rule Chapter 46 applies to relational measures of consciousness: a bispectral index must outperform the power spectrum it is computed from, and a coupling measure must outperform the band powers of the coupled signals.

J.2 Shuffle Surrogates

The simplest surrogate permutes the samples of a series at random. It preserves the distribution of values exactly and destroys all temporal structure. The null it tests is that the series is independent and identically distributed. Shuffling is appropriate when the unit of analysis has no internal time structure (for example, trial labels in a between-trial design) and inappropriate for testing relations between continuous signals, where it destroys the autocorrelation that any realistic null must keep.

J.3 Phase-Randomized Surrogates

Let x_n, $n = 0, \dots, N-1$, be a real series with discrete Fourier transform $\hat{x}_k = |\hat{x}_k|\, \mathrm{e}^{\mathrm{i}\phi_k}$. A Fourier-transform surrogate keeps the amplitudes and replaces the phases:

$$\hat{x}^{\mathrm{s}}_k = |\hat{x}_k|\, \mathrm{e}^{\mathrm{i}\phi'_k}, \qquad \phi'_k \sim \mathrm{Uniform}[0, 2\pi), \qquad \phi'_{N-k} = -\phi'_k, \tag{J.2}$$

where the last condition keeps the inverse transform real (the zero-frequency term and, for even N, the Nyquist term keep their original phases). The surrogate has exactly the periodogram of the data, hence its sample autocorrelation, and random phases. It realizes the null hypothesis that the data come from a stationary linear Gaussian process (Theiler et al. 1992). Any structure carried by the Fourier phases, including nonlinearity, transient events, and phase relations among frequencies, is destroyed.

Two cautions apply. The discrete Fourier transform treats the series as periodic, so a mismatch between the first and last values introduces a jump that spreads power across frequencies; trimming the series so that its ends and their slopes match reduces the artifact (Schreiber and Schmitz 2000). And phase randomization makes a nonstationary series stationary, so a statistic sensitive to nonstationarity will reject the null for reasons unrelated to the question.

J.4 Amplitude-Adjusted Surrogates

Real signals are rarely Gaussian. The amplitude-adjusted Fourier transform (AAFT) surrogate of Theiler et al. (1992) tests the null that the data are a static, monotone nonlinear transformation of a linear Gaussian process: rescale the data to a Gaussian by rank order, phase-randomize, and rescale back to the original values by rank order. The result preserves the distribution of values exactly but the spectrum only approximately.

The iterative AAFT (IAAFT) of Schreiber and Schmitz (1996) improves the match by alternating two projections until they converge:

1. Start from a random shuffle of the data.
2. Replace the Fourier amplitudes of the current series with those of the data, keeping its phases.
3. Replace the values of the result, by rank order, with the sorted values of the data.
4. Repeat steps 2 and 3 until the spectrum stops changing.

The final series has exactly the data's distribution of values and very nearly its spectrum. Schreiber and Schmitz (2000) and Lancaster et al. (2018) review the construction, its convergence, and its failure modes.

J.5 Multivariate Surrogates

Relational measures between two or more signals need surrogates that respect each signal's own structure. Two constructions answer different questions.

Definition J.2 (Common-phase and independent-phase surrogates)**.** Given signals x_j, $j = 1, \dots, J$, with transforms $\hat{x}_{j,k}$:

1. a *common-phase* surrogate adds the same random phase to every signal at each frequency, $\hat{x}^{\mathrm{s}}_{j,k} = \hat{x}_{j,k}\, \mathrm{e}^{\mathrm{i}\phi'_k}$;
2. an *independent-phase* surrogate draws a separate random phase for each signal, $\hat{x}^{\mathrm{s}}_{j,k} = \hat{x}_{j,k}\, \mathrm{e}^{\mathrm{i}\phi'_{j,k}}$.

The common-phase surrogate preserves every auto-spectrum and every cross-spectrum, because a phase added equally to two signals cancels in their cross-spectrum (Prichard and Theiler 1994). It therefore preserves the linear relation between the signals and tests whether their relation contains anything

beyond what a linear Gaussian model with the same cross-spectrum would produce. The independent-phase surrogate preserves every auto-spectrum and destroys the cross-spectrum. It tests whether the signals are related at all beyond what their individual spectra would produce by chance.

A test of phase synchrony or coherence between two brain signals needs the second construction. A test of whether the relation between two signals is nonlinear needs the first. Using the wrong one produces either a trivial rejection or a test with no power.

J.6 Time-Shift and Trial-Shuffle Surrogates

Fourier surrogates impose stationarity and Gaussian phase statistics on each signal. When the signals are nonstationary or strongly non-Gaussian, a weaker and often better null keeps each signal intact and destroys only their alignment. A *time-shift* surrogate circularly shifts one signal relative to the other by a random lag longer than the autocorrelation time of either; every property of each signal is preserved, and so is everything about their relation except its timing. For phase–amplitude coupling, the standard null shifts the amplitude series relative to the phase series (Tort et al. 2010; Aru et al. 2015).

In trial-based experiments, a *trial-shuffle* surrogate pairs signal x from trial i with signal y from trial $j \neq i$. It preserves each signal's structure within trials and every component locked to the stimulus, and destroys only trial-specific co-fluctuation. The difference between the within-trial and the shuffled-trial value of a measure separates coupling that arises from a shared stimulus from coupling intrinsic to each trial. The distinction matters for frequency-tagging designs, where both inputs are locked to the stimulus by construction (Appendix K).

J.7 Relational Measures and Their Nulls

Table J.1 pairs the relational measures used in this book with nulls that satisfy Principle J.1.

Two entries repay comment. For the phase-locking value computed from N independent, uniformly distributed phase differences, $\mathbb{E}[\mathrm{PLV}^2] = 1/N$ exactly, and $\mathbb{E}[\mathrm{PLV}] \approx \sqrt{\pi}/(2\sqrt{N})$ for large N, so a naive comparison against zero is biased in favor of synchrony at small sample sizes; Appendix K gives the unbiased estimator. For bicoherence, the matched constituent measure is the power spectrum: Miller et al. (2004) compared the bispectral component of a commercial anesthesia monitor with an analogous parameter computed from the power spectrum during induction of anesthesia in 39 patients, found the two correlated ($r = 0.84$) with neither superior in separating awake from anesthetized states, and could not show that the bispectrum added information.

Table J.1: Relational measures, what can produce them without the relation, and matched nulls.

Measure	Produced without the relation by	Matched null
Phase-locking value, coherence	Finite samples; shared stimulus drive; volume conduction	Independent-phase or trial-shuffle surrogates; lag-sensitive estimators (Appendix K)
Phase–amplitude coupling	Nonsinusoidal waveforms; sharp transients; filter settings; power changes	Time-shifted amplitude series; waveform-shape controls (Aru et al. 2015; Cole and Voytek 2017)
Bicoherence	Finite samples; amplitude changes in the component bands	Phase-randomized surrogates; comparison with the power spectrum (Miller et al. 2004)
Intermodulation responses	Local nonlinearities acting on each input separately; stimulus artifacts	Configural controls that keep each input's stimulation and remove its integration (Boremanse et al. 2013)
Order parameters, synchrony indices	Finite ensembles; common input	Independent-phase surrogates; ensembles of matched size

J.8 Multiple Comparisons and Nonstationarity

Relational measures are usually computed over many channels, frequency pairs, and time windows, and a surrogate test at each point multiplies the chance of false positives. Cluster-based permutation tests control the family-wise error rate while exploiting the fact that true effects extend over neighboring channels, frequencies, and times: the statistic is the size or mass of the largest cluster of neighboring points exceeding a threshold, and its null distribution is obtained by permuting condition labels (Maris and Oostenveld 2007).

Nonstationarity is the most common source of spurious results. A surrogate that preserves the data's global spectrum but not its changes over time will judge any time-varying coupling significant. The remedies are to compute the test within windows short enough to be approximately stationary, to use time-shift or trial-shuffle surrogates that preserve each signal's nonstationarity, or to model the nonstationarity explicitly and test against the model.

J.9 Notes and References

The method of surrogate data was introduced by Theiler et al. (1992); the iterative amplitude-adjusted construction is from Schreiber and Schmitz (1996),

and multivariate surrogates that preserve cross-spectra from Prichard and Theiler (1994). Schreiber and Schmitz (2000) is the standard review, and Lancaster et al. (2018) a comprehensive recent one covering physical systems. For phase–amplitude coupling, Tort et al. (2010) give the modulation index with its surrogate procedure and Aru et al. (2015) the pitfalls; Cole and Voytek (2017) show how waveform shape produces spurious coupling. Nonparametric cluster-based testing for electrophysiological data is from Maris and Oostenveld (2007); resampling methods in general are covered by Efron and Tibshirani (1993). The comparison of bispectral and power-spectral parameters in anesthesia is Miller et al. (2004), and the intermodulation study with configural controls is Boremanse et al. (2013).

K

EEG and MEG Spectral Analysis

Most of the relational measures of brain activity discussed in Part X and Part XIII are computed from electroencephalography (EEG) or magnetoencephalography (MEG). This appendix collects what a reader needs to evaluate those measures: what the sensors record, how spectra are estimated, why band power can change without any oscillation changing, how synchrony between signals is measured and how volume conduction fakes it, how cross-frequency coupling is estimated and how waveform shape fakes it, what steady-state and intermodulation responses show, and how statistics are done across many sensors and frequencies. Surrogate nulls for each measure are in Appendix J.

K.1 What the Sensors Record

EEG records potential differences between electrodes on the scalp; MEG records the magnetic fields just outside the head. Both are dominated by the summed postsynaptic currents of large populations of spatially aligned cortical pyramidal neurons, and both see such a population only when many of its members are active together (Nunez and Srinivasan 2006; Hämäläinen et al. 1993). Each sensor records a weighted sum of many sources, with weights fixed by the geometry and conductivity of the head. The mixing is effectively instantaneous at the frequencies of interest, which has a consequence for every measure of coupling between sensors (Section K.5).

Estimating the sources from the sensors is the inverse problem. It has no unique solution without constraints, and every source-reconstruction method adds constraints of its own; residual mixing among reconstructed sources, called leakage, remains (Hämäläinen et al. 1993; Bastos and Schoffelen 2016). A relational measure computed between two sensors or two reconstructed sources is therefore always partly a measure of the mixing.

K.2 Estimating a Spectrum

For a series x_n sampled at interval Δt, the periodogram

$$P(f) = \frac{\Delta t}{N} \left| \sum_{n=0}^{N-1} x_n \, \mathrm{e}^{-\mathrm{i}2\pi f n \Delta t} \right|^2 \tag{K.1}$$

is the natural estimate of the power spectrum, and it is a poor one. Its variance does not decrease as N grows, and the implicit rectangular window leaks power from strong peaks into neighboring frequencies. Tapering the data with a smooth window reduces leakage at the cost of resolution.

Welch's method divides the record into overlapping segments, tapers each, and averages their periodograms, reducing variance in proportion to the number of effectively independent segments at the cost of the frequency resolution of a single segment (Welch 1967). The multitaper method of Thomson (1982) instead computes several periodograms of the whole record with a family of orthogonal tapers, the discrete prolate spheroidal (Slepian) sequences, chosen to concentrate as much energy as possible within a band of half-width W; averaging roughly $2NW - 1$ of them reduces variance while fixing the resolution at $2W$ (Percival and Walden 1993). For signals whose spectrum changes over time, short-time Fourier transforms and wavelet convolution trade temporal against spectral resolution; no choice escapes the trade-off, and the choice should be reported (M. X. Cohen 2014).

K.3 Periodic and Aperiodic Components

The power spectrum of neural field recordings has two parts: peaks at the frequencies of oscillations, such as the alpha rhythm, and an aperiodic background whose power falls roughly as a power law in frequency. The aperiodic component has its own parameters, an offset and an exponent, and they change with age, task, and state. A change in power within a fixed frequency band can therefore reflect a change in the aperiodic component with no change in any oscillation. Donoghue et al. (2020) introduced a method that fits the spectrum as an aperiodic function plus a set of peaks, and showed that apparent changes in band power, and in ratios of band powers, can arise from aperiodic changes alone.

The point matters for relational measures in two ways. A ratio of band powers (a relation between two frequencies) can change because the aperiodic slope changes, which is a property of neither band's oscillation. And measures of phase are meaningful only where an oscillation is present: the phase of a band that contains only aperiodic power is the phase of filtered noise. Any relational claim about oscillations should report that the oscillations exist, by showing a peak above the aperiodic fit.

K.4 Phase and the Analytic Signal

To speak of the phase of a band-limited signal, one forms its analytic signal. After band-pass filtering $x(t)$, let $\mathcal{H}[x]$ be its Hilbert transform and

$$z(t) = x(t) + \mathrm{i}\,\mathcal{H}[x](t) = a(t)\,\mathrm{e}^{\mathrm{i}\varphi(t)}, \tag{K.2}$$

with instantaneous amplitude $a(t)$ and phase $\varphi(t)$. The construction is well behaved only for narrow bands; for broadband signals the instantaneous phase can behave erratically. Wavelet convolution with a complex kernel gives equivalent estimates (M. X. Cohen 2014).

K.5 Coupling Between Signals

K.5.1 Phase locking

Given phase differences $\Delta\varphi_k$ between two signals at N time points or trials, the phase-locking value is

$$\mathrm{PLV} = \Big|\frac{1}{N}\sum_{k=1}^{N}\mathrm{e}^{\mathrm{i}\Delta\varphi_k}\Big| \tag{K.3}$$

(Lachaux et al. 1999); the same quantity applied to a single pair of signals over time is the mean phase coherence of Mormann et al. (2000). The PLV is biased upward at small N: for independent uniform phase differences $\mathbb{E}[\mathrm{PLV}^2] = 1/N$. Expanding $|\sum_k \mathrm{e}^{\mathrm{i}\Delta\varphi_k}|^2 = N + 2\sum_{j<k}\cos(\Delta\varphi_j - \Delta\varphi_k)$ shows that

$$\mathrm{PPC} = \frac{2}{N(N-1)}\sum_{j<k}\cos(\Delta\varphi_j - \Delta\varphi_k) = \frac{N\,\mathrm{PLV}^2 - 1}{N-1} \tag{K.4}$$

has expectation zero under independence; this pairwise phase consistency is an unbiased estimator of the squared population phase locking (Vinck, Wingerden, et al. 2010).

K.5.2 Coherence and volume conduction

The coherency between signals x and y at frequency f is the normalized cross-spectrum $C_{xy}(f) = S_{xy}(f)/\sqrt{S_{xx}(f)S_{yy}(f)}$, and coherence is its squared magnitude. Instantaneous mixing produces coherence without any interaction between sources.

Proposition K.1 (Instantaneous mixing leaves the imaginary part of coherency untouched)**.** *Let independent sources s_i be mixed instantaneously into two sensors, $x = \sum_i a_i s_i$ and $y = \sum_i b_i s_i$, with real weights a_i, b_i. Then $S_{xy}(f) = \sum_i a_i b_i\, S_{s_i s_i}(f)$, which is real, so $\operatorname{Im} C_{xy}(f) = 0$. A nonzero imaginary part of coherency therefore cannot arise from instantaneous mixing of independent sources, and indicates an interaction with a time lag.*

Proof. Independence gives $S_{s_i s_j} = 0$ for $i \neq j$, and each auto-spectrum $S_{s_i s_i}$ is real and nonnegative; with real weights, every term of the sum is real. □

Nolte et al. (2004) proposed the imaginary part of coherency on this basis. The phase lag index measures the consistency of the sign of the phase difference, discarding zero-lag contributions (Stam et al. 2007), and the weighted phase lag index weights by the magnitude of the imaginary component to reduce sensitivity to noise (Vinck, Oostenveld, et al. 2011). The price of all such measures is that genuine interactions with zero lag are invisible to them. Bastos and Schoffelen (2016) review these measures and the other ways coupling estimates mislead: common input from a third source, differences in signal-to-noise ratio between signals, and residual leakage after source reconstruction.

K.6 Cross-Frequency Coupling

Phase–amplitude coupling (PAC) asks whether the amplitude of a fast rhythm varies with the phase of a slow one. With the amplitude envelope $a_{\mathrm{H}}(t)$ of a high band and the phase $\varphi_{\mathrm{L}}(t)$ of a low band, the mean vector length

$$\mathrm{MVL} = \left| \frac{1}{N} \sum_{n} a_{\mathrm{H}}(t_n)\, \mathrm{e}^{\mathrm{i}\varphi_{\mathrm{L}}(t_n)} \right| \tag{K.5}$$

measures the coupling (Canolty, Edwards, et al. 2006); the modulation index of Tort et al. (2010) measures instead how far the distribution of high-band amplitude over low-band phase bins departs from uniform, using the Kullback–Leibler divergence. Canolty and Knight (2010) review the proposed functional roles of such coupling.

Cross-frequency coupling is unusually easy to fake. A nonsinusoidal oscillation has harmonics whose phases are locked to the fundamental, and filtering such a waveform into a low and a high band produces apparent coupling between them with no second rhythm present (Cole and Voytek 2017). Sharp transients in the signal produce broadband power locked to the phase of any slower component that contains them (Kramer et al. 2008). Changes in the power of either band change the estimates. Aru et al. (2015) list the conditions a coupling claim should meet: an oscillation must be present in each band, the waveform shape must be examined, the filter bandwidth for the amplitude band must be wide enough to contain the sidebands produced by the modulation, and the result must survive a surrogate test that preserves each signal (Appendix J).

K.7 Steady-State Responses and Intermodulation

Stimulating the visual system at a fixed frequency f_1 evokes a steady-state response at f_1 and its harmonics, which can be measured with high signal-to-noise ratio because all its power lies at known frequencies (Norcia et al. 2015).

Responses to flicker from 1 to 100 Hz show resonance-like peaks at preferred frequencies (Herrmann 2001).

Presenting two inputs tagged at different frequencies f_1 and f_2 turns the method into a test of integration. A linear system, or any system that processes the two inputs separately, responds only at multiples of each tag, nf_1 and mf_2. A response at an intermodulation frequency $nf_1 + mf_2$ with both n and m nonzero requires a nonlinearity that combines the two inputs, the brain's version of the difference frequencies of Chapter 4 (M. Regan and D. Regan 1988). Boremanse et al. (2013) tagged the left and right halves of a face at 5.88 Hz and 7.14 Hz and found intermodulation components, for example at their difference of 1.26 Hz, over the right occipito-temporal cortex. The responses at the fundamental frequencies did not depend on how the halves were arranged, whereas the intermodulation components were nearly absent when the halves were misaligned and separated and were reduced when the face was inverted. The configural controls are what make the result informative: local nonlinearities acting on each half would survive misalignment, so their loss identifies integration of the halves into a whole.

K.8 Higher-Order Spectra

The power spectrum discards phase. The bispectrum,

$$B(f_1, f_2) = \mathbb{E}\big[\hat{x}(f_1)\,\hat{x}(f_2)\,\hat{x}^*(f_1 + f_2)\big], \tag{K.6}$$

is nonzero when the phase at $f_1 + f_2$ is consistently related to the phases at f_1 and f_2, as quadratic nonlinearities produce; bicoherence normalizes it to remove the dependence on amplitude (Nikias and Mendel 1993). Sigl and Chamoun (1994) introduced bispectral analysis of the EEG for monitoring anesthesia, and a bispectral component enters a widely used commercial index. The caution of Appendix J applies with full force: in the comparison of Miller et al. (2004), the bispectral parameter carried no more information about the anesthetized state than an analogous power-spectral parameter, and an amplitude-independent bicoherence parameter did not change during induction at all. A higher-order spectral measure earns its place only by what it adds to the spectrum it is computed from.

K.9 Statistics Across Sensors, Frequencies, and Time

A typical analysis computes a measure at many sensors, frequencies, and time points, and so performs thousands of tests. Cluster-based permutation tests address the problem by forming clusters of neighboring points whose statistic exceeds a threshold, summarizing each cluster by its size or summed

statistic, and comparing the largest cluster with its distribution under random permutation of condition labels; the procedure controls the family-wise error rate without assuming a parametric distribution (Maris and Oostenveld 2007). Open software implements these steps together with the spectral and coupling measures of this appendix (Gramfort 2013).

K.10 Notes and References

The physics of EEG is treated by Nunez and Srinivasan (2006) and of MEG by Hämäläinen et al. (1993). Welch (1967) introduced averaged modified periodograms and Thomson (1982) the multitaper method; Percival and Walden (1993) is the standard text on both. M. X. Cohen (2014) is a practical guide to time–frequency analysis of neural data. On aperiodic components, see Donoghue et al. (2020). On phase synchrony, Lachaux et al. (1999), Mormann et al. (2000), and Vinck, Wingerden, et al. (2010); on volume conduction and lag-based measures, Nolte et al. (2004), Stam et al. (2007), and Vinck, Oostenveld, et al. (2011); on pitfalls of connectivity analysis in general, Bastos and Schoffelen (2016). On cross-frequency coupling, Canolty, Edwards, et al. (2006), Tort et al. (2010), and Canolty and Knight (2010), with the cautions of Kramer et al. (2008), Aru et al. (2015), and Cole and Voytek (2017). On steady-state responses and intermodulation, Norcia et al. (2015), M. Regan and D. Regan (1988), and Boremanse et al. (2013); on flicker resonance, Herrmann (2001). On higher-order spectra, Nikias and Mendel (1993); on the bispectral index, Sigl and Chamoun (1994) and Miller et al. (2004). Cluster-based permutation testing is from Maris and Oostenveld (2007); Gramfort (2013) describe an open implementation.

L

Comparative Glossary

Each entry gives a term, the field in which it is at home, and the sense in which this book uses it, with the chapter where it is developed. The glossary is comparative on purpose: its second column records who already owns each concept. A final section lists terms the book considered and declined to coin, with the established term used instead.

L.1 Terms

Term	Home field	Use in this book
Access consciousness	Philosophy of mind	Availability of a content for reasoning, report, and the control of action; distinguished from phenomenal consciousness (Chapter 43).
Active sensing	Sensory neuroscience	Sensing in which the animal emits the energy it senses, or shapes its input by its own movement (Chapter 19).
Agent-centered measurement frame	This book (label for an established structure)	A reading loop whose policy acts on self-indexed variables; model-based or carrier-based (Chapter 22).
Arnold tongue	Nonlinear dynamics	Region of parameter space in which a driven oscillator locks at a given rational frequency ratio (Chapter 7, Appendix F).
Attention schema	Cognitive neuroscience	A simplified internal model of one's own attention, proposed as the basis of reports of awareness (Chapter 42).
Autopoiesis	Theoretical biology	Organization of a system that continuously produces the components that produce it, including its boundary (Chapter 27).
Beat	Acoustics, signal theory	Slow amplitude modulation of the sum of two nearby frequencies at their difference frequency (Chapter 3).

Term	Home field	Use in this book
Bicoherence	Signal processing	Normalized bispectrum; measures phase coupling among frequency components (Chapter 46, Appendix B).
Carrier	Communications engineering	A periodic signal on which information is impressed by modulation; for the wave-type fish, its own discharge (Chapter 19).
Causal emergence	Complex systems	The claim that a coarse-grained description can have more effective information than the micro-description; contested (Chapter 14).
Chimera state	Nonlinear dynamics	Coexistence of synchronized and desynchronized groups among identical coupled oscillators (Chapter 12).
Closed loop	Control theory	A configuration in which a system's output affects its own input (Chapters 20 and 59).
Commensurability	Nonlinear dynamics, condensed matter	A rational ratio of frequencies or periods; distinct from resonance and from magnitude balance (Chapter 7).
Contamination ratio	Electric fish neuroethology	Relative amplitude m of a neighbor's field at a patch of the fish's skin (Chapter 20).
Coordination dynamics	Movement science	Study of the dynamics of relative phase and related collective variables in behavior and brain (Chapter 10).
Corollary discharge	Neuroscience	Any internal signal derived from a motor command and sent to non-motor structures (Chapter 22).
Counterfactual	Philosophy, causal inference	What would have happened under a different condition; the test of genuine implementation (Chapters 17 and 32).
Demodulation	Signal theory	Extraction of a modulation from a modulated carrier; a paradigmatic physical reader (Chapter 4).
Detached representation	Cognitive science	A representation that can be used in the absence of the stimulus it represents (Chapter 31).
Detector	Physics, engineering	A device whose output depends systematically on some property of its input; not yet a reader unless something uses the output (Chapter 64).
Differential phase	Electric fish neuroethology	Phase of the signal at one patch of skin relative to another; frame-independent (Chapter 21).

Term	Home field	Use in this book
Disjunction problem	Philosophy of mind	The difficulty, for causal theories of content, of distinguishing misrepresentation from correct representation of a disjunctive cause (Chapters 23 and 29).
Effective information	Complex systems	Mutual information between a maximum-entropy intervention and its effect; the measure used in causal-emergence arguments (Chapter 14).
Efference copy	Neuroscience	A corollary discharge used to predict and cancel the sensory consequences of one's own action (Chapter 22).
Electric image	Electric fish neuroethology	Pattern of fractional amplitude change produced on the fish's skin by a nearby object (Chapter 19).
Epistemic cut	Theoretical biology (Pattee)	The necessary distinction between the dynamics of a system and the measurements or symbols that describe or control it (Chapter 25).
Essential indexical	Philosophy of language	A self-locating belief that cannot be replaced by a non-indexical one without changing its role in action (Chapter 37).
Exafference	Neuroscience	Sensory input caused by events in the world rather than by the animal's own action (Chapter 22).
Explanatory gap	Philosophy of mind	The apparent absence of an explanation of why a physical process is accompanied by a particular experience (Chapter 43).
Frame-independent observable	Physics	A quantity unchanged by transformations of the reference frame; by Theorem 1.5, necessarily relational when the frame group acts transitively (Chapter 1).
Functional perspective	This book (label)	A set of agent-centered variables that a system reads and uses (Chapters 43 and 64).
Generative model	Machine learning, neuroscience	A probabilistic model of how hidden causes produce observations, inverted in perception (Chapter 33).
Global workspace	Cognitive neuroscience	A capacity-limited stage whose contents are broadcast to many specialized processes (Chapter 44).
Heterodyning	Radio engineering	Mixing a signal with a local oscillator to shift it to a difference frequency (Chapters 4 and 48).
Higher-order theory	Philosophy of mind	The view that a mental state is conscious when it is the object of a suitable higher-order representation (Chapter 40).

Term	Home field	Use in this book
Hypergraph	Mathematics, network science	Generalization of a graph in which edges join any number of nodes (Appendix G).
Implementation	Philosophy of computation	The relation between a physical system and a computation it performs, constrained by causal and counterfactual structure (Chapter 17).
Integrated information	Consciousness science	A measure of the irreducibility of a system's cause–effect structure, proposed by IIT as identical to consciousness (Chapter 44).
Intentionality	Philosophy of mind	The aboutness of mental states (Chapter 29).
Interoception	Physiology, neuroscience	Sensing of the physiological condition of the body (Chapter 36).
Intermodulation	Signal theory	Components at combinations of input frequencies produced by a nonlinear system; a signature of joint processing (Chapters 4 and 46).
Jamming avoidance response (JAR)	Electric fish neuroethology	Shift of a wave-type fish's discharge frequency away from a neighbor's (Chapters 20 and 21).
Lock-in detection	Physics instrumentation	Phase-sensitive detection of a signal at a known reference frequency (Chapter 4).
Magic angle	Condensed matter physics	Twist angle of bilayer graphene at which the lowest bands become flat (Chapter 8).
Metacognition	Cognitive science	Monitoring and control of one's own cognitive states, including confidence (Chapter 39).
Metastability	Dynamical systems, neuroscience	Transient coordination without permanent locking (Chapter 11).
Metaxics	This book	A research perspective on relational observables and their readers; not, as yet, a discipline (Preface, Chapter 56).
Misrepresentation	Philosophy of mind	A representation whose content does not match the world, with functional consequences (Chapter 30).
Moiré	Optics, physics	Large-scale pattern produced by the superposition of two similar periodic structures (Chapter 2).
Normativity	Philosophy, theoretical biology	The property of being better or worse for a system; grounded in viability or selection history (Chapter 28).
Order parameter	Statistical physics, synergetics	A macroscopic variable that describes the collective state of a system near a transition (Chapter 13).

Term	Home field	Use in this book
Phantom jamming	Electric fish neuroethology	Frequency shifts evoked by non-jamming stimuli that mimic the jamming cues (Chapter 23).
Phase reduction	Nonlinear dynamics	Description of a weakly perturbed limit-cycle oscillator by its phase alone (Appendix D).
Phenomenal consciousness	Philosophy of mind	There being something it is like to be in a state (Chapter 43).
Polycomputation	Computer science, biology	The performance of several computations by one substrate at once, relative to different observers or readouts (Chapter 18).
Reafference	Neuroscience	Sensory input caused by the animal's own action (Chapter 22).
Reader	This book (label for an established notion)	A physical operation that makes a variable of its input dynamically explicit; transducer, classifier, controller, interpreter, or subject (Chapter 15).
Reading loop	This book (label)	The closed structure $\mathcal{L} = (\mathsf{M}, \mathsf{D}, \pi, \mathsf{F})$ in which a reader acts on the relation it reads (Chapter 15).
Relational observable	Physics, philosophy of science	A variable of a configuration that is a function of no single component (Chapter 1).
Relative phase	Nonlinear dynamics, movement science	Difference of the phases of two oscillators; the collective variable of coordination dynamics (Chapter 10).
Resonance	Physics	Peak response of a driven system near its natural frequency (Chapter 6).
Rotation number	Nonlinear dynamics	Average phase advance per iteration of a circle map (Chapter 7, Appendix F).
Self-indexed variable	This book (label)	A sensory variable that depends on the world only through a relation to something the agent produces (Chapter 19).
Self-model	Philosophy of mind, neuroscience	A representation of the system itself, including its body and states (Chapter 41).
Semantic closure	Theoretical biology (Pattee)	Organization in which a system constructs the symbols and measuring devices that control its own construction (Chapter 26).
Sense-making	Enactive cognitive science	The organism's evaluation of its encounters in terms of their significance for its viability (Chapter 28).
Slaving principle	Synergetics	Near an instability, fast modes follow the slow order parameters (Chapter 13).

Term	Home field	Use in this book
Specious present	Psychology, phenomenology	The short interval experienced as present (Chapter 35).
Surrogate data	Time-series analysis	Artificial data sharing chosen properties with the original, used to build null models (Appendix J).
Synchronization	Nonlinear dynamics	Adjustment of rhythms of interacting oscillators (Chapter 9).
Synergy	Information theory	Information carried jointly by several variables and by none alone (Chapter 1, Appendix H).
Teleosemantics	Philosophy of mind	Theories that ground content in biological function (Chapter 29).
Umwelt	Biology (von Uexküll)	The world as it is available to a particular organism through its senses and actions (Chapter 27).

L.2 Terms Declined

The method of the book (Preface) forbids coining a term unless it names a distinction, prediction, or operation that existing vocabulary cannot express. The following candidates were considered while the book was planned and were declined.

Declined term	Established term used instead, and why
Perspectival symmetry breaking	*Reafferent active sensing* and the *agent-centered measurement frame*. Occupying a location breaks translation symmetry for a rock too; what distinguishes the fish is that its measurement variable is indexed to its own emission (Chapter 22).
Metaxics as a new science	*Research perspective*. No new formalism, transferable theorem, experimental operation, or distinctive confirmed prediction has yet been shown (Chapter 56).
Reader-dependent existence	*Reader-relative variables*: which relational variables become dynamically explicit depends on the mixture and the transduction; nothing about what exists depends on a reader (Chapter 16).
Relational consciousness	The major theories already make relational claims (integration, broadcast, recurrence, higher-order representation); a new label would add nothing to them (Chapter 44).
The right amount of misalignment	Separated into *magnitude balance*, *resonance*, and *commensurability*, three different phenomena (Part II).
Moiré consciousness	*Neural frequency relations*, which must outperform constituent spectral measures to earn a place (Chapter 46).

L.3 Notes and References

Definitions in the first column follow the sources cited in the chapters indicated. For information-theoretic terms see Cover and Thomas (2006); for nonlinear dynamics, Strogatz (2015) and Pikovsky et al. (2001); for the philosophy of mind, Chalmers (1996b) and Shea (2018); for electric fish, Heiligenberg (1991a).

References

Abbott, A. (2001). *Chaos of Disciplines*. Chicago: University of Chicago Press.

Abrams, D. M., R. Mirollo, S. H. Strogatz, and D. A. Wiley (2008). "Solvable Model for Chimera States of Coupled Oscillators". In: *Physical Review Letters* 101.8, 084103. DOI: 10.1103/physrevlett.101.084103.

Abrams, D. M. and S. H. Strogatz (2004). "Chimera States for Coupled Oscillators". In: *Physical Review Letters* 93.17, 174102. DOI: 10.1103/physrevlett.93.174102.

Acebrón, J. A., L. L. Bonilla, C. J. Pérez Vicente, F. Ritort, and R. Spigler (2005). "The Kuramoto model: A simple paradigm for synchronization phenomena". In: *Reviews of Modern Physics* 77.1, pp. 137–185. DOI: 10.1103/revmodphys.77.137.

Adler, J. (1966). "Chemotaxis in Bacteria". In: *Science* 153.3737, pp. 708–716. DOI: 10.1126/science.153.3737.708.

Adler, J. (1969). "Chemoreceptors in Bacteria". In: *Science* 166.3913, pp. 1588–1597. DOI: 10.1126/science.166.3913.1588.

Adler, R. (1946). "A Study of Locking Phenomena in Oscillators". In: *Proceedings of the IRE* 34.6, pp. 351–357. DOI: 10.1109/jrproc.1946.229930.

Albantakis, L., L. Barbosa, G. Findlay, M. Grasso, A. M. Haun, W. Marshall, W. G. P. Mayner, A. Zaeemzadeh, M. Boly, B. E. Juel, S. Sasai, K. Fujii, I. David, J. Hendren, J. P. Lang, and G. Tononi (2023). "Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms". In: *PLOS Computational Biology* 19.10, e1011465. DOI: 10.1371/journal.pcbi.1011465.

Alexander, H. G., ed. (1956). *The Leibniz–Clarke Correspondence*. Manchester: Manchester University Press.

Ali, M. M., K. K. Sellers, and F. Frohlich (2013). "Transcranial Alternating Current Stimulation Modulates Large-Scale Cortical Network Activity by Network Resonance". In: *Journal of Neuroscience* 33.27, pp. 11262–11275. DOI: 10.1523/jneurosci.5867-12.2013.

Alkire, M. T., A. G. Hudetz, and G. Tononi (2008). "Consciousness and Anesthesia". In: *Science* 322.5903, pp. 876–880. DOI: 10.1126/science.1149213.

Alon, U., M. G. Surette, N. Barkai, and S. Leibler (1999). "Robustness in bacterial chemotaxis". In: *Nature* 397.6715, pp. 168–171. DOI: 10.1038/16483.

Amidror, I. (2009). *The Theory of the Moiré Phenomenon, Volume I: Periodic Layers*. 2nd ed. London: Springer.

Andersen, R. A., G. K. Essick, and R. M. Siegel (1985). "Encoding of Spatial Location by Posterior Parietal Neurons". In: *Science* 230.4724, pp. 456–458. DOI: 10.1126/science.4048942.

Andersen, R. A., L. H. Snyder, D. C. Bradley, and J. Xing (1997). "Multimodal representation of space in the posterior parietal cortex and its use in planning movements". In: *Annual Review of Neuroscience* 20.1, pp. 303–330. DOI: 10.1146/annurev.neuro.20.1.303.

Anderson, P. W. (1972). "More Is Different". In: *Science* 177.4047, pp. 393–396. DOI: 10.1126/science.177.4047.393.

Andrei, E. Y. and A. H. MacDonald (2020). "Graphene bilayers with a twist". In: *Nature Materials* 19.12, pp. 1265–1275. DOI: 10.1038/s41563-020-00840-0.

Andrews, M. (2021). "The math is not the territory: navigating the free energy principle". In: *Biology & Philosophy* 36.3, 30. DOI: 10.1007/s10539-021-09807-0.

Aristotle (1984). "On the Soul". In: *The Complete Works of Aristotle: The Revised Oxford Translation*. Ed. by J. Barnes. Trans. by J. A. Smith. Vol. 1. Princeton: Princeton University Press.

Armstrong, D. M. (1968). *A Materialist Theory of the Mind*. London: Routledge & Kegan Paul.

Armstrong, E. (1921). "A New System of Short Wave Amplification". In: *Proceedings of the IRE* 9.1, pp. 3–11. DOI: 10.1109/jrproc.1921.220092.

Arnold, V. I. (1983). *Geometrical Methods in the Theory of Ordinary Differential Equations*. New York: Springer.

Aru, J., J. Aru, V. Priesemann, M. Wibral, L. Lana, G. Pipa, W. Singer, and R. Vicente (2015). "Untangling cross-frequency coupling in neuroscience". In: *Current Opinion in Neurobiology* 31, pp. 51–61. DOI: 10.1016/j.conb.2014.08.002.

Asamoah, B., A. Khatoun, and M. Mc Laughlin (2019). "tACS motor system effects can be caused by transcutaneous stimulation of peripheral nerves". In: *Nature Communications* 10.1, 266. DOI: 10.1038/s41467-018-08183-w.

Ashby, W. R. (1956). *An Introduction to Cybernetics*. London: Chapman & Hall.

Ashwin, P. and A. Rodrigues (2016). "Hopf normal form with S_N symmetry and reduction to systems of nonlinearly coupled phase oscillators". In: *Physica D: Nonlinear Phenomena* 325, pp. 14–24. DOI: 10.1016/j.physd.2016.02.009.

Avidan, M. S., L. Zhang, B. A. Burnside, K. J. Finkel, A. C. Searleman, J. A. Selvidge, L. Saager, M. S. Turner, S. Rao, M. Bottros, C. Hantler, E. Jacobsohn, and A. S. Evers (2008). "Anesthesia Awareness and the Bispectral Index". In: *New England Journal of Medicine* 358.11, pp. 1097–1108. DOI: 10.1056/nejmoa0707361.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Bachmann, P. A., P. L. Luisi, and J. Lang (1992). "Autocatalytic self-replicating micelles as models for prebiotic structures". In: *Nature* 357.6373, pp. 57–59. DOI: 10.1038/357057a0.

Bak, P. (1982). "Commensurate phases, incommensurate phases and the devil's staircase". In: *Reports on Progress in Physics* 45.6, pp. 587–629. DOI: 10.1088/0034-4885/45/6/001.

Bak, P. (1986). "The Devil's Staircase". In: *Physics Today* 39.12, pp. 38–45. DOI: 10.1063/1.881047.

Barandiaran, X. E., E. Di Paolo, and M. Rohde (2009). "Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action". In: *Adaptive Behavior* 17.5, pp. 367–386. DOI: 10.1177/1059712309343819.

Barbour, J. B. and B. Bertotti (1982). "Mach's principle and the structure of dynamical theories". In: *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 382.1783, pp. 295–306. DOI: 10.1098/rspa.1982.0102.

Barenblatt, G. I. (1996). *Scaling, Self-Similarity, and Intermediate Asymptotics*. Cambridge: Cambridge University Press.

Barkai, N. and S. Leibler (1997). "Robustness in simple biochemical networks". In: *Nature* 387.6636, pp. 913–917. DOI: 10.1038/43199.

Barnett, L., A. B. Barrett, and A. K. Seth (2009). "Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables". In: *Physical Review Letters* 103.23, 238701. DOI: 10.1103/physrevlett.103.238701.

Barron, A. B. and C. Klein (2016). "What insects can tell us about the origins of consciousness". In: *Proceedings of the National Academy of Sciences* 113.18, pp. 4900–4908. DOI: 10.1073/pnas.1520084113.

Bastian, J. and W. Heiligenberg (1980). "Phase-sensitive midbrain neurons in *Eigenmannia*: neural correlates of the jamming avoidance response". In: *Science* 209.4458, pp. 828–831. DOI: 10.1126/science.7403849.

Bastos, A. M. and J.-M. Schoffelen (2016). "A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls". In: *Frontiers in Systems Neuroscience* 9, 175. DOI: 10.3389/fnsys.2015.00175.

Battiston, F., G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri (2020). "Networks beyond pairwise interactions: Structure and dynamics". In: *Physics Reports* 874, pp. 1–92. DOI: 10.1016/j.physrep.2020.05.004.

Baumgartner, M. and A. Gebharter (2016). "Constitutive Relevance, Mutual Manipulability, and Fat-Handedness". In: *The British Journal for the Philosophy of Science* 67.3, pp. 731–756. DOI: 10.1093/bjps/axv003.

Bechtel, W. and J. Mundale (1999). "Multiple Realizability Revisited: Linking Cognitive and Neural States". In: *Philosophy of Science* 66.2, pp. 175–207. DOI: 10.1086/392683.

Beggs, J. M. and D. Plenz (2003). "Neuronal Avalanches in Neocortical Circuits". In: *The Journal of Neuroscience* 23.35, pp. 11167–11177. DOI: 10.1523/jneurosci.23-35-11167.2003.

Bell, C., D. Bodznick, J. Montgomery, and J. Bastian (1997). "The Generation and Subtraction of Sensory Expectations within Cerebellum-Like Structures". In: *Brain, Behavior and Evolution* 50.Suppl. 1, pp. 17–31. DOI: 10.1159/000113352.

Bell, C. C. (1981). "An Efference Copy Which Is Modified by Reafferent Input". In: *Science* 214.4519, pp. 450–453. DOI: 10.1126/science.7291985.

Berg, H. C. and E. M. Purcell (1977). "Physics of chemoreception". In: *Biophysical Journal* 20.2, pp. 193–219. DOI: 10.1016/S0006-3495(77)85544-6.

Berg, H. C. and D. A. Brown (1972). "Chemotaxis in *Escherichia coli* analysed by Three-dimensional Tracking". In: *Nature* 239.5374, pp. 500–504. DOI: 10.1038/239500a0.

Bertalanffy, L. von (1968). *General System Theory: Foundations, Development, Applications*. New York: George Braziller.

Bick, C., E. Gross, H. A. Harrington, and M. T. Schaub (2023). "What Are Higher-Order Networks?" In: *SIAM Review* 65.3, pp. 686–731. DOI: 10.1137/21m1414024.

Binder, K. and A. P. Young (1986). "Spin glasses: Experimental facts, theoretical concepts, and open questions". In: *Reviews of Modern Physics* 58.4, pp. 801–976. DOI: 10.1103/revmodphys.58.801.

Birch, J., A. K. Schnell, and N. S. Clayton (2020). "Dimensions of Animal Consciousness". In: *Trends in Cognitive Sciences* 24.10, pp. 789–801. DOI: 10.1016/j.tics.2020.07.007.

Bistritzer, R. and A. H. MacDonald (2011). "Moiré bands in twisted double-layer graphene". In: *Proceedings of the National Academy of Sciences* 108.30, pp. 12233–12237. DOI: 10.1073/pnas.1108174108.

Biswas, D., L. A. Arend, S. A. Stamper, B. P. Vágvölgyi, E. S. Fortune, and N. J. Cowan (2018). "Closed-Loop Control of Active Sensing Movements Regulates Sensory Slip". In: *Current Biology* 28.24, 4029–4036.e4. DOI: 10.1016/j.cub.2018.11.002.

Blair, H. T., A. C. Welday, and K. Zhang (2007). "Scale-Invariant Memory Representations Emerge from Moiré Interference between Grid Fields That Produce Theta Oscillations: A Computational Model". In: *The Journal of Neuroscience* 27.12, pp. 3211–3229. DOI: 10.1523/jneurosci.4724-06.2007.

Blakemore, S.-J., D. M. Wolpert, and C. D. Frith (1998). "Central cancellation of self-produced tickle sensation". In: *Nature Neuroscience* 1.7, pp. 635–640. DOI: 10.1038/2870.

Blanke, O. and T. Metzinger (2009). "Full-body illusions and minimal phenomenal selfhood". In: *Trends in Cognitive Sciences* 13.1, pp. 7–13. DOI: 10.1016/j.tics.2008.10.003.

Block, N. (1995). "On a confusion about a function of consciousness". In: *Behavioral and Brain Sciences* 18.2, pp. 227–247. DOI: 10.1017/s0140525x00038188.

Blumenfeld, H. (2012). "Impaired consciousness in epilepsy". In: *The Lancet Neurology* 11.9, pp. 814–826. DOI: 10.1016/s1474-4422(12)70188-6.

Boly, M., M. Massimini, N. Tsuchiya, B. R. Postle, C. Koch, and G. Tononi (2017). "Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence". In: *The Journal of Neuroscience* 37.40, pp. 9603–9613. DOI: 10.1523/jneurosci.3218-16.2017.

Bongard, J., V. Zykov, and H. Lipson (2006). "Resilient Machines Through Continuous Self-Modeling". In: *Science* 314.5802, pp. 1118–1121. DOI: 10.1126/science.1133687.

Bongard, J. and M. Levin (2023). "There's Plenty of Room Right Here: Biological Systems as Evolved, Overloaded, Multi-Scale Machines". In: *Biomimetics* 8.1, p. 110. DOI: 10.3390/biomimetics8010110.

Boremanse, A., A. M. Norcia, and B. Rossion (2013). "An objective signature for visual binding of face parts in the human brain". In: *Journal of Vision* 13.11, p. 6. DOI: 10.1167/13.11.6.

Botvinick, M. and J. Cohen (1998). "Rubber hands 'feel' touch that eyes see". In: *Nature* 391.6669, p. 756. DOI: 10.1038/35784.

Bracewell, R. N. (2000). *The Fourier Transform and Its Applications*. 3rd ed. Boston: McGraw-Hill.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press.

Braithwaite, V. (2010). *Do Fish Feel Pain?* Oxford: Oxford University Press.

Brandon, M. P., A. R. Bogaard, C. P. Libby, M. A. Connerney, K. Gupta, and M. E. Hasselmo (2011). "Reduction of Theta Rhythm Dissociates Grid Cell Spatial Periodicity from Directional Tuning". In: *Science* 332.6029, pp. 595–599. DOI: 10.1126/science.1201652.

Brenner, N., W. Bialek, and R. de Ruyter van Steveninck (2000). "Adaptive Rescaling Maximizes Information Transmission". In: *Neuron* 26.3, pp. 695–702. DOI: 10.1016/s0896-6273(00)81205-2.

Brentano, F. (1995). *Psychology from an Empirical Standpoint*. Trans. by A. C. Rancurello, D. B. Terrell, and L. L. McAlister. London: Routledge.

Brown, E. N., R. Lydic, and N. D. Schiff (2010). "General Anesthesia, Sleep, and Coma". In: *New England Journal of Medicine* 363.27, pp. 2638–2650. DOI: 10.1056/nejmra0808281.

Brown, R., H. Lau, and J. E. LeDoux (2019). "Understanding the Higher-Order Approach to Consciousness". In: *Trends in Cognitive Sciences* 23.9, pp. 754–768. DOI: 10.1016/j.tics.2019.06.009.

Buck, J. (1988). "Synchronous Rhythmic Flashing of Fireflies. II". In: *The Quarterly Review of Biology* 63.3, pp. 265–289. DOI: 10.1086/415929.

Buckingham, E. (1914). "On Physically Similar Systems; Illustrations of the Use of Dimensional Equations". In: *Physical Review* 4.4, pp. 345–376. DOI: 10.1103/physrev.4.345.

Bullock, T. H., R. H. Hamstra, and H. Scheich (1972a). "The jamming avoidance response of high frequency electric fish". In: *Journal of comparative physiology* 77.1, pp. 1–22. DOI: 10.1007/bf00696517.

Bullock, T. H., R. H. Hamstra, and H. Scheich (1972b). "The jamming avoidance response of high frequency electric fish". In: *Journal of comparative physiology* 77.1, pp. 23–48. DOI: 10.1007/bf00696518.

Bullock, T. H., C. D. Hopkins, A. N. Popper, and R. R. Fay, eds. (2005). *Electroreception*. Springer Handbook of Auditory Research 21. New York: Springer.

Burak, Y. and I. R. Fiete (2009). "Accurate Path Integration in Continuous Attractor Network Models of Grid Cells". In: *PLoS Computational Biology* 5.2, e1000291. DOI: 10.1371/journal.pcbi.1000291.

Burgess, N. (2006). "Spatial memory: how egocentric and allocentric combine". In: *Trends in Cognitive Sciences* 10.12, pp. 551–557. DOI: 10.1016/j.tics.2006.10.005.

Burgess, N., C. Barry, and J. O'Keefe (2007). "An oscillatory interference model of grid cell firing". In: *Hippocampus* 17.9, pp. 801–812. DOI: 10.1002/hipo.20327.

Buzsáki, G. (2006). *Rhythms of the Brain*. New York: Oxford University Press.

Buzsáki, G. and A. Draguhn (2004). "Neuronal Oscillations in Cortical Networks". In: *Science* 304.5679, pp. 1926–1929. DOI: 10.1126/science.1099745.

Canolty, R. T., E. Edwards, S. S. Dalal, M. Soltani, S. S. Nagarajan, H. E. Kirsch, M. S. Berger, N. M. Barbaro, and R. T. Knight (2006). "High Gamma Power Is Phase-Locked to Theta Oscillations in Human Neocortex". In: *Science* 313.5793, pp. 1626–1628. DOI: 10.1126/science.1128115.

Canolty, R. T. and R. T. Knight (2010). "The functional role of cross-frequency coupling". In: *Trends in Cognitive Sciences* 14.11, pp. 506–515. DOI: 10.1016/j.tics.2010.09.001.

Cao, Y., V. Fatemi, A. Demir, S. Fang, S. L. Tomarken, J. Y. Luo, J. D. Sanchez-Yamagishi, K. Watanabe, T. Taniguchi, E. Kaxiras, R. C. Ashoori, and P. Jarillo-Herrero (2018). "Correlated insulator behaviour at half-filling in magic-angle graphene superlattices". In: *Nature* 556.7699, pp. 80–84. DOI: 10.1038/nature26154.

Cao, Y., V. Fatemi, S. Fang, K. Watanabe, T. Taniguchi, E. Kaxiras, and P. Jarillo-Herrero (2018). "Unconventional superconductivity in magic-angle graphene superlattices". In: *Nature* 556.7699, pp. 43–50. DOI: 10.1038/nature26160.

Caputi, A. A. and R. Budelli (2006). "Peripheral electrosensory imaging by weakly electric fish". In: *Journal of Comparative Physiology A* 192.6, pp. 587–600. DOI: 10.1007/s00359-006-0100-2.

Caputi, A. A. (2004). "Contributions of electric fish to the understanding of sensory processing by reafferent systems". In: *Journal of Physiology-Paris* 98.1–3, pp. 81–97. DOI: 10.1016/j.jphysparis.2004.03.002.

Carlson, B. A. and M. Kawasaki (2004). "Nonlinear Response Properties of Combination-Sensitive Electrosensory Neurons in the Midbrain of *Gymnarchus niloticus*". In: *The Journal of Neuroscience* 24.37, pp. 8039–8048. DOI: 10.1523/jneurosci.2021-04.2004.

Carlson, B. A. and M. Kawasaki (2007). "Behavioral responses to jamming and 'phantom' jamming stimuli in the weakly electric fish *Eigenmannia*". In: *Journal of Comparative Physiology A* 193.9, pp. 927–941. DOI: 10.1007/s00359-007-0246-6.

Carr, C., W. Heiligenberg, and G. Rose (1986). "A time-comparison circuit in the electric fish midbrain. I. Behavior and physiology". In: *The Journal of Neuroscience* 6.1, pp. 107–119. DOI: 10.1523/jneurosci.06-01-00107.1986.

Carr, S., D. Massatt, S. Fang, P. Cazeaux, M. Luskin, and E. Kaxiras (2017). "Twistronics: Manipulating the electronic properties of two-dimensional layered structures through their twist angle". In: *Physical Review B* 95.7, 075420. DOI: 10.1103/physrevb.95.075420.

Carruthers, P. (2000). *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge: Cambridge University Press.

Casali, A. G., O. Gosseries, M. Rosanova, M. Boly, S. Sarasso, K. R. Casali, S. Casarotto, M.-A. Bruno, S. Laureys, G. Tononi, and M. Massimini (2013). "A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior". In: *Science Translational Medicine* 5.198. DOI: 10.1126/scitranslmed.3006294.

Casarotto, S., A. Comanducci, M. Rosanova, S. Sarasso, M. Fecchio, M. Napolitani, A. Pigorini, A. G. Casali, P. D. Trimarchi, M. Boly, O. Gosseries, O. Bodart, F. Curto, C. Landi, M. Mariotti, G. Devalle, S. Laureys, G. Tononi, and M. Massimini (2016). "Stratification of unresponsive patients by an independently validated index of brain complexity". In: *Annals of Neurology* 80.5, pp. 718–729. DOI: 10.1002/ana.24779.

Cassirer, E. (1923). *Substance and Function and Einstein's Theory of Relativity*. Trans. by W. C. Swabey and M. C. Swabey. Chicago: Open Court.

Castañeda, H.-N. (1966). "'He': A Study in the Logic of Self-Consciousness". In: *Ratio* 8, pp. 130–157.

Castro Neto, A. H., F. Guinea, N. M. R. Peres, K. S. Novoselov, and A. K. Geim (2009). "The electronic properties of graphene". In: *Reviews of Modern Physics* 81.1, pp. 109–162. DOI: 10.1103/revmodphys.81.109.

Chalmers, D. J. (1995). "Facing Up to the Problem of Consciousness". In: *Journal of Consciousness Studies* 2.3, pp. 200–219.

Chalmers, D. J. (1996a). "Does a rock implement every finite-state automaton?" In: *Synthese* 108.3, pp. 309–333. DOI: 10.1007/bf00413692.

Chalmers, D. J. (1996b). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Chalmers, D. J. (2011). "A Computational Foundation for the Study of Cognition". In: *Journal of Cognitive Science* 12.4, pp. 325–359. DOI: 10.17791/jcs.2011.12.4.325.

Chialvo, D. R. (2010). "Emergent complex neural dynamics". In: *Nature Physics* 6.10, pp. 744–750. DOI: 10.1038/nphys1803.

Clark, A. (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science". In: *Behavioral and Brain Sciences* 36.3, pp. 181–204. DOI: 10.1017/s0140525x12000477.

Clark, A. and J. Toribio (1994). "Doing without representing?" In: *Synthese* 101.3, pp. 401–431. DOI: 10.1007/bf01063896.

Cogitate Consortium et al. (2025). "Adversarial testing of global neuronal workspace and integrated information theories of consciousness". In: *Nature* 642.8066, pp. 133–142. DOI: 10.1038/s41586-025-08888-1.

Cohen, M. X. (2014). *Analyzing Neural Time Series Data: Theory and Practice*. Cambridge, MA: MIT Press.

Cole, S. R. and B. Voytek (2017). "Brain Oscillations and the Importance of Waveform Shape". In: *Trends in Cognitive Sciences* 21.2, pp. 137–149. DOI: 10.1016/j.tics.2016.12.008.

Colombo, M. and C. Wright (2021). "First principles in the life sciences: the free-energy principle, organicism, and mechanism". In: *Synthese* 198.S14, pp. 3463–3488. DOI: 10.1007/s11229-018-01932-w.

Conant, R. C. and W. Ross Ashby (1970). "Every good regulator of a system must be a model of that system". In: *International Journal of Systems Science* 1.2, pp. 89–97. DOI: 10.1080/00207727008920220.

Copeland, B. J. (1996). "What is computation?" In: *Synthese* 108.3, pp. 335–359. DOI: 10.1007/bf00413693.

Couey, J. J., A. Witoelar, S.-J. Zhang, K. Zheng, J. Ye, B. Dunn, R. Czajkowski, M.-B. Moser, E. I. Moser, Y. Roudi, and M. P. Witter (2013). "Recurrent inhibitory circuitry as a mechanism for grid formation". In: *Nature Neuroscience* 16.3, pp. 318–324. DOI: 10.1038/nn.3310.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory*. 2nd ed. Hoboken, NJ: Wiley.

Cowan, N. J., M. M. Ankarali, J. P. Dyhr, M. S. Madhav, E. Roth, S. Sefati, S. Sponberg, S. A. Stamper, E. S. Fortune, and T. L. Daniel (2014). "Feedback Control as a Framework for Understanding Tradeoffs in Biology". In: *Integrative and Comparative Biology* 54.2, pp. 223–237. DOI: 10.1093/icb/icu050.

Cowey, A. and P. Stoerig (1995). "Blindsight in monkeys". In: *Nature* 373.6511, pp. 247–249. DOI: 10.1038/373247a0.

Coxeter, H. S. M. (2003). *Projective Geometry*. 2nd ed. New York: Springer.

Craig, A. D. (2002). "How do you feel? Interoception: the sense of the physiological condition of the body". In: *Nature Reviews Neuroscience* 3.8, pp. 655–666. DOI: 10.1038/nrn894.

Craik, K. J. W. (1943). *The Nature of Explanation*. Cambridge: Cambridge University Press.

Crapse, T. B. and M. A. Sommer (2008). "Corollary discharge across the animal kingdom". In: *Nature Reviews Neuroscience* 9.8, pp. 587–600. DOI: 10.1038/nrn2457.

Crick, F. and C. Koch (1990). "Towards a Neurobiological Theory of Consciousness". In: *Seminars in the Neurosciences* 2, pp. 263–275.

Crutchfield, J. P. and K. Young (1989). "Inferring statistical complexity". In: *Physical Review Letters* 63.2, pp. 105–108. DOI: 10.1103/physrevlett.63.105.

Cummins, R. (1996). *Representations, Targets, and Attitudes*. Cambridge, MA: MIT Press.

Damasio, A. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt Brace.

Dean, C. R., L. Wang, P. Maher, C. Forsythe, F. Ghahari, Y. Gao, J. Katoch, M. Ishigami, P. Moon, M. Koshino, T. Taniguchi, K. Watanabe, K. L. Shepard, J. Hone, and P. Kim (2013). "Hofstadter's butterfly and the fractal quantum Hall effect in moiré superlattices". In: *Nature* 497.7451, pp. 598–602. DOI: 10.1038/nature12186.

Deco, G. and M. L. Kringelbach (2016). "Metastability and Coherence: Extending the Communication through Coherence Hypothesis Using A Whole-Brain Computational Perspective". In: *Trends in Neurosciences* 39.3, pp. 125–135. DOI: 10.1016/j.tins.2016.01.001.

Dehaene, S. and J.-P. Changeux (2011). "Experimental and Theoretical Approaches to Conscious Processing". In: *Neuron* 70.2, pp. 200–227. DOI: 10.1016/j.neuron.2011.03.018.

Dehaene, S., M. Kerszberg, and J.-P. Changeux (1998). "A neuronal model of a global workspace in effortful cognitive tasks". In: *Proceedings of the National Academy of Sciences* 95.24, pp. 14529–14534. DOI: 10.1073/pnas.95.24.14529.

Dennett, D. C. (1991). *Consciousness Explained*. Boston: Little, Brown.

Dennett, D. C. (1995). *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon & Schuster.

Dennett, D. C. (1996). *Kinds of Minds: Toward an Understanding of Consciousness*. New York: Basic Books.

Dennett, D. C. (2003). "Who's On First? Heterophenomenology Explained". In: *Journal of Consciousness Studies* 10.9–10, pp. 19–30.

Dennett, D. C. (2016). "Illusionism as the Obvious Default Theory of Consciousness". In: *Journal of Consciousness Studies* 23.11–12, pp. 65–72.

Desimone, R. and J. Duncan (1995). "Neural Mechanisms of Selective Visual Attention". In: *Annual Review of Neuroscience* 18.1, pp. 193–222. DOI: 10.1146/annurev.ne.18.030195.001205.

Dewhurst, J. and Philosophy Documentation Center (2021). "Causal emergence from effective information: Neither causal nor emergent?" In: *Thought: A Journal of Philosophy* 10.3, pp. 158–168. DOI: 10.1002/tht3.489.

Di Paolo, E. A. (2005). "Autopoiesis, Adaptivity, Teleology, Agency". In: *Phenomenology and the Cognitive Sciences* 4.4, pp. 429–452. DOI: 10.1007/s11097-005-9002-y.

Dicke, R. H. (1946). "The Measurement of Thermal Radiation at Microwave Frequencies". In: *Review of Scientific Instruments* 17.7, pp. 268–275. DOI: 10.1063/1.1770483.

Domnisoru, C., A. A. Kinkhabwala, and D. W. Tank (2013). "Membrane potential dynamics of grid cells". In: *Nature* 495.7440, pp. 199–204. DOI: 10.1038/nature11973.

Donoghue, T., M. Haller, E. J. Peterson, P. Varma, P. Sebastian, R. Gao, T. Noto, A. H. Lara, J. D. Wallis, R. T. Knight, A. Shestyuk, and B. Voytek (2020). "Parameterizing neural power spectra into periodic and aperiodic components". In: *Nature Neuroscience* 23.12, pp. 1655–1665. DOI: 10.1038/s41593-020-00744-x.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Dretske, F. I. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Eagleman, D. M. and T. J. Sejnowski (2000). "Motion Integration and Postdiction in Visual Awareness". In: *Science* 287.5460, pp. 2036–2038. DOI: 10.1126/science.287.5460.2036.

Eberhardt, F. and L. L. Lee (2022). "Causal Emergence: When Distortions in a Map Obscure the Territory". In: *Philosophies* 7.2, p. 30. DOI: 10.3390/philosophies7020030.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Egbert, M. D., X. E. Barandiaran, and E. A. Di Paolo (2010). "A Minimal Model of Metabolism-Based Chemotaxis". In: *PLoS Computational Biology* 6.12, e1001004. DOI: 10.1371/journal.pcbi.1001004.

Ehrsson, H. H. (2007). "The Experimental Induction of Out-of-Body Experiences". In: *Science* 317.5841, p. 1048. DOI: 10.1126/science.1142175.

Emde, G. von der (1999). "Active electrolocation of objects in weakly electric fish". In: *Journal of Experimental Biology* 202.10, pp. 1205–1215. DOI: 10.1242/jeb.202.10.1205.

Enikolopov, A. G., L. Abbott, and N. B. Sawtell (2018). "Internally Generated Predictions Enhance Neural and Behavioral Detection of Sensory Stimuli in an Electric Fish". In: *Neuron* 99.1, 135–146.e3. DOI: 10.1016/j.neuron.2018.06.006.

Ermentrout, G. B. and D. H. Terman (2010). *Mathematical Foundations of Neuroscience*. New York: Springer.

Ernst, M. O. and M. S. Banks (2002). "Humans integrate visual and haptic information in a statistically optimal fashion". In: *Nature* 415.6870, pp. 429–433. DOI: 10.1038/415429a.

Faraday, M. (1831). "On a peculiar class of acoustical figures; and on certain forms assumed by groups of particles upon vibrating elastic surfaces". In: *Philosophical Transactions of the Royal Society of London* 121.121, pp. 299–340. DOI: 10.1098/rstl.1831.0018.

Feinberg, T. E. and J. M. Mallatt (2016). *The Ancient Origins of Consciousness: How the Brain Created Experience*. Cambridge, MA: MIT Press.

Flavell, J. H. (1979). "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry". In: *American Psychologist* 34.10, pp. 906–911. DOI: 10.1037/0003-066x.34.10.906.

Fleming, S. M., R. J. Dolan, and C. D. Frith (2012). "Metacognition: computation, biology and function". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1594, pp. 1280–1286. DOI: 10.1098/rstb.2012.0021.

Fodor, J. A. (1974). "Special sciences (or: The disunity of science as a working hypothesis)". In: *Synthese* 28.2, pp. 97–115. DOI: 10.1007/bf00485230.

Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (1990). *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Francis, B. and W. Wonham (1976). "The internal model principle of control theory". In: *Automatica* 12.5, pp. 457–465. DOI: 10.1016/0005-1098(76)90006-6.

Frankish, K. (2016). "Illusionism as a Theory of Consciousness". In: *Journal of Consciousness Studies* 23.11–12, pp. 11–39.

Fries, P. (2005). "A mechanism for cognitive dynamics: neuronal communication through neuronal coherence". In: *Trends in Cognitive Sciences* 9.10, pp. 474–480. DOI: 10.1016/j.tics.2005.08.011.

Fries, P. (2015). "Rhythms for Cognition: Communication through Coherence". In: *Neuron* 88.1, pp. 220–235. DOI: 10.1016/j.neuron.2015.09.034.

Friston, K. (2010). "The free-energy principle: a unified brain theory?" In: *Nature Reviews Neuroscience* 11.2, pp. 127–138. DOI: 10.1038/nrn2787.

Friston, K. J. (1997). "Transients, Metastability, and Neuronal Dynamics". In: *NeuroImage* 5.2, pp. 164–171. DOI: 10.1006/nimg.1997.0259.

Frith, C. D., S.-J. Blakemore, and D. M. Wolpert (2000). "Abnormalities in the awareness and control of action". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 355.1404, pp. 1771–1788. DOI: 10.1098/rstb.2000.0734.

Fuhs, M. C. and D. S. Touretzky (2006). "A Spin Glass Model of Path Integration in Rat Medial Entorhinal Cortex". In: *The Journal of Neuroscience* 26.16, pp. 4266–4276. DOI: 10.1523/jneurosci.4353-05.2006.

Gallagher, S. (2000). "Philosophical conceptions of the self: implications for cognitive science". In: *Trends in Cognitive Sciences* 4.1, pp. 14–21. DOI: 10.1016/s1364-6613(99)01417-5.

Gallant, J. R., L. L. Traeger, J. D. Volkening, H. Moffett, P.-H. Chen, C. D. Novina, G. N. Phillips, R. Anand, G. B. Wells, M. Pinch, R. Güth, G. A. Unguez, J. S. Albert, H. H. Zakon, M. P. Samanta, and M. R. Sussman (2014). "Genomic basis for the convergent evolution of electric organs". In: *Science* 344.6191, pp. 1522–1525. DOI: 10.1126/science.1254432.

Gärdenfors, P. (1996). "Cued and detached representations in animal cognition". In: *Behavioural Processes* 35.1–3, pp. 263–273. DOI: 10.1016/0376-6357(95)00043-7.

Gardner, R. J., E. Hermansen, M. Pachitariu, Y. Burak, N. A. Baas, B. A. Dunn, M.-B. Moser, and E. I. Moser (2022). "Toroidal topology of population activity in grid cells". In: *Nature* 602.7895, pp. 123–128. DOI: 10.1038/s41586-021-04268-7.

Giacino, J., S. Ashwal, N. Childs, R. Cranford, B. Jennett, D. Katz, J. Kelly, J. Rosenberg, J. Whyte, R. Zafonte, and N. Zasler (2002). "The minimally conscious state". In: *Neurology* 58.3, pp. 349–353. DOI: 10.1212/wnl.58.3.349.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Ginsburg, S. and E. Jablonka (2019). *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*. Cambridge, MA: MIT Press.

Giocomo, L. M., M.-B. Moser, and E. I. Moser (2011). "Computational Models of Grid Cells". In: *Neuron* 71.4, pp. 589–603. DOI: 10.1016/j.neuron.2011.07.023.

Glass, L. and M. C. Mackey (1988). *From Clocks to Chaos: The Rhythms of Life*. Princeton: Princeton University Press.

Godfrey-Smith, P. (2009). "Triviality arguments against functionalism". In: *Philosophical Studies* 145.2, pp. 273–295. DOI: 10.1007/s11098-008-9231-3.

Godfrey-Smith, P. (2016). *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. New York: Farrar, Straus and Giroux.

Goldberg, J. M. and P. B. Brown (1969). "Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization". In: *Journal of Neurophysiology* 32.4, pp. 613–636. DOI: 10.1152/jn.1969.32.4.613.

Goldstein, J. L. (1967). "Auditory Nonlinearity". In: *The Journal of the Acoustical Society of America* 41.3, pp. 676–699. DOI: 10.1121/1.1910396.

Gramfort, A. (2013). "MEG and EEG data analysis with MNE-Python". In: *Frontiers in Neuroscience* 7. DOI: 10.3389/fnins.2013.00267.

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3, pp. 424–438. DOI: 10.2307/1912791.

Graziano, M. S. A. (2013). *Consciousness and the Social Brain*. New York: Oxford University Press.

Graziano, M. S. A. and T. W. Webb (2015). "The attention schema theory: a mechanistic account of subjective awareness". In: *Frontiers in Psychology* 06. DOI: 10.3389/fpsyg.2015.00500.

Green, D. M. and J. A. Swets (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Grilli, J., G. Barabás, M. J. Michalska-Smith, and S. Allesina (2017). "Higher-order interactions stabilize dynamics in competitive network models". In: *Nature* 548.7666, pp. 210–213. DOI: 10.1038/nature23273.

Grush, R. (2004). "The emulation theory of representation: Motor control, imagery, and perception". In: *Behavioral and Brain Sciences* 27.3, pp. 377–396. DOI: 10.1017/s0140525x04000093.

Guckenheimer, J. (1975). "Isochrons and phaseless sets". In: *Journal of Mathematical Biology* 1.3, pp. 259–273. DOI: 10.1007/bf01273747.

Guckenheimer, J. and P. Holmes (1983). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. New York: Springer.

Guevara, M. R., L. Glass, and A. Shrier (1981). "Phase Locking, Period-Doubling Bifurcations, and Irregular Dynamics in Periodically Stimulated Cardiac Cells". In: *Science* 214.4527, pp. 1350–1353. DOI: 10.1126/science.7313693.

Guo, Y.-X. and M. Kawasaki (1997). "Representation of Accurate Temporal Information in the Electrosensory System of the African Electric Fish, *Gymnarchus niloticus*". In: *The Journal of Neuroscience* 17.5, pp. 1761–1768. DOI: 10.1523/jneurosci.17-05-01761.1997.

Ha, D. and J. Schmidhuber (2018). *World Models*. arXiv: 1803.10122.

Hafting, T., M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser (2005). "Microstructure of a spatial map in the entorhinal cortex". In: *Nature* 436.7052, pp. 801–806. DOI: 10.1038/nature03721.

Haggard, P. (2017). "Sense of agency in the human brain". In: *Nature Reviews Neuroscience* 18.4, pp. 196–207. DOI: 10.1038/nrn.2017.14.

Haken, H. (1975). "Analogy between higher instabilities in fluids and lasers". In: *Physics Letters A* 53.1, pp. 77–78. DOI: 10.1016/0375-9601(75)90353-9.

Haken, H., J. A. S. Kelso, and H. Bunz (1985). "A theoretical model of phase transitions in human hand movements". In: *Biological Cybernetics* 51.5, pp. 347–356. DOI: 10.1007/bf00336922.

Haken, H. (1983a). *Advanced Synergetics: Instability Hierarchies of Self-Organizing Systems and Devices*. Berlin: Springer.

Haken, H. (1983b). *Synergetics: An Introduction*. 3rd ed. Berlin: Springer.

Hämäläinen, M., R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa (1993). "Magnetoencephalography—theory, instrumentation, and applica-

tions to noninvasive studies of the working human brain". In: *Reviews of Modern Physics* 65.2, pp. 413–497. DOI: 10.1103/revmodphys.65.413.

Harder, M., C. Salge, and D. Polani (2013). "Bivariate measure of redundant information". In: *Physical Review E* 87.1, 012130. DOI: 10.1103/physreve.87.012130.

Hauser, M. D., N. Chomsky, and W. T. Fitch (2002). "The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?" In: *Science* 298.5598, pp. 1569–1579. DOI: 10.1126/science.298.5598.1569.

Haykin, S. (2005). "Cognitive radio: brain-empowered wireless communications". In: *IEEE Journal on Selected Areas in Communications* 23.2, pp. 201–220. DOI: 10.1109/jsac.2004.839380.

Heiligenberg, W. and J. Bastian (1984). "The Electric Sense of Weakly Electric Fish". In: *Annual Review of Physiology* 46.1, pp. 561–583. DOI: 10.1146/annurev.ph.46.030184.003021.

Heiligenberg, W. and G. Rose (1985). "Phase and amplitude computations in the midbrain of an electric fish: intracellular studies of neurons participating in the jamming avoidance response of *Eigenmannia*". In: *The Journal of Neuroscience* 5.2, pp. 515–531. DOI: 10.1523/jneurosci.05-02-00515.1985.

Heiligenberg, W. (1980). "The jamming avoidance response in the weakly electric fish *Eigenmannia*". In: *Naturwissenschaften* 67.10, pp. 499–507. DOI: 10.1007/bf01047630.

Heiligenberg, W. (1991a). *Neural Nets in Electric Fish*. Cambridge, MA: MIT Press.

Heiligenberg, W. (1991b). "The jamming avoidance response of the electric fish, *Eigenmannia*: computational rules and their neuronal implementation". In: *Seminars in Neuroscience* 3.1, pp. 3–18. DOI: 10.1016/1044-5765(91)90062-s.

Heiligenberg, W., C. Baker, and J. Matsubara (1978). "The Jamming Avoidance Response in *Eigenmannia* revisited: The structure of a neuronal democracy". In: *Journal of Comparative Physiology ? A* 127.3, pp. 267–286. DOI: 10.1007/bf01350117.

Heiligenberg, W. and G. Rose (1986). "Gating of sensory information: Joint computations of phase and amplitude data in the midbrain of the electric fish,*Eigenmannia*". In: *Journal of Comparative Physiology A* 159.3, pp. 311–324. DOI: 10.1007/bf00603977.

Held, R. and A. Hein (1963). "Movement-produced stimulation in the development of visually guided behavior". In: *Journal of Comparative and Physiological Psychology* 56.5, pp. 872–876. DOI: 10.1037/h0040546.

Helfrich, R. F., H. Knepper, G. Nolte, D. Strüber, S. Rach, C. S. Herrmann, T. R. Schneider, and A. K. Engel (2014). "Selective Modulation of Interhemispheric Functional Connectivity by HD-tACS Shapes Perception". In: *PLoS Biology* 12.12, e1002031. DOI: 10.1371/journal.pbio.1002031.

Helmholtz, H. v. (1867). *Handbuch der physiologischen Optik*. Leipzig: Leopold Voss.

Helmholtz, H. v. (1885). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Trans. by A. J. Ellis. 2nd ed. London: Longmans, Green.

Herrmann, C. S. (2001). "Human EEG responses to 1–100 Hz flicker: resonance phenomena in visual cortex and their potential correlation to cognitive phenomena". In: *Experimental Brain Research* 137.3–4, pp. 346–353. DOI: 10.1007/s002210100682.

Herrmann, C. S., M. M. Murray, S. Ionta, A. Hutt, and J. Lefebvre (2016). "Shaping Intrinsic Neural Oscillations with Periodic Stimulation". In: *The Journal of Neuroscience* 36.19, pp. 5328–5337. DOI: 10.1523/jneurosci.0236-16.2016.

Herzog, M. H., T. Kammer, and F. Scharnowski (2016). "Time Slices: What Is the Duration of a Percept?" In: *PLOS Biology* 14.4, e1002433. DOI: 10.1371/journal.pbio.1002433.

Hoel, E. (2017). "When the Map Is Better Than the Territory". In: *Entropy* 19.5, p. 188. DOI: 10.3390/e19050188.

Hoel, E. P., L. Albantakis, and G. Tononi (2013). "Quantifying causal emergence shows that macro can beat micro". In: *Proceedings of the National Academy of Sciences* 110.49, pp. 19790–19795. DOI: 10.1073/pnas.1314922110.

Hofstadter, D. R. (2007). *I Am a Strange Loop*. New York: Basic Books.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Holst, E. von and H. Mittelstaedt (1950). "Das Reafferenzprinzip". In: *Naturwissenschaften* 37.20, pp. 464–476. DOI: 10.1007/bf00622503.

Horowitz, P. and W. Hill (2015). *The Art of Electronics*. 3rd ed. Cambridge: Cambridge University Press.

Horsman, D., S. Stepney, R. C. Wagner, and V. Kendon (2014). "When does a physical system compute?" In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 470.2169, 20140182. DOI: 10.1098/rspa.2014.0182.

Husserl, E. (1991). *On the Phenomenology of the Consciousness of Internal Time (1893–1917)*. Trans. by J. B. Brough. Dordrecht: Kluwer.

Hutcheon, B. and Y. Yarom (2000). "Resonance, oscillation and the intrinsic frequency preferences of neurons". In: *Trends in Neurosciences* 23.5, pp. 216–222. DOI: 10.1016/s0166-2236(00)01547-2.

Iacopini, I., G. Petri, A. Barrat, and V. Latora (2019). "Simplicial models of social contagion". In: *Nature Communications* 10.1, 2485. DOI: 10.1038/s41467-019-10431-6.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Izhikevich, E. M. (2001). "Resonate-and-fire neurons". In: *Neural Networks* 14.6–7, pp. 883–894. DOI: 10.1016/s0893-6080(01)00078-8.

Izhikevich, E. M. (2007). *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. Cambridge, MA: MIT Press.

Jackson, F. (1982). "Epiphenomenal Qualia". In: *The Philosophical Quarterly* 32.127, pp. 127–136. DOI: 10.2307/2960077.

James, R. G., C. J. Ellison, and J. P. Crutchfield (2011). "Anatomy of a bit: Information in a time series observation". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 21.3, 037109. DOI: 10.1063/1.3637494.

James, W. (1890). *The Principles of Psychology*. New York: Henry Holt.

Jensen, M. H., P. Bak, and T. Bohr (1983). "Complete Devil's Staircase, Fractal Dimension, and Universality of Mode- Locking Structure in the Circle Map". In: *Physical Review Letters* 50.21, pp. 1637–1639. DOI: 10.1103/physrevlett.50.1637.

Johnson, A. and A. D. Redish (2007). "Neural Ensembles in CA3 Transiently Encode Paths Forward of the Animal at a Decision Point". In: *The Journal of Neuroscience* 27.45, pp. 12176–12189. DOI: 10.1523/jneurosci.3761-07.2007.

Jonas, H. (1966). *The Phenomenon of Life: Toward a Philosophical Biology*. New York: Harper & Row.

Kaplan, D. (1989). "Demonstratives". In: *Themes from Kaplan*. Ed. by J. Almog, J. Perry, and H. Wettstein. New York: Oxford University Press, pp. 481–563.

Kaschube, M., M. Schnabel, S. Löwel, D. M. Coppola, L. E. White, and F. Wolf (2010). "Universality in the Evolution of Orientation Columns in the Visual Cortex". In: *Science* 330.6007, pp. 1113–1116. DOI: 10.1126/science.1194869.

Kawasaki, M. and Y. Guo (1996). "Neuronal circuitry for comparison of timing in the electrosensory lateral line lobe of the African wave-type electric fish *Gymnarchus niloticus*". In: *The Journal of Neuroscience* 16.1, pp. 380–391. DOI: 10.1523/jneurosci.16-01-00380.1996.

Kawasaki, M. and W. Heiligenberg (1990). "Different classes of glutamate receptors and GABA mediate distinct modulations of a neuronal oscillator, the medullary pacemaker of a gymnotiform electric fish". In: *The Journal of Neuroscience* 10.12, pp. 3896–3904. DOI: 10.1523/jneurosci.10-12-03896.1990.

Kawasaki, M. (1993). "Independently evolved jamming avoidance responses employ identical computational algorithms: a behavioral study of the African electric fish, *Gymnarchus niloticus*". In: *Journal of Comparative Physiology A* 173.1, pp. 9–22. DOI: 10.1007/bf00209614.

Kawasaki, M. (1994). "The African wave-type electric fish, *Gymnarchus niloticus*, lacks corollary discharge mechanisms for electrosensory gating". In: *Journal of Comparative Physiology A* 174.2, pp. 133–144. DOI: 10.1007/bf00193781.

Kawasaki, M. (1996). "Comparative Analysis of the Jamming Avoidance Response in African and South American Wave-Type Electric Fishes". In: *The Biological Bulletin* 191.1, pp. 103–108. DOI: 10.2307/1543070.

Kawasaki, M., L. Maler, G. J. Rose, and W. Heiligenberg (1988). "Anatomical and functional organization of the prepacemaker nucleus in gymnotiform electric fish: The accommodation of two behaviors in one nucleus". In: *Journal of Comparative Neurology* 276.1, pp. 113–131. DOI: 10.1002/cne.902760108.

Kawasaki, M., G. Rose, and W. Heiligenberg (1988). "Temporal hyperacuity in single neurons of electric fish". In: *Nature* 336.6195, pp. 173–176. DOI: 10.1038/336173a0.

Kawato, M. (1999). "Internal models for motor control and trajectory planning". In: *Current Opinion in Neurobiology* 9.6, pp. 718–727. DOI: 10.1016/s0959-4388(99)00028-8.

Keller, C. H., L. Maler, and W. Heiligenberg (1990). "Structural and functional organization of a diencephalic sensory-motor interface in the gymnotiform fish, *Eigenmannia*". In: *Journal of Comparative Neurology* 293.3, pp. 347–376. DOI: 10.1002/cne.902930304.

Kelso, J. A. (1984). "Phase transitions and critical behavior in human bimanual coordination". In: *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 246.6, R1000–R1004. DOI: 10.1152/ajpregu.1984.246.6.r1000.

Kelso, J. A. S., J. P. Scholz, and G. Schöner (1986). "Nonequilibrium phase transitions in coordinated biological motion: critical fluctuations". In: *Physics Letters A* 118.6, pp. 279–284. DOI: 10.1016/0375-9601(86)90359-2.

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press.

Kemp, D. T. (1978). "Stimulated acoustic emissions from within the human auditory system". In: *The Journal of the Acoustical Society of America* 64.5, pp. 1386–1391. DOI: 10.1121/1.382104.

Kentridge, R. W., C. A. Heywood, and L. Weiskrantz (1999). "Attention without awareness in blindsight". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 266.1430, pp. 1805–1811. DOI: 10.1098/rspb.1999.0850.

Kepecs, A., N. Uchida, H. A. Zariwala, and Z. F. Mainen (2008). "Neural correlates, computation and behavioural impact of decision confidence". In: *Nature* 455.7210, pp. 227–231. DOI: 10.1038/nature07200.

Key, B. (2016). "Why fish do not feel pain". In: *Animal Sentience* 1.3. DOI: 10.51291/2377-7478.1011.

Kiani, R. and M. N. Shadlen (2009). "Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex". In: *Science* 324.5928, pp. 759–764. DOI: 10.1126/science.1169405.

Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind–Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Klimesch, W. (2018). "The frequency architecture of brain and brain body oscillations: an analysis". In: *European Journal of Neuroscience* 48.7, pp. 2431–2453. DOI: 10.1111/ejn.14192.

Knill, D. C. and A. Pouget (2004). "The Bayesian brain: the role of uncertainty in neural coding and computation". In: *Trends in Neurosciences* 27.12, pp. 712–719. DOI: 10.1016/j.tins.2004.10.007.

Koch, C., M. Massimini, M. Boly, and G. Tononi (2016). "Neural correlates of consciousness: progress and problems". In: *Nature Reviews Neuroscience* 17.5, pp. 307–321. DOI: 10.1038/nrn.2016.22.

Koch, C. and N. Tsuchiya (2007). "Attention and consciousness: two distinct brain processes". In: *Trends in Cognitive Sciences* 11.1, pp. 16–22. DOI: 10.1016/j.tics.2006.10.012.

Koenig, J., A. N. Linder, J. K. Leutgeb, and S. Leutgeb (2011). "The Spatial Periodicity of Grid Cells Is Not Sustained During Reduced Theta Oscillations". In: *Science* 332.6029, pp. 592–595. DOI: 10.1126/science.1201685.

Kolata, G. B. (1977). "Catastrophe Theory: The Emperor Has No Clothes". In: *Science* 196.4287, pp. 287, 350–351. DOI: 10.1126/science.196.4287.287.

Kramer, M. A., A. B. L. Tort, and N. J. Kopell (2008). "Sharp edge artifacts and spurious coupling in EEG frequency comodulation measures". In: *Journal of Neuroscience Methods* 170.2, pp. 352–357. DOI: 10.1016/j.jneumeth.2008.01.020.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kuramoto, Y. (1975). "Self-entrainment of a population of coupled non-linear oscillators". In: *International Symposium on Mathematical Problems in Theoretical Physics*. Ed. by H. Araki. Vol. 39. Lecture Notes in Physics. Berlin: Springer, pp. 420–422. DOI: 10.1007/bfb0013365.

Kuramoto, Y. (1984). *Chemical Oscillations, Waves, and Turbulence*. Berlin: Springer.

Kuramoto, Y. and D. Battogtokh (2002). "Coexistence of Coherence and Incoherence in Nonlocally Coupled Phase Oscillators". In: *Nonlinear Phenomena in Complex Systems* 5.4, pp. 380–385.

Kuwada, S., T. C. T. Yin, and R. E. Wickesberg (1979). "Response of Cat Inferior Colliculus Neurons to Binaural Beat Stimuli: Possible Mechanisms for Sound Localization". In: *Science* 206.4418, pp. 586–588. DOI: 10.1126/science.493964.

Kuznetsov, Y. A. (2004). *Elements of Applied Bifurcation Theory*. 3rd ed. New York: Springer.

Lachaux, J.-P., E. Rodriguez, J. Martinerie, and F. J. Varela (1999). "Measuring phase synchrony in brain signals". In: *Human Brain Mapping* 8.4, pp. 194–208. DOI: 10.1002/(sici)1097-0193(1999)8:4<194::aid-hbm4>3.0.co;2-c.

Ladyman, J. and D. Ross (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Lakatos, I. (1970). "Falsification and the Methodology of Scientific Research Programmes". In: *Criticism and the Growth of Knowledge*. Cambridge University Press, pp. 91–196. DOI: 10.1017/cbo9781139171434.009.

Lamme, V. A. (2006). "Towards a true neural stance on consciousness". In: *Trends in Cognitive Sciences* 10.11, pp. 494–501. DOI: 10.1016/j.tics.2006.09.001.

Lamme, V. A. and P. R. Roelfsema (2000). "The distinct modes of vision offered by feedforward and recurrent processing". In: *Trends in Neurosciences* 23.11, pp. 571–579. DOI: 10.1016/s0166-2236(00)01657-x.

Lancaster, G., D. Iatsenko, A. Pidde, V. Ticcinelli, and A. Stefanovska (2018). "Surrogate data for hypothesis testing of physical systems". In: *Physics Reports* 748, pp. 1–60. DOI: 10.1016/j.physrep.2018.06.001.

Landau, L. D. and E. M. Lifshitz (1976). *Mechanics*. 3rd ed. Course of Theoretical Physics 1. Oxford: Pergamon.

Landauer, R. (1991). "Information is Physical". In: *Physics Today* 44.5, pp. 23–29. DOI: 10.1063/1.881299.

Langton, C. G. (1990). "Computation at the edge of chaos: Phase transitions and emergent computation". In: *Physica D: Nonlinear Phenomena* 42.1–3, pp. 12–37. DOI: 10.1016/0167-2789(90)90064-v.

Lau, H. and D. Rosenthal (2011). "Empirical support for higher-order theories of conscious awareness". In: *Trends in Cognitive Sciences* 15.8, pp. 365–373. DOI: 10.1016/j.tics.2011.05.009.

Lavoué, S., M. Miya, M. E. Arnegard, J. P. Sullivan, C. D. Hopkins, and M. Nishida (2012). "Comparable Ages for the Independent Origins of Electrogenesis in African and South American Weakly Electric Fishes". In: *PLoS ONE* 7.5, e36287. DOI: 10.1371/journal.pone.0036287.

Lee, B. S. (1950). "Effects of Delayed Speech Feedback". In: *The Journal of the Acoustical Society of America* 22.6, pp. 824–826. DOI: 10.1121/1.1906696.

Lendner, J. D., R. F. Helfrich, B. A. Mander, L. Romundstad, J. J. Lin, M. P. Walker, P. G. Larsson, and R. T. Knight (2020). "An electrophysiological marker of arousal level in humans". In: *eLife* 9, e55092. DOI: 10.7554/elife.55092.

Lenggenhager, B., T. Tadi, T. Metzinger, and O. Blanke (2007). "Video Ergo Sum: Manipulating Bodily Self-Consciousness". In: *Science* 317.5841, pp. 1096–1099. DOI: 10.1126/science.1143439.

León, I. and D. Pazó (2019). "Phase reduction beyond the first order: The case of the mean-field complex Ginzburg-Landau equation". In: *Physical Review E* 100.1, 012211. DOI: 10.1103/physreve.100.012211.

Lettvin, J. Y., H. R. Maturana, W. S. McCulloch, and W. H. Pitts (1959). "What the Frog's Eye Tells the Frog's Brain". In: *Proceedings of the IRE* 47.11, pp. 1940–1951. DOI: 10.1109/jrproc.1959.287207.

Levin, M. (2019). "The Computational Boundary of a "Self": Developmental Bioelectricity Drives Multicellularity and Scale-Free Cognition". In: *Frontiers in Psychology* 10, 2688. DOI: 10.3389/fpsyg.2019.02688.

Levin, M. (2022). "Technological Approach to Mind Everywhere: An Experimentally-Grounded Framework for Understanding Diverse Bodies and Minds". In: *Frontiers in Systems Neuroscience* 16, 768201. DOI: 10.3389/fnsys.2022.768201.

Levine, J. (1983). "Materialism and qualia: the explanatory gap". In: *Pacific Philosophical Quarterly* 64.4, pp. 354–361. DOI: 10.1111/j.1468-0114.1983.tb00207.x.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1979). "Attitudes De Dicto and De Se". In: *The Philosophical Review* 88.4, pp. 513–543. DOI: 10.2307/2184843.

Licklider, J. C. R., J. C. Webster, and J. M. Hedlun (1950). "On the Frequency Limits of Binaural Beats". In: *The Journal of the Acoustical Society of America* 22.4, pp. 468–473. DOI: 10.1121/1.1906629.

Lim, L.-H. (2020). "Hodge Laplacians on Graphs". In: *SIAM Review* 62.3, pp. 685–715. DOI: 10.1137/18m1223101.

Lissmann, H. W. (1951). "Continuous Electrical Signals from the Tail of a Fish, *Gymnarchus niloticus* Cuv". In: *Nature* 167.4240, pp. 201–202. DOI: 10.1038/167201a0.

Lissmann, H. W. (1958). "On the Function and Evolution of Electric Organs in Fish". In: *Journal of Experimental Biology* 35.1, pp. 156–191. DOI: 10.1242/jeb.35.1.156.

Lissmann, H. W. and K. E. Machin (1958). "The Mechanism of Object Location in *Gymnarchus* Niloticus and Similar Fish". In: *Journal of Experimental Biology* 35.2, pp. 451–486. DOI: 10.1242/jeb.35.2.451.

Lopes dos Santos, J. M. B., N. M. R. Peres, and A. H. Castro Neto (2007). "Graphene Bilayer with a Twist: Electronic Structure". In: *Physical Review Letters* 99.25, 256802. DOI: 10.1103/physrevlett.99.256802.

Luisi, P. L. (2003). "Autopoiesis: a review and a reappraisal". In: *Naturwissenschaften* 90.2, pp. 49–59. DOI: 10.1007/s00114-002-0389-9.

Lycan, W. G. (1996). *Consciousness and Experience*. Cambridge, MA: MIT Press.

Lyon, P. (2006). "The biogenic approach to cognition". In: *Cognitive Processing* 7.1, pp. 11–29. DOI: 10.1007/s10339-005-0016-8.

Maass, W., T. Natschläger, and H. Markram (2002). "Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations". In: *Neural Computation* 14.11, pp. 2531–2560. DOI: 10.1162/089976602760407955.

Macnab, R. M. and D. E. Koshland (1972). "The Gradient-Sensing Mechanism in Bacterial Chemotaxis". In: *Proceedings of the National Academy of Sciences* 69.9, pp. 2509–2512. DOI: 10.1073/pnas.69.9.2509.

Madhav, M. S., S. A. Stamper, E. S. Fortune, and N. J. Cowan (2013). "Closed-loop stabilization of the jamming avoidance response reveals its locally unstable and globally nonlinear dynamics". In: *Journal of Experimental Biology* 216.22, pp. 4272–4284. DOI: 10.1242/jeb.088922.

Maris, E. and R. Oostenveld (2007). "Nonparametric statistical testing of EEG- and MEG-data". In: *Journal of Neuroscience Methods* 164.1, pp. 177–190. DOI: 10.1016/j.jneumeth.2007.03.024.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman.

Martens, E. A., S. Thutupalli, A. Fourrière, and O. Hallatschek (2013). "Chimera states in mechanical oscillator networks". In: *Proceedings of the National Academy of Sciences* 110.26, pp. 10563–10567. DOI: 10.1073/pnas.1302880110.

Mashour, G. A., P. Roelfsema, J.-P. Changeux, and S. Dehaene (2020). "Conscious Processing and the Global Neuronal Workspace Hypothesis". In: *Neuron* 105.5, pp. 776–798. DOI: 10.1016/j.neuron.2020.01.026.

Massimini, M., F. Ferrarelli, R. Huber, S. K. Esser, H. Singh, and G. Tononi (2005). "Breakdown of Cortical Effective Connectivity During Sleep". In: *Science* 309.5744, pp. 2228–2232. DOI: 10.1126/science.1117256.

Mathieson, W. B., W. Heiligenberg, and L. Maler (1987). "Ultrastructural studies of physiologically identified electrosensory afferent synapses in the gymnotiform fish, *Eigenmannia*". In: *Journal of Comparative Neurology* 255.4, pp. 526–537. DOI: 10.1002/cne.902550405.

Maturana, H. R. and F. J. Varela (1980). *Autopoiesis and Cognition: The Realization of the Living*. Boston Studies in the Philosophy of Science 42. Dordrecht: Reidel.

McGill, W. J. (1954). "Multivariate Information Transmission". In: *Psychometrika* 19.2, pp. 97–116. DOI: 10.1007/bf02289159.

Mediano, P. A. M., F. E. Rosas, A. I. Luppi, R. L. Carhart-Harris, D. Bor, A. K. Seth, and A. B. Barrett (2021). *Towards an Extended Taxonomy of Information Dynamics via Integrated Information Decomposition*. arXiv: 2109.13186.

Mellers, B., R. Hertwig, and D. Kahneman (2001). "Do Frequency Representations Eliminate Conjunction Effects? An Exercise in Adversarial Collaboration". In: *Psychological Science* 12.4, pp. 269–275. DOI: 10.1111/1467-9280.00350.

Melloni, L., L. Mudrik, M. Pitts, K. Bendtz, O. Ferrante, U. Gorska, R. Hirschhorn, A. Khalaf, C. Kozma, A. Lepauvre, L. Liu, D. Mazumder, D. Richter, H. Zhou, H. Blumenfeld, M. Boly, D. J. Chalmers, S. Devore, F. Fallon, F. P. de Lange, O. Jensen, G. Kreiman, H. Luo, T. I. Panagiotaropoulos, S. Dehaene, C. Koch, and G. Tononi (2023). "An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory". In: *PLOS ONE* 18.2, e0268577. DOI: 10.1371/journal.pone.0268577.

Merker, B. (2005). "The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution". In: *Consciousness and Cognition* 14.1, pp. 89–114. DOI: 10.1016/s1053-8100(03)00002-3.

Merker, B. (2007). "Consciousness without a cerebral cortex: A challenge for neuroscience and medicine". In: *Behavioral and Brain Sciences* 30.1, pp. 63–81. DOI: 10.1017/s0140525x07000891.

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.

Metzinger, T. (2009). *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. New York: Basic Books.

Metzner, W. (1993a). "An audio-vocal interface in echolocating horseshoe bats". In: *The Journal of Neuroscience* 13.5, pp. 1899–1915. DOI: 10.1523/jneurosci.13-05-01899.1993.

Metzner, W. (1993b). "The jamming avoidance response in *Eigenmannia* is controlled by two separate motor pathways". In: *The Journal of Neuroscience* 13.5, pp. 1862–1878. DOI: 10.1523/jneurosci.13-05-01862.1993.

Metzner, W. and W. Heiligenberg (1991). "The coding of signals in the electric communication of the gymnotiform fish *Eigenmannia*: From electroreceptors to neurons in the torus semicircularis of the midbrain". In: *Journal of Comparative Physiology A* 169.2, pp. 135–150. DOI: 10.1007/bf00215861.

Miłkowski, M. (2013). *Explaining the Computational Mind*. Cambridge, MA: MIT Press.

Millán, A. P., J. J. Torres, and G. Bianconi (2020). "Explosive Higher-Order Kuramoto Dynamics on Simplicial Complexes". In: *Physical Review Letters* 124.21, 218301. DOI: 10.1103/physrevlett.124.218301.

Miller, A., J. Sleigh, J. Barnard, and D. Steyn-Ross (2004). "Does bispectral analysis of the electroencephalogram add anything but complexity? This article is accompanied by Editorial I". In: *British Journal of Anaesthesia* 92.1, pp. 8–13. DOI: 10.1093/bja/aeh003.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press.

Millikan, R. G. (1989). "Biosemantics". In: *The Journal of Philosophy* 86.6, pp. 281–297. DOI: 10.2307/2027123.

Mitchell, M. (2009). *Complexity: A Guided Tour*. New York: Oxford University Press.

Mitchell, M., P. T. Hraber, and J. P. Crutchfield (1993). "Revisiting the Edge of Chaos: Evolving Cellular Automata to Perform Computations". In: *Complex Systems* 7, pp. 89–130.

Mitola, J. and G. Maguire (1999). "Cognitive radio: making software radios more personal". In: *IEEE Personal Communications* 6.4, pp. 13–18. DOI: 10.1109/98.788210.

Montévil, M. and M. Mossio (2015). "Biological organisation as closure of constraints". In: *Journal of Theoretical Biology* 372, pp. 179–191. DOI: 10.1016/j.jtbi.2015.02.029.

Moreno, A. and M. Mossio (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Dordrecht: Springer.

Mormann, F., K. Lehnertz, P. David, and C. E. Elger (2000). "Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients". In: *Physica D: Nonlinear Phenomena* 144.3–4, pp. 358–369. DOI: 10.1016/s0167-2789(00)00087-7.

Mossio, M., C. Saborido, and A. Moreno (2009). "An Organizational Account of Biological Functions". In: *The British Journal for the Philosophy of Science* 60.4, pp. 813–841. DOI: 10.1093/bjps/axp036.

Mukamel, E. A., E. Pirondini, B. Babadi, K. F. K. Wong, E. T. Pierce, P. G. Harrell, J. L. Walsh, A. F. Salazar-Gomez, S. S. Cash, E. N. Eskandar, V. S. Weiner, E. N. Brown, and P. L. Purdon (2014). "A Transition in Brain State during Propofol-Induced Unconsciousness". In: *The Journal of Neuroscience* 34.3, pp. 839–845. DOI: 10.1523/jneurosci.5813-12.2014.

Myles, P., K. Leslie, J. McNeil, A. Forbes, and M. Chan (2004). "Bispectral index monitoring to prevent awareness during anaesthesia: the B-Aware randomised controlled trial". In: *The Lancet* 363.9423, pp. 1757–1763. DOI: 10.1016/s0140-6736(04)16300-9.

Nagel, T. (1974). "What Is It Like to Be a Bat?" In: *The Philosophical Review* 83.4, pp. 435–450. DOI: 10.2307/2183914.

Nagel, T. (1986). *The View from Nowhere*. New York: Oxford University Press.

Nakao, H. (2016). "Phase reduction approach to synchronisation of nonlinear oscillators". In: *Contemporary Physics* 57.2, pp. 188–214. DOI: 10.1080/00107514.2015.1094987.

Nayfeh, A. H. and D. T. Mook (1979). *Nonlinear Oscillations*. New York: Wiley.

Neander, K. (2017). *A Mark of the Mental: In Defense of Informational Teleosemantics*. Cambridge, MA: MIT Press.

Nelson, M. E. and M. A. MacIver (2006). "Sensory acquisition in active sensing systems". In: *Journal of Comparative Physiology A* 192.6, pp. 573–586. DOI: 10.1007/s00359-006-0099-4.

Nelson, T. O. and L. Narens (1990). "Metamemory: A Theoretical Framework and New Findings". In: *Psychology of Learning and Motivation*. Ed. by G. H. Bower. Vol. 26. San Diego: Academic Press, pp. 125–173. DOI: 10.1016/s0079-7421(08)60053-5.

Netoff, T. I., M. I. Banks, A. D. Dorval, C. D. Acker, J. S. Haas, N. Kopell, and J. A. White (2005). "Synchronization in Hybrid Neuronal Networks of the Hippocampal Formation". In: *Journal of Neurophysiology* 93.3, pp. 1197–1208. DOI: 10.1152/jn.00982.2004.

Neumann, J. von (1955). *Mathematical Foundations of Quantum Mechanics*. Trans. by R. T. Beyer. Princeton: Princeton University Press.

Neumann, J. von (1966). *Theory of Self-Reproducing Automata*. Ed. by A. W. Burks. Urbana: University of Illinois Press.

Newman, M. (2010). *Networks: An Introduction*. Oxford: Oxford University Press.

Ngo, H.-V. V., T. Martinetz, J. Born, and M. Mölle (2013). "Auditory Closed-Loop Stimulation of the Sleep Slow Oscillation Enhances Memory". In: *Neuron* 78.3, pp. 545–553. DOI: 10.1016/j.neuron.2013.03.006.

Nikias, C. and J. Mendel (1993). "Signal processing with higher-order spectra". In: *IEEE Signal Processing Magazine* 10.3, pp. 10–37. DOI: 10.1109/79.221324.

Nisbett, R. E. and T. D. Wilson (1977). "Telling more than we can know: Verbal reports on mental processes". In: *Psychological Review* 84.3, pp. 231–259. DOI: 10.1037/0033-295x.84.3.231.

Noë, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.

Nolte, G., O. Bai, L. Wheaton, Z. Mari, S. Vorbach, and M. Hallett (2004). "Identifying true brain interaction from EEG data using the imaginary part of coherency". In: *Clinical Neurophysiology* 115.10, pp. 2292–2307. DOI: 10.1016/j.clinph.2004.04.029.

Norcia, A. M., L. G. Appelbaum, J. M. Ales, B. R. Cottereau, and B. Rossion (2015). "The steady-state visual evoked potential in vision research: A review". In: *Journal of Vision* 15.6, p. 4. DOI: 10.1167/15.6.4.

Nosek, B. A., C. R. Ebersole, A. C. DeHaven, and D. T. Mellor (2018). "The preregistration revolution". In: *Proceedings of the National Academy of Sciences* 115.11, pp. 2600–2606. DOI: 10.1073/pnas.1708274114.

Notbohm, A., J. Kurths, and C. S. Herrmann (2016). "Modification of Brain Oscillations via Rhythmic Light Stimulation Provides Evidence for Entrainment but Not for Superposition of Event-Related Responses". In: *Frontiers in Human Neuroscience* 10. DOI: 10.3389/fnhum.2016.00010.

Nunez, P. L. and R. Srinivasan (2006). *Electric Fields of the Brain: The Neurophysics of EEG*. 2nd ed. New York: Oxford University Press.

O'Keefe, J. and N. Burgess (2005). "Dual phase and rate coding in hippocampal place cells: Theoretical significance and relationship to entorhinal grid cells". In: *Hippocampus* 15.7, pp. 853–866. DOI: 10.1002/hipo.20115.

O'Keefe, J. and L. Nadel (1978). *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press.

O'Keefe, J. and M. L. Recce (1993). "Phase relationship between hippocampal place units and the EEG theta rhythm". In: *Hippocampus* 3.3, pp. 317–330. DOI: 10.1002/hipo.450030307.

O'Regan, J. K. and A. Noë (2001). "A sensorimotor account of vision and visual consciousness". In: *Behavioral and Brain Sciences* 24.5, pp. 939–973. DOI: 10.1017/s0140525x01000115.

Odegaard, B., R. T. Knight, and H. Lau (2017). "Should a Few Null Findings Falsify Prefrontal Theories of Conscious Perception?" In: *The Journal of Neuroscience* 37.40, pp. 9593–9602. DOI: 10.1523/jneurosci.3217-16.2017.

Oizumi, M., L. Albantakis, and G. Tononi (2014). "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0".

In: *PLoS Computational Biology* 10.5, e1003588. DOI: 10.1371/journal.pcbi.1003588.

Oppenheim, A. V. and R. W. Schafer (2010). *Discrete-Time Signal Processing*. 3rd ed. Upper Saddle River, NJ: Pearson.

Oster, G. (1973). "Auditory Beats in the Brain". In: *Scientific American* 229.4, pp. 94–102. DOI: 10.1038/scientificamerican1073-94.

Oster, G. and Y. Nishijima (1963). "Moiré Patterns". In: *Scientific American* 208.5, pp. 54–63. DOI: 10.1038/scientificamerican0563-54.

Oster, G., M. Wasserman, and C. Zwerling (1964). "Theoretical Interpretation of Moiré Patterns". In: *Journal of the Optical Society of America* 54.2, pp. 169–175. DOI: 10.1364/josa.54.000169.

Ott, E. and T. M. Antonsen (2008). "Low dimensional behavior of large systems of globally coupled oscillators". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 18.3, 037113. DOI: 10.1063/1.2930766.

Owen, A. M., M. R. Coleman, M. Boly, M. H. Davis, S. Laureys, and J. D. Pickard (2006). "Detecting Awareness in the Vegetative State". In: *Science* 313.5792, p. 1402. DOI: 10.1126/science.1130197.

Packard, N. H. (1988). "Adaptation Toward the Edge of Chaos". In: *Dynamic Patterns in Complex Systems*. Ed. by J. A. S. Kelso, A. J. Mandell, and M. F. Shlesinger. Singapore: World Scientific, pp. 293–301.

Paik, S.-B. and D. L. Ringach (2011). "Retinal origin of orientation maps in visual cortex". In: *Nature Neuroscience* 14.7, pp. 919–925. DOI: 10.1038/nn.2824.

Panaggio, M. J. and D. M. Abrams (2015). "Chimera states: coexistence of coherence and incoherence in networks of coupled oscillators". In: *Nonlinearity* 28.3, R67–R87. DOI: 10.1088/0951-7715/28/3/r67.

Parr, T., G. Pezzulo, and K. J. Friston (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. Cambridge, MA: MIT Press.

Parsa, A., S. Witthaus, N. Pashine, C. O'Hern, R. Kramer-Bottiglio, and J. Bongard (2023). "Universal Mechanical Polycomputation in Granular Matter". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, pp. 193–201. DOI: 10.1145/3583131.3590520.

Pastalkova, E., V. Itskov, A. Amarasingham, and G. Buzsáki (2008). "Internally Generated Cell Assembly Sequences in the Rat Hippocampus". In: *Science* 321.5894, pp. 1322–1327. DOI: 10.1126/science.1159775.

Pattee, H. H. (2015). "Cell phenomenology: The first phenomenon". In: *Progress in Biophysics and Molecular Biology* 119.3, pp. 461–468. DOI: 10.1016/j.pbiomolbio.2015.06.010.

Pattee, H. H. (1969). "How Does a Molecule Become a Message?" In: *Developmental Biology Supplement* 3, pp. 1–16.

Pattee, H. H. (1982). "Cell Psychology: An Evolutionary Approach to the Symbol–Matter Problem". In: *Cognition and Brain Theory* 5.4, pp. 325–341.

Pattee, H. H. (1995). "Evolving Self-Reference: Matter, Symbols, and Semantic Closure". In: *Communication and Cognition – Artificial Intelligence* 12.1–2, pp. 9–27.

Pattee, H. H. (2001). "The physics of symbols: bridging the epistemic cut". In: *Biosystems* 60.1–3, pp. 5–21. DOI: 10.1016/s0303-2647(01)00104-6.

Pattee, H. H. and J. Rączaszek-Leonardi (2012). *Laws, Language and Life: Howard Pattee's Classic Papers on the Physics of Symbols with Contemporary Commentary*. Biosemiotics 7. Dordrecht: Springer.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press.

Pecora, L. M., F. Sorrentino, A. M. Hagerstrom, T. E. Murphy, and R. Roy (2014). "Cluster synchronization and isolated desynchronization in complex networks with symmetries". In: *Nature Communications* 5.1, 4079. DOI: 10.1038/ncomms5079.

Percival, D. B. and A. T. Walden (1993). *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge: Cambridge University Press.

Perry, J. (1979). "The Problem of the Essential Indexical". In: *Noûs* 13.1, pp. 3–21. DOI: 10.2307/2214792.

Petkova, V. I. and H. H. Ehrsson (2008). "If I Were You: Perceptual Illusion of Body Swapping". In: *PLoS ONE* 3.12, e3832. DOI: 10.1371/journal.pone.0003832.

Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. Oxford: Oxford University Press.

Pikovsky, A., M. Rosenblum, and J. Kurths (2001). *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge: Cambridge University Press.

Plato (1989). *Symposium*. Trans. by A. Nehamas and P. Woodruff. Indianapolis: Hackett.

Pletzer, B., H. Kerschbaum, and W. Klimesch (2010). "When frequencies never synchronize: The golden mean and the resting EEG". In: *Brain Research* 1335, pp. 91–102. DOI: 10.1016/j.brainres.2010.03.074.

Plomp, R. and W. J. M. Levelt (1965). "Tonal Consonance and Critical Bandwidth". In: *The Journal of the Acoustical Society of America* 38.4, pp. 548–560. DOI: 10.1121/1.1909741.

Polanía, R., M. A. Nitsche, C. Korman, G. Batsikadze, and W. Paulus (2012). "The Importance of Timing in Segregated Theta Phase-Coupling for Cognitive Performance". In: *Current Biology* 22.14, pp. 1314–1318. DOI: 10.1016/j.cub.2012.05.021.

Pöppel, E. (1997). "A hierarchical model of temporal perception". In: *Trends in Cognitive Sciences* 1.2, pp. 56–61. DOI: 10.1016/s1364-6613(97)01008-5.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Posner, M. I. (1980). "Orienting of Attention". In: *Quarterly Journal of Experimental Psychology* 32.1, pp. 3–25. DOI: 10.1080/00335558008248231.

Post, D., B. Han, and P. Ifju (1994). *High Sensitivity Moiré: Experimental Analysis for Mechanics and Materials*. New York: Springer.

Potter, S. M., A. El Hady, and E. E. Fetz (2014). "Closed-loop neuroscience and neuroengineering". In: *Frontiers in Neural Circuits* 8. DOI: 10.3389/fncir.2014.00115.

Pouget, A. and T. J. Sejnowski (1997). "Spatial Transformations in the Parietal Cortex Using Basis Functions". In: *Journal of Cognitive Neuroscience* 9.2, pp. 222–237. DOI: 10.1162/jocn.1997.9.2.222.

Prichard, D. and J. Theiler (1994). "Generating surrogate data for time series with several simultaneously measured variables". In: *Physical Review Letters* 73.7, pp. 951–954. DOI: 10.1103/PhysRevLett.73.951.

Purdon, P. L., E. T. Pierce, E. A. Mukamel, M. J. Prerau, J. L. Walsh, K. F. K. Wong, A. F. Salazar-Gomez, P. G. Harrell, A. L. Sampson, A. Cimenser, S. Ching, N. J. Kopell, C. Tavares-Stoeckel, K. Habeeb, R. Merhar, and E. N. Brown (2013). "Electroencephalogram signatures of loss and recovery of consciousness from propofol". In: *Proceedings of the National Academy of Sciences* 110.12, E1142–E1151. DOI: 10.1073/pnas.1221180110.

Putnam, H. (1967). "Psychological Predicates". In: *Art, Mind, and Religion*. University of Pittsburgh Press, pp. 37–48. DOI: 10.2307/jj.6380610.6.

Putnam, H. (1988). *Representation and Reality*. Cambridge, MA: MIT Press.

Rabinovich, M., R. Huerta, and G. Laurent (2008). "Transient Dynamics for Neural Processing". In: *Science* 321.5885, pp. 48–50. DOI: 10.1126/science.1155564.

Rampil, I. J. (1998). "A Primer for EEG Signal Processing in Anesthesia". In: *Anesthesiology* 89.4, pp. 980–1002. DOI: 10.1097/00000542-199810000-00023.

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Rao, R. P. N. and D. H. Ballard (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects". In: *Nature Neuroscience* 2.1, pp. 79–87. DOI: 10.1038/4580.

Rattenborg, N., C. Amlaner, and S. Lima (2000). "Behavioral, neurophysiological and evolutionary perspectives on unihemispheric sleep". In: *Neuroscience & Biobehavioral Reviews* 24.8, pp. 817–842. DOI: 10.1016/s0149-7634(00)00039-7.

Rayleigh, L. (1874). "On the manufacture and theory of diffraction-gratings". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 47.310, pp. 81–93. DOI: 10.1080/14786447408640996.

Redish, A. D. (2016). "Vicarious trial and error". In: *Nature Reviews Neuroscience* 17.3, pp. 147–159. DOI: 10.1038/nrn.2015.30.

Regan, M. and D. Regan (1988). "A frequency domain technique for characterizing nonlinearities in biological systems". In: *Journal of Theoretical Biology* 133.3, pp. 293–317. DOI: 10.1016/s0022-5193(88)80323-0.

Requarth, T. and N. B. Sawtell (2011). "Neural mechanisms for filtering self-generated sensory signals in cerebellum-like circuits". In: *Current Opinion in Neurobiology* 21.4, pp. 602–608. DOI: 10.1016/j.conb.2011.05.031.

Reynolds, O. (1883). "An experimental investigation of the circumstances which determine whether the motion of water shall be direct or sinuous, and of the law of resistance in parallel channels". In: *Philosophical Transactions of the Royal Society of London* 174.174, pp. 935–982. DOI: 10.1098/rstl.1883.0029.

Ribeiro-Palau, R., C. Zhang, K. Watanabe, T. Taniguchi, J. Hone, and C. R. Dean (2018). "Twistable electronics with dynamically rotatable heterostructures". In: *Science* 361.6403, pp. 690–693. DOI: 10.1126/science.aat6981.

Ringach, D. L. (2007). "On the Origin of the Functional Architecture of the Cortex". In: *PLoS ONE* 2.2, e251. DOI: 10.1371/journal.pone.0000251.

Robles, L. and M. A. Ruggero (2001). "Mechanics of the Mammalian Cochlea". In: *Physiological Reviews* 81.3, pp. 1305–1352. DOI: 10.1152/physrev.2001.81.3.1305.

Rosas, F. E., P. A. M. Mediano, H. J. Jensen, A. K. Seth, A. B. Barrett, R. L. Carhart-Harris, and D. Bor (2020). "Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data". In: *PLOS Computational Biology* 16.12, e1008289. DOI: 10.1371/journal.pcbi.1008289.

Rose, G. and W. Heiligenberg (1985a). "Structure and function of electrosensory neurons in the torus semicircularis of *Eigenmannia*: morphological correlates of phase and amplitude sensitivity". In: *The Journal of Neuroscience* 5.8, pp. 2269–2280. DOI: 10.1523/jneurosci.05-08-02269.1985.

Rose, G. and W. Heiligenberg (1985b). "Temporal hyperacuity in the electric sense of fish". In: *Nature* 318.6042, pp. 178–180. DOI: 10.1038/318178a0.

Rose, G. and W. Heiligenberg (1986a). "Limits of phase and amplitude sensitivity in the torus semicircularis of *Eigenmannia*". In: *Journal of Comparative Physiology A* 159.6, pp. 813–822. DOI: 10.1007/bf00603734.

Rose, G. and W. Heiligenberg (1986b). "Neural coding of difference frequencies in the midbrain of the electric fish *Eigenmannia*: Reading the sense of rotation in an amplitude-phase plane". In: *Journal of Comparative Physiology A* 158.5, pp. 613–624. DOI: 10.1007/bf00603818.

Rose, G., C. Keller, and W. Heiligenberg (1987). "'Ancestral' neural mechanisms of electrolocation suggest a substrate for the evolution of the jamming avoidance response". In: *Journal of Comparative Physiology A* 160.4, pp. 491–500. DOI: 10.1007/bf00615082.

Rose, G. J. (2004). "Insights into neural mechanisms and evolution of behaviour from electric fish". In: *Nature Reviews Neuroscience* 5.12, pp. 943–951. DOI: 10.1038/nrn1558.

Rosen, R. (1991). *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life*. New York: Columbia University Press.

Rosenthal, D. M. (2005). *Consciousness and Mind*. Oxford: Clarendon Press.

Rovelli, C. (1996). "Relational quantum mechanics". In: *International Journal of Theoretical Physics* 35.8, pp. 1637–1678. DOI: 10.1007/bf02302261.

Rudrauf, D., D. Bennequin, I. Granic, G. Landini, K. Friston, and K. Williford (2017). "A mathematical model of embodied consciousness". In: *Journal of Theoretical Biology* 428, pp. 106–131. DOI: 10.1016/j.jtbi.2017.05.032.

Sakaguchi, H. and Y. Kuramoto (1986). "A Soluble Active Rotator Model Showing Phase Transitions via Mutual Entrainment". In: *Progress of Theoretical Physics* 76.3, pp. 576–581. DOI: 10.1143/ptp.76.576.

Sarasso, S., M. Boly, M. Napolitani, O. Gosseries, V. Charland-Verville, S. Casarotto, M. Rosanova, A. G. Casali, J.-F. Brichant, P. Boveroux, S. Rex, G. Tononi, S. Laureys, and M. Massimini (2015). "Consciousness and Complexity during Unresponsiveness Induced by Propofol, Xenon, and Ketamine". In: *Current Biology* 25.23, pp. 3099–3105. DOI: 10.1016/j.cub.2015.10.014.

Schacter, D. L., D. R. Addis, and R. L. Buckner (2007). "Remembering the past to imagine the future: the prospective brain". In: *Nature Reviews Neuroscience* 8.9, pp. 657–661. DOI: 10.1038/nrn2213.

Scheich, H., T. H. Bullock, and R. H. Hamstra (1973). "Coding properties of two classes of afferent nerve fibers: high-frequency electroreceptors in the electric fish, *Eigenmannia*". In: *Journal of Neurophysiology* 36.1, pp. 39–60. DOI: 10.1152/jn.1973.36.1.39.

Schetzen, M. (1980). *The Volterra and Wiener Theories of Nonlinear Systems*. New York: Wiley.

Schmidt, R. C., C. Carello, and M. T. Turvey (1990). "Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people". In: *Journal of Experimental Psychology: Human Perception and Performance* 16.2, pp. 227–247. DOI: 10.1037/0096-1523.16.2.227.

Schmidt-Hieber, C. and M. Häusser (2013). "Cellular mechanisms of spatial navigation in the medial entorhinal cortex". In: *Nature Neuroscience* 16.3, pp. 325–331. DOI: 10.1038/nn.3340.

Schnitzler, H.-U. (1968). "Die Ultraschall-Ortungslaute der Hufeisen-Fledermäuse (Chiroptera-Rhinolophidae) in verschiedenen Orientierungssituationen". In: *Zeitschrift für Vergleichende Physiologie* 57.4, pp. 376–408. DOI: 10.1007/bf00303062.

Schnitzler, H.-U. and A. Denzinger (2011). "Auditory fovea and Doppler shift compensation: adaptations for flutter detection in echolocating bats using CF-FM signals". In: *Journal of Comparative Physiology A* 197.5, pp. 541–559. DOI: 10.1007/s00359-010-0569-6.

Scholz, J. P., J. A. S. Kelso, and G. Schöner (1987). "Nonequilibrium phase transitions in coordinated biological motion: Critical slowing down and switching time". In: *Physics Letters A* 123.8, pp. 390–394. DOI: 10.1016/0375-9601(87)90038-7.

Schöner, G. and J. A. S. Kelso (1988). "Dynamic Pattern Generation in Behavioral and Neural Systems". In: *Science* 239.4847, pp. 1513–1520. DOI: 10.1126/science.3281253.

Schottdorf, M., W. Keil, D. Coppola, L. E. White, and F. Wolf (2015). "Random Wiring, Ganglion Cell Mosaics, and the Functional Architecture of the Visual Cortex". In: *PLOS Computational Biology* 11.11, e1004602. DOI: 10.1371/journal.pcbi.1004602.

Schreiber, T. (2000). "Measuring Information Transfer". In: *Physical Review Letters* 85.2, pp. 461–464. DOI: 10.1103/physrevlett.85.461.

Schreiber, T. and A. Schmitz (1996). "Improved Surrogate Data for Nonlinearity Tests". In: *Physical Review Letters* 77.4, pp. 635–638. DOI: 10.1103/PhysRevLett.77.635.

Schreiber, T. and A. Schmitz (2000). "Surrogate time series". In: *Physica D: Nonlinear Phenomena* 142.3–4, pp. 346–382. DOI: 10.1016/s0167-2789(00)00043-9.

Schwitzgebel, E. (2008). "The Unreliability of Naive Introspection". In: *The Philosophical Review* 117.2, pp. 245–273. DOI: 10.1215/00318108-2007-037.

Scofield, J. H. (1994). "Frequency-domain description of a lock-in amplifier". In: *American Journal of Physics* 62.2, pp. 129–133. DOI: 10.1119/1.17629.

Searle, J. R. (1980). "Minds, brains, and programs". In: *Behavioral and Brain Sciences* 3.3, pp. 417–424. DOI: 10.1017/s0140525x00005756.

Searle, J. R. (1990). "Is the Brain a Digital Computer?" In: *Proceedings and Addresses of the American Philosophical Association* 64.3, pp. 21–37. DOI: 10.2307/3130074.

Searle, J. R. (1992). *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Seth, A. K. (2013). "Interoceptive inference, emotion, and the embodied self". In: *Trends in Cognitive Sciences* 17.11, pp. 565–573. DOI: 10.1016/j.tics.2013.09.007.

Seth, A. K. and T. Bayne (2022). "Theories of consciousness". In: *Nature Reviews Neuroscience* 23.7, pp. 439–452. DOI: 10.1038/s41583-022-00587-4.

Shagrir, O. (2020). "In defense of the semantic view of computation". In: *Synthese* 197.9, pp. 4083–4108. DOI: 10.1007/s11229-018-01921-z.

Shalizi, C. R. and J. P. Crutchfield (2001). "Computational Mechanics: Pattern and Prediction, Structure and Simplicity". In: *Journal of Statistical Physics* 104.3–4, pp. 817–879. DOI: 10.1023/a:1010388907793.

Shannon, C. E. (1948). "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Sharp, A. A., M. B. O'Neil, L. F. Abbott, and E. Marder (1993). "Dynamic clamp: computer-generated conductances in real neurons". In: *Journal of Neurophysiology* 69.3, pp. 992–995. DOI: 10.1152/jn.1993.69.3.992.

Shea, N. (2018). *Representation in Cognitive Science*. Oxford: Oxford University Press.

Sherrington, C. S. (1906). *The Integrative Action of the Nervous System*. New Haven: Yale University Press.

Sigl, J. C. and N. G. Chamoun (1994). "An introduction to bispectral analysis for the electroencephalogram". In: *Journal of Clinical Monitoring* 10.6, pp. 392–404. DOI: 10.1007/bf01618421.

Skardal, P. S. and A. Arenas (2019). "Abrupt Desynchronization and Extensive Multistability in Globally Coupled Oscillator Simplexes". In: *Physical Review Letters* 122.24, 248301. DOI: 10.1103/physrevlett.122.248301.

Skardal, P. S. and A. Arenas (2020). "Higher order interactions in complex networks of phase oscillators promote abrupt synchronization switching". In: *Communications Physics* 3.1, 218. DOI: 10.1038/s42005-020-00485-0.

Smith, J. D., W. E. Shields, and D. A. Washburn (2003). "The comparative psychology of uncertainty monitoring and metacognition". In: *Behavioral and Brain Sciences* 26.3, pp. 317–339. DOI: 10.1017/s0140525x03000086.

Sneddon, L. U., V. A. Braithwaite, and M. J. Gentle (2003). "Do fishes have nociceptors? Evidence for the evolution of a vertebrate sensory system". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270.1520, pp. 1115–1121. DOI: 10.1098/rspb.2003.2349.

Soodak, R. E. (1987). "The retinal ganglion cell mosaic defines orientation columns in striate cortex". In: *Proceedings of the National Academy of Sciences* 84.11, pp. 3936–3940. DOI: 10.1073/pnas.84.11.3936.

Specification of the Bluetooth System, Version 1.2 (2003). Bluetooth SIG.

Sperry, R. W. (1950). "Neural basis of the spontaneous optokinetic response produced by visual inversion". In: *Journal of Comparative and Physiological Psychology* 43.6, pp. 482–489. DOI: 10.1037/h0055479.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search*. 2nd ed. Cambridge, MA: MIT Press.

Sprevak, M. (2018). "Triviality arguments about computational implementation". In: *The Routledge Handbook of the Computational Mind*. Routledge, pp. 175–191. DOI: 10.4324/9781315643670-14.

Stam, C. J., G. Nolte, and A. Daffertshofer (2007). "Phase lag index: Assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources". In: *Human Brain Mapping* 28.11, pp. 1178–1193. DOI: 10.1002/hbm.20346.

Stamper, S. A., E. Carrera-G, E. W. Tan, V. Fugère, R. Krahe, and E. S. Fortune (2010). "Species differences in group size and electrosensory interference in weakly electric fishes: Implications for electrosensory processing". In: *Behavioural Brain Research* 207.2, pp. 368–376. DOI: 10.1016/j.bbr.2009.10.023.

Stamper, S. A., E. Roth, N. J. Cowan, and E. S. Fortune (2012). "Active sensing via movement shapes spatiotemporal patterns of sensory feedback".

In: *Journal of Experimental Biology* 215.9, pp. 1567–1574. DOI: 10.1242/jeb.068007.

Stensola, H., T. Stensola, T. Solstad, K. Frøland, M.-B. Moser, and E. I. Moser (2012). "The entorhinal grid map is discretized". In: *Nature* 492.7427, pp. 72–78. DOI: 10.1038/nature11649.

Sterling, P. (2012). "Allostasis: A model of predictive regulation". In: *Physiology & Behavior* 106.1, pp. 5–15. DOI: 10.1016/j.physbeh.2011.06.004.

Strogatz, S. H. (2000). "From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators". In: *Physica D: Nonlinear Phenomena* 143.1–4, pp. 1–20. DOI: 10.1016/s0167-2789(00)00094-4.

Strogatz, S. H. (2015). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. 2nd ed. Boulder, CO: Westview Press.

Suárez Morell, E., J. D. Correa, P. Vargas, M. Pacheco, and Z. Barticevic (2010). "Flat bands in slightly twisted bilayer graphene: Tight-binding calculations". In: *Physical Review B* 82.12, 121407. DOI: 10.1103/physrevb.82.121407.

Suddendorf, T. and M. C. Corballis (2007). "The evolution of foresight: What is mental time travel, and is it unique to humans?" In: *Behavioral and Brain Sciences* 30.3, pp. 299–313. DOI: 10.1017/s0140525x07001975.

Sussmann, H. J. and R. S. Zahler (1978). "Catastrophe theory as applied to the social and biological sciences: A critique". In: *Synthese* 37.2, pp. 117–216. DOI: 10.1007/bf00869575.

Swoyer, C. (1991). "Structural representation and surrogative reasoning". In: *Synthese* 87.3, pp. 449–508. DOI: 10.1007/bf00499820.

Tagliazucchi, E., D. R. Chialvo, M. Siniatchkin, E. Amico, J.-F. Brichant, V. Bonhomme, Q. Noirhomme, H. Laufs, and S. Laureys (2016). "Large-scale signatures of unconsciousness are consistent with a departure from critical dynamics". In: *Journal of The Royal Society Interface* 13.114, p. 20151027. DOI: 10.1098/rsif.2015.1027.

Tanaka, T. and T. Aoyagi (2011). "Multistable Attractors in a Network of Phase Oscillators with Three-Body Interactions". In: *Physical Review Letters* 106.22, 224101. DOI: 10.1103/physrevlett.106.224101.

Tarnopolsky, G., A. J. Kruchkov, and A. Vishwanath (2019). "Origin of Magic Angles in Twisted Bilayer Graphene". In: *Physical Review Letters* 122.10, 106405. DOI: 10.1103/physrevlett.122.106405.

The New York Declaration on Animal Consciousness (2024). New York University. URL: https://sites.google.com/nyu.edu/nydeclaration/declaration (visited on 09/18/2026).

Theiler, J., S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer (1992). "Testing for nonlinearity in time series: the method of surrogate data". In: *Physica D: Nonlinear Phenomena* 58.1–4, pp. 77–94. DOI: 10.1016/0167-2789(92)90102-s.

Thom, R. (1975). *Structural Stability and Morphogenesis: An Outline of a General Theory of Models*. Trans. by D. H. Fowler. Reading, MA: W. A. Benjamin.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.

Thomson, D. (1982). "Spectrum estimation and harmonic analysis". In: *Proceedings of the IEEE* 70.9, pp. 1055–1096. DOI: 10.1109/proc.1982.12433.

Thut, G., D. Veniero, V. Romei, C. Miniussi, P. Schyns, and J. Gross (2011). "Rhythmic TMS Causes Local Entrainment of Natural Oscillatory Signatures". In: *Current Biology* 21.14, pp. 1176–1185. DOI: 10.1016/j.cub.2011.05.049.

Timme, N., W. Alford, B. Flecker, and J. M. Beggs (2014). "Synergy, redundancy, and multivariate information measures: an experimentalist's perspective". In: *Journal of Computational Neuroscience* 36.2, pp. 119–140. DOI: 10.1007/s10827-013-0458-4.

Tinsley, M. R., S. Nkomo, and K. Showalter (2012). "Chimera and phase-cluster states in populations of coupled chemical oscillators". In: *Nature Physics* 8.9, pp. 662–665. DOI: 10.1038/nphys2371.

Tishby, N., F. C. Pereira, and W. Bialek (2000). *The Information Bottleneck Method*. arXiv: physics/0004057.

Tognoli, E. and J. A. S. Kelso (2014). "The Metastable Brain". In: *Neuron* 81.1, pp. 35–48. DOI: 10.1016/j.neuron.2013.12.022.

Tolman, E. C. (1948). "Cognitive maps in rats and men". In: *Psychological Review* 55.4, pp. 189–208. DOI: 10.1037/h0061626.

Tomasello, M. (2014). *A Natural History of Human Thinking*. Cambridge, MA: Harvard University Press.

Tononi, G., O. Sporns, and G. M. Edelman (1994). "A measure for brain complexity: relating functional segregation and integration in the nervous system". In: *Proceedings of the National Academy of Sciences* 91.11, pp. 5033–5037. DOI: 10.1073/pnas.91.11.5033.

Tononi, G. (1998). "Consciousness and Complexity". In: *Science* 282.5395, pp. 1846–1851. DOI: 10.1126/science.282.5395.1846.

Tononi, G. (2004). "An information integration theory of consciousness". In: *BMC Neuroscience* 5.1, 42. DOI: 10.1186/1471-2202-5-42.

Tononi, G., M. Boly, M. Massimini, and C. Koch (2016). "Integrated information theory: from consciousness to its physical substrate". In: *Nature Reviews Neuroscience* 17.7, pp. 450–461. DOI: 10.1038/nrn.2016.44.

Tononi, G. and O. Sporns (2003). "Measuring information integration". In: *BMC Neuroscience* 4.1, 31. DOI: 10.1186/1471-2202-4-31.

Tort, A. B. L., R. Komorowski, H. Eichenbaum, and N. Kopell (2010). "Measuring Phase-Amplitude Coupling Between Neuronal Oscillations of Different Frequencies". In: *Journal of Neurophysiology* 104.2, pp. 1195–1210. DOI: 10.1152/jn.00106.2010.

Touboul, J. and A. Destexhe (2010). "Can Power-Law Scaling and Neuronal Avalanches Arise from Stochastic Dynamics?" In: *PLoS ONE* 5.2, e8982. DOI: 10.1371/journal.pone.0008982.

Tsakiris, M. (2010). "My body in the brain: A neurocognitive model of body-ownership". In: *Neuropsychologia* 48.3, pp. 703–712. DOI: 10.1016/j.neuropsychologia.2009.09.034.

Turing, A. M. (1937). "On Computable Numbers, with an Application to the Entscheidungsproblem". In: *Proceedings of the London Mathematical Society* s2-42.1, pp. 230–265. DOI: 10.1112/plms/s2-42.1.230.

Uexküll, J. von (2010). *A Foray into the Worlds of Animals and Humans, with A Theory of Meaning*. Trans. by J. D. O'Neil. Minneapolis: University of Minnesota Press.

VanRullen, R. and C. Koch (2003). "Is perception discrete or continuous?" In: *Trends in Cognitive Sciences* 7.5, pp. 207–213. DOI: 10.1016/s1364-6613(03)00095-0.

Varela, F., H. Maturana, and R. Uribe (1974). "Autopoiesis: The organization of living systems, its characterization and a model". In: *Biosystems* 5.4, pp. 187–196. DOI: 10.1016/0303-2647(74)90031-8.

Varela, F., J.-P. Lachaux, E. Rodriguez, and J. Martinerie (2001). "The brainweb: Phase synchronization and large-scale integration". In: *Nature Reviews Neuroscience* 2.4, pp. 229–239. DOI: 10.1038/35067550.

Varela, F. J., E. Thompson, and E. Rosch (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Viete, S. and W. Heiligenberg (1991). "The development of the Jamming Avoidance Response (JAR) in *Eigenmannia*: An innate behavior indeed". In: *Journal of Comparative Physiology A* 169.1, pp. 15–23. DOI: 10.1007/bf00198169.

Vinck, M., R. Oostenveld, M. van Wingerden, F. Battaglia, and C. M. A. Pennartz (2011). "An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias". In: *NeuroImage* 55.4, pp. 1548–1565. DOI: 10.1016/j.neuroimage.2011.01.055.

Vinck, M., M. van Wingerden, T. Womelsdorf, P. Fries, and C. M. Pennartz (2010). "The pairwise phase consistency: A bias-free measure of rhythmic neuronal synchronization". In: *NeuroImage* 51.1, pp. 112–122. DOI: 10.1016/j.neuroimage.2010.01.073.

Voegelin, E. (1974). *Order and History, Volume IV: The Ecumenic Age*. Baton Rouge: Louisiana State University Press.

Wannier, G. H. (1950). "Antiferromagnetism. The Triangular Ising Net". In: *Physical Review* 79.2, pp. 357–364. DOI: 10.1103/physrev.79.357.

Wark, B., B. N. Lundstrom, and A. Fairhall (2007). "Sensory adaptation". In: *Current Opinion in Neurobiology* 17.4, pp. 423–429. DOI: 10.1016/j.conb.2007.07.001.

Watanabe, A. and K. Takeda (1963). "The Change of Discharge Frequency by A.C. Stimulus in A Weak Electric Fish". In: *Journal of Experimental Biology* 40.1, pp. 57–66. DOI: 10.1242/jeb.40.1.57.

Webb, T. W., H. H. Kean, and M. S. A. Graziano (2016). "Effects of Awareness on the Control of Attention". In: *Journal of Cognitive Neuroscience* 28.6, pp. 842–851. DOI: 10.1162/jocn_a_00931.

Weber, A. and F. J. Varela (2002). "Life after Kant: Natural purposes and the autopoietic foundations of biological individuality". In: *Phenomenology and the Cognitive Sciences* 1.2, pp. 97–125. DOI: 10.1023/a:1020368120174.

Weil, S. (2002). *Gravity and Grace*. Trans. by E. Crawford and M. von der Ruhr. London: Routledge.

Weiskrantz, L. (1986). *Blindsight: A Case Study and Implications*. Oxford: Clarendon Press.

Welch, P. (1967). "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms". In: *IEEE Transactions on Audio and Electroacoustics* 15.2, pp. 70–73. DOI: 10.1109/tau.1967.1161901.

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press.

Wiley, D. A., S. H. Strogatz, and M. Girvan (2006). "The size of the sync basin". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 16.1, 015103. DOI: 10.1063/1.2165594.

Williams, P. L. and R. D. Beer (2010). *Nonnegative Decomposition of Multivariate Information*. arXiv: 1004.2515.

Williford, K., D. Bennequin, K. Friston, and D. Rudrauf (2018). "The Projective Consciousness Model and Phenomenal Selfhood". In: *Frontiers in Psychology* 9, 2571. DOI: 10.3389/fpsyg.2018.02571.

Wilterson, A. I. and M. S. A. Graziano (2021). "The attention schema theory in a neural network agent: Controlling visuospatial attention using a descriptive model of attention". In: *Proceedings of the National Academy of Sciences* 118.33, e2102421118. DOI: 10.1073/pnas.2102421118.

Winfree, A. T. (1967). "Biological rhythms and the behavior of populations of coupled oscillators". In: *Journal of Theoretical Biology* 16.1, pp. 15–42. DOI: 10.1016/0022-5193(67)90051-3.

Winfree, A. T. (2001). *The Geometry of Biological Time*. 2nd ed. New York: Springer.

Wittmann, M. (2011). "Moments in Time". In: *Frontiers in Integrative Neuroscience* 5. DOI: 10.3389/fnint.2011.00066.

Wolfrum, M. and O. E. Omel'chenko (2011). "Chimera states are chaotic transients". In: *Physical Review E* 84.1, 015201. DOI: 10.1103/physreve.84.015201.

Wolpert, D. M., Z. Ghahramani, and M. I. Jordan (1995). "An Internal Model for Sensorimotor Integration". In: *Science* 269.5232, pp. 1880–1882. DOI: 10.1126/science.7569931.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Yankowitz, M., S. Chen, H. Polshyn, Y. Zhang, K. Watanabe, T. Taniguchi, D. Graf, A. F. Young, and C. R. Dean (2019). "Tuning superconductivity in twisted bilayer graphene". In: *Science* 363.6431, pp. 1059–1064. DOI: 10.1126/science.aav1910.

Yartsev, M. M., M. P. Witter, and N. Ulanovsky (2011). "Grid cells without theta oscillations in the entorhinal cortex of bats". In: *Nature* 479.7371, pp. 103–107. DOI: 10.1038/nature10583.

Yi, T.-M., Y. Huang, M. I. Simon, and J. Doyle (2000). "Robust perfect adaptation in bacterial chemotaxis through integral feedback control". In: *Proceedings of the National Academy of Sciences* 97.9, pp. 4649–4653. DOI: 10.1073/pnas.97.9.4649.

Zeeman, E. C. (1977). *Catastrophe Theory: Selected Papers, 1972–1977*. Reading, MA: Addison-Wesley.

Zipser, D. and R. A. Andersen (1988). "A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons". In: *Nature* 331.6158, pp. 679–684. DOI: 10.1038/331679a0.

Index

adaptivity, 281
agent-centered measurement frame, 229
agent-indexed variable, 343
allocentric coordinates, 339
Arnold tongue, 68
attention schema, 414
autopoiesis, 272

beat
  envelope, 24
  frequency, 24
bispectrum, 672
Bistritzer–MacDonald model, 79
Buckingham Π theorem, 47

causal state, 139
centered world, 369
character, 368
chimera state, 118
circle map, 67
closed loop, 599
closure
  of constraints, 264
coarse-graining, 135
code, 265
combinatorial-state automaton, 175
commensurability, 65
constituent measure, 609
content, 289, 368
control, 255
control parameter, 102
corollary discharge, 225
counterfactual, 320
counterfactual encoding, 165
crossover, 48

detachment, 312
disciplinary amnesia, 555

effective information, 136
efference copy, 225
egocentric coordinates, 339
electric image, 201
epistemic cut, 255
explicit variable, 163
exploited mapping, 188

failure condition, 632
finite-state automaton, 172
free energy, variational, 329
frustration, 120
functional perspective, 425
functional present, 349

gain field, 341

higher-order theory, 396
HKB model, 102

implementation
  counterfactual, 174
  simple mapping account, 172
intermodulation, 32
invariance, 6

Kuramoto model, 92

latent computation, 188
linear time-invariant system, 32

measurement, 255
metacognition, 387
metastability, 110
misclassification, 234

misrepresentation, 300
moiré
  period, 14

neural complexity, 447
normativity, 280

open loop, 599
order parameter, 102, 128
  Kuramoto, 92
organization, 271

perspective
  functional, 345
phase–amplitude coupling, 458
phenomenal perspective, 426
planning, 321
polycomputing, 185
precariousness, 280

quality factor, 57

reader
  rungs of reading, 150
reader intervention, 589
reading loop, 152
record, 254
reference-free reafferent reader, 537
relation
  dynamically explicit, 39
  implicit, 39
relation clamp, 583
relational coordinates, 5
relational intervention, 580
relational observable, 4
relatum-changing response, 210
rotation number, 67

salience, 281
second-order reader, 387
self-indexed variable, 203
self-locator, 371
self-model, 404
semantic closure, 265
structural causal model, 319
structural coupling, 274
structural representation, 293

temporally centered variable, 352
theory of consciousness
  profile, 435
transparency, 405

Umwelt, 275
unification, candidate, 533

viability, 279