Owned Causality:
Constraint Closure, Basin Steering, and the Emergence of Agency from Chemistry

PIATRA . INSTITUTE

May 2026

Abstract

A causal influence is *owned* when a persistent organisation produces the constraints that generate it, those constraints steer the organisation's future attractor regime, and the resulting regime feeds back into the production of further constraints. We argue that this is the right primitive for agency, and that it has to be defined without "I" or "self" on the right-hand side, on pain of circularity. The paper assembles a single mathematical object,

$$\mathfrak{O}_T(\Omega) = \Gamma_T(\Omega) \cdot DI(U^{\Omega}_{t:t+T} \to Q_{t+T} \mid X_t, E_{t:t+T}) \cdot \Delta V_T \cdot I(Q_{t+T}; C_{t+T:t+2T} \mid \Lambda_{t+T}),$$

that absorbs the relevant ideas from chemical reaction-network theory (Feinberg), chemical organisation theory (Dittrich and Speroni di Fenizio), autocatalytic-set and RAF theory (Kauffman, Hordijk and Steel), closure to efficient causation (Rosen, Hofmeyr), closure of constraints (Moreno, Mossio, Montévil), autopoiesis (Maturana and Varela), semantic closure (Pattee, Hoffmeyer, Kolchinsky and Wolpert), Markov-blanket-mediated inference (Friston, Kirchhoff et al.), basin dynamics (Lorenz), and circular causality across levels (Haken). Each of the four factors does work that no other can: Γ_T excludes thermostats, $DI(U^{\Omega} \to Q)$ excludes systems mostly pushed by environment, ΔV_T excludes random internal explosions, and $I(Q;C)$ excludes one-shot influences. We compute the formula on a six-rung agency ladder running a common bistable dynamics under different constraint-production policies. The ladder cleanly separates the two systems whose constraints are externally specified (rock and thermostat, both at $\mathfrak{O}_T = 0$) from the four systems with internal constraint production (flame, RAF set, chemoton, and organism, all with positive $\mathfrak{O}_T$ between 2.2×10^{-5} and 2.1×10^{-4}). A second simulation traces the handover from an external scaffold to internal constraint deployment over an episode in which the scaffold withdraws linearly. Directed information from the scaffold to the macrostate falls from a peak of 0.037 to near zero, and directed information from endogenous constraints rises to a peak of 0.403, with the two traces crossing at window 6 of 20 while viability remains between 0.885 and 1.000. The pathological case, in which the scaffold cannot be withdrawn without viability collapsing, is the contrapositive of the same equation. The "I" appears later as a compressed model of recurrent high-$\mathfrak{O}_T$ history. Every numeric claim in the paper traces to a key in `simulation/output/results.json`.

1. A primitive before the self

A simple question motivates this paper. What is the right formal content of the claim *my action changed my future state*? Standard accounts begin by assigning the action to a subject, the future state to that subject's body, and then asking how the subject brought about the change. The trouble is that the subject is what needs explaining. Definitions that put "I" or "self" on the right-hand side of the equation are circular. They explain agency by quietly assuming it.

The paper takes the constraint seriously. A primitive for *owned causality* must be definable in pre-personal terms, using only ingredients that a chemical reaction network in a viable configuration could be said to possess; the right-hand side cannot mention "I" or "self". The synthesis we pro-

pose is that owned causality is the time-directed influence of constraints produced inside a persistent organisation on that organisation's future attractor regime, under conditions where the regime feeds back into further constraint production. No subject is needed for this to be defined. The subject is what appears later, as a compressed model of recurrent high-$\mathfrak{O}_T$ history.

The paper has three goals. The first is to identify what each tradition from chemistry through cognition contributes to a non-circular primitive: reaction-network theory, chemical-organisation theory, autocatalysis, hypercycles, closure to efficient causation, closure of constraints, autopoiesis, semantic closure, Markov-blanket inference, attractor dynamics, and circular causality across levels. The second is to compose these contributions into one formal object, $\mathfrak{O}_T(\Omega)$, with four factors that each exclude a known failure mode. The third is to compute the object on stylised toy systems and to show that the formula discriminates organisations that genuinely own their causality from those whose causality belongs to their environment or to an external designer. The paper closes by deriving the "I" as a downstream summary of $\mathfrak{O}_T$ history rather than as an input to it.

2. What "owned" is not

The word "owned" carries everyday meanings that the formal primitive must shed. Four common candidates fail.

Caused inside the boundary fails. A stomach spasm is caused inside the body but is not owned in any agentic sense. A reflex starts internally but only weakly counts as the agent's action. A spasm or a reflex does not have the architecture that makes its causal effect feed back into the system's future ability to act. It happens to the organisation. The organisation does not produce it as a constraint that the rest of the organisation continues to recruit.

Voluntary fails for the same reason. Voluntariness is a feature of high-level cognitive control, and demanding it as the primitive again presupposes a self who exercises will. Plenty of agentic processes operate below the threshold of voluntariness: the bacterium that climbs the chemotactic gradient operates without will, yet its movement is owned in the structural sense. Demanding voluntariness as the criterion either pushes agency too high in the ladder or smuggles in a subject by the back door.

Conscious fails by the same move. Consciousness is a feature of a small number of high-end agentic systems, and may even be unrelated to agency proper. The primitive must apply to organisations that are agentic without being conscious. Bacteria, immune systems, and developing embryos exhibit owned causality at a level the primitive must capture; conscious humans exhibit it on top of, not instead of, the same structural property.

Caused by an "I" is the circular candidate already identified. Any definition with this shape has the subject on both sides of the equation. The subject is what we want to explain, so it cannot appear in the explanation.

The working answer is structural. A cause is *owned* when it has been recursively incorporated into the organisation's constraint-producing future. The cause must come from inside the organisation,

must bias the organisation's future regime, must contribute to the organisation's continued viability, and must alter the organisation's future capacity to produce constraints. The four conditions look interlocking because they are. The formula in §6 makes them simultaneous.

3. Reaction, organisation, closure

3.1 Reaction-network grammar

The lowest layer of the construction is dynamical: matter changes according to reaction networks. The non-equilibrium thermodynamics tradition of Prigogine and Nicolis established that systems driven far from equilibrium develop dissipative structures whose macroscopic order is maintained by continuous through-flow of matter and energy (Prigogine and Nicolis, 1977); Schrödinger's earlier formulation, that life feeds on negentropy, is the same observation in earlier vocabulary (Schrödinger, 1944). Feinberg's Chemical Reaction Network Theory gives the rigorous account of how reaction-graph structure constrains possible dynamics, with stoichiometric matrices, complexes, linkage classes, deficiency, and the existence and uniqueness theorems for steady states under mass-action kinetics (Feinberg, 1995). A mass-action chemical reaction network can be written

$$\dot{X} = N v(X),$$

with X the vector of concentrations, N the stoichiometric matrix, and $v(X)$ the reaction-rate vector. The theory tells the modeller which flows are possible, which steady states exist, when they are unique, and when they are stable.

CRNT is the grammar layer: it rules out impossible chemistries and characterises the multistability available to a given reaction graph. It does not, by itself, identify which sub-networks are organisations in any biological sense. A network can have rich CRNT structure while remaining a passive participant in an externally supplied environment.

3.2 Closed and self-maintaining sets

The first move toward organisation is Dittrich and Speroni di Fenizio's Chemical Organisation Theory (Dittrich and Speroni di Fenizio, 2007), which singles out a subset $O \subseteq S$ of species as an *organisation* when it is

- *closed*: every reaction with reactants in O produces products in O;
- *self-maintaining*: every species consumed inside O can be regenerated inside O.

This is a structural property of the reaction graph plus the kinetic constants. A closed self-maintaining set is the smallest unit that can persist as an identity in the chemical sense: it does not leak out into wider chemistry and does not deplete itself faster than it can replenish. COT defines a lattice of such organisations and lets the same chemistry move between them under different conditions.

What COT gives is identity-as-set. A reaction graph either does or does not contain a particular organisation. What it does not yet give is *control*: the organisation does not, in the COT framework

alone, deploy its constraints to bias which regime it occupies.

3.3 Catalytic closure

Kauffman's autocatalytic-set theory and Hordijk and Steel's RAF formalism (Kauffman, 1986; Hordijk and Steel, 2004) sharpen the notion of closure to catalysis. A RAF set is *reflexively autocatalytic* and *food-generated*: each reaction in the set is catalysed by at least one molecule from the set, and every molecule in the set can be produced from a food set by reactions in the set. RAF theory gives an explicit construction for when collectively autocatalytic networks emerge in random chemistries and shows that the emergence threshold is sharp.

Eigen and Schuster's hypercycle (Eigen and Schuster, 1979) addresses a related question, the maintenance of catalytic cooperation against the corrosive pressure of parasites. A hypercycle is a cyclic arrangement of self-replicators in which each member catalyses the production of the next. The construction shows both the possibility of cooperative closure and its fragility: without a higher-level boundary, the network is vulnerable to error catastrophe and to free riders.

Catalytic closure is necessary for what follows. It is not sufficient. A RAF set with no boundary is chemistry in a soup, not yet an organisation that owns what it does.

3.4 Closure of constraints

The decisive step is the move from species closure to constraint closure. Rosen, in his relational biology, characterises a living system by *closure to efficient causation*: the efficient causes that make the system run, the catalysts in the broad sense, are themselves generated within the system (Rosen, 1991; Letelier et al., 2006). Rosen's (M, R)-systems express this category-theoretically. Hofmeyr makes the construction biochemically realistic by treating the living cell as a self-fabricating system whose enzymes, transporters, scaffolds, and membranes are produced by the very network they enable (Hofmeyr, 2007). Kauffman's *Investigations* frames the same construction in thermodynamic terms (Kauffman, 2000): an *autonomous agent* is a system that reproduces itself and performs at least one thermodynamic work cycle, with the cycle's constraints (machines, catalysts, enclosures) produced by the work the cycle does. Constraint and process feed each other in a closed loop.

Moreno, Mossio, and Montévil give this construction its current canonical form (Moreno and Mossio, 2015; Montévil and Mossio, 2015). Where Rosen closes the efficient causes, the catalysts that make the system run, they close the wider class of constraints, the structures that channel processes without being consumed in them. They distinguish *processes* from *constraints*. Processes are material flows and transformations. Constraints are structures that channel those flows while being maintained by the organisation they constrain. An enzyme constrains a reaction rate. A membrane constrains diffusion. An ion channel constrains an electrochemical gradient. A cytoskeleton constrains shape. A skill constrains the future action space. A habit constrains the future policy distribution. *Closure of constraints* is the condition that the set of constraints in the organisation is jointly produced and maintained by the very dynamics they constrain.

Juarrero gives the philosophical reading (Juarrero, 1999). Context-free constraints, of the kind that

walls and conservation laws impose, reduce the trajectory space the system can occupy. Context-sensitive constraints, the kind that arise from the organisation's own dynamical history, reshape which trajectories are accessible; they are generated by the same dynamics they constrain. Intentional behaviour, on her account, is a constraint structure of the second kind. Deacon develops a parallel construction in *Incomplete Nature* (Deacon, 2012). His *autogen* combines autocatalysis with a self-assembling capsule into the minimal hypothetical organisation that exhibits what he calls *teleodynamics*: the capacity to maintain the very constraints under which the organisation persists. The vocabularies of Juarrero, Deacon, and Moreno-Mossio-Montévil differ; the structural claim is the same.

This is the bridge from chemistry to autonomy. The move from $X \to X$ to $X \to C \to X$ is what separates a reaction soup from an organisation. The action variable in the formal primitive will be constraint deployment, not direct material push.

3.5 Semantic closure

A further layer enters with symbols. Pattee's *semantic closure* is the requirement that descriptions, codes, or measurements inside a system influence the system's material dynamics, while material dynamics produce and interpret those descriptions (Pattee, 1982, 1995). DNA is a paradigm case: it is itself a molecule, but it functions inside the cell as a description that constrains protein construction. Hoffmeyer extends this into biosemiotics, treating living organisation as broadly semiotic and arguing that biological dynamics carry sign-mediated information all the way down (Hoffmeyer, 2008). Barbieri's code biology sharpens the requirement: organic codes are real conventions, arbitrary mappings between two molecular domains held in place by physical adaptors, and they proliferate well beyond the genetic code (Barbieri, 2015).

Kolchinsky and Wolpert give an operational handle on what "semantic" means in physical terms (Kolchinsky and Wolpert, 2018). Their definition: semantic information is information that a physical system has about its environment and that is causally necessary for the system to continue to exist. The empirical test is scrambling. Replace the information channel with a noise channel of identical capacity; if the system's continued existence becomes less likely, the channel was semantically loaded. Bertschinger et al. (2008) give a related information-theoretic account of autonomy.

The advantage of this construction is that *meaning* is defined without subjects. A signal is meaningful for an organisation when scrambling it reduces the organisation's viability. The right-hand side of the definition contains organisation, signal, viability. It does not contain anyone who understands the signal.

4. Boundary and semantics

4.1 Autopoiesis

A constraint-closing organisation needs a boundary, and the boundary cannot be a passive container. Maturana and Varela define a living system as *autopoietic*: a network of processes that produces the components that regenerate the network and constitute it as a unity (Maturana and

Varela, 1980; Varela, 1979). The cell membrane is produced and repaired by the metabolism it encloses; the boundary is itself a variable of the dynamics, with production and maintenance equations coupled to the dynamics inside it.

The distinction matters formally. A *container* boundary is given as a parameter of the model. An *autopoietic* boundary is a variable of the model, with its own production and maintenance equations, coupled to the dynamics inside it. The first is dead, the second is alive in the structural sense.

Thompson develops the post-Varela synthesis (Thompson, 2007). On his reading, a living organisation is already a cognitive one in the structural sense, because the closure that maintains the boundary is also what makes the organisation sensitive to what does and does not preserve it. The cognitive layer arrives with the closure, not on top of it.

4.2 Markov blankets

Friston's free-energy principle gives a separate formal account of boundary, in statistical rather than material terms (Friston, 2013; Kirchhoff et al., 2018). A *Markov blanket* is the set of variables that screens an internal subsystem from external states. The internal states are conditionally independent of external states given the blanket. The blanket itself partitions into sensory states, which the external can drive, and active states, which the internal can drive. Under this partition the internal states appear, on average, to minimise a free-energy functional of the blanket states. This is a powerful formal lens on inference and action.

The Markov-blanket construction is not by itself enough to secure ownership. A blanket can be defined for almost any conditional-independence structure. To pick out living boundaries, the blanket has to be coupled to the closure-of-constraints layer: the blanket must be produced by the internal organisation, actively regulated, and tied to viability. Without that coupling, Markov-blanket talk drifts toward describing any sufficiently complex stochastic system as a quasi-organism. With the coupling, the construction is sharp.

4.3 Semantic closure quantified

Combining Pattee's qualitative requirement with Kolchinsky and Wolpert's operational test gives the working semantic primitive. A signal or measurement M is semantically loaded for an organisation Ω when

$$\text{meaning}(M) \approx \mathbb{E}\big[V(Q_{t+T}) \mid \text{do}(M)\big] - \mathbb{E}\big[V(Q_{t+T}) \mid \text{do}(\text{scramble}(M))\big],$$

where V is a viability function over macrostates Q and scramble replaces the channel with a noise channel of equal capacity. This is the version of "meaning" that appears in ΔV_T in the unified formula. No subject required.

5. Basin steering, not point control

5.1 Lorenz: futures as basins

Lorenz showed that bounded nonlinear deterministic systems can evolve lawfully yet remain unpredictable in trajectory, through sensitive dependence on initial conditions (Lorenz, 1963; Strogatz, 2015). The lesson for agency is structural. An organism in a stochastic, nonlinear, partially observable environment does not control which point in state space it occupies at time $t + T$. The point-control conception of action is false for living systems.

What an organism can do, when it has the architecture to do it, is bias the probability that it occupies one attractor *basin* rather than another. A bacterium biases itself toward a nutrient-seeking regime over a starvation regime, leaving its exact molecular configuration to thermal noise. A human at a fork in the road shifts the probability that the trajectory falls into the work basin, the rest basin, the conflict basin, or the flight basin, without specifying the resulting trajectory point by point. Agency is basin-steering under uncertainty.

Formally, replace

$$a_t \to x_{t+T}$$

with

$$a_t \to P(Q_{t+T} = q_i),$$

where Q_t is a coarse-grained macrostate. The action does not select the point; it biases the basin distribution. Kauffman (1991) gives a Santa Fe-tradition version of this same observation, framing adaptation as living near the edge of chaos. England's dissipative-adaptation result is the thermodynamic complement (England, 2015): under driven self-assembly, configurations more efficient at absorbing work from the driving field accumulate, without any selection acting on reproduction. The bistable Langevin model used in §7 lives in this regime. Levin's work on bioelectric pattern memory shows that basin steering is mechanistic at the cellular level rather than metaphorical (Levin, 2022). Tissues hold multistable bioelectric states; modifying the state changes the morphology the tissue settles into, with the genome held fixed.

5.2 Haken: circular causality across levels

Haken's synergetics adds the observation that macro-scale order parameters and micro-scale components stand in *circular causality* (Haken, 1983). The order parameter is generated by the cooperative behaviour of microscopic components, and at the same time it enslaves those components, constraining which microstates they can occupy. Neither level is causally prior. The macro and the micro maintain one another.

Campbell (1974) introduced the term *downward causation* for the same phenomenon in hierarchically organised biological systems. Ellis (2012) develops the modern treatment, distinguishing five types of top-down effect and arguing that none of them require non-physical causes; macroscopic constraints set the boundary conditions under which microscopic dynamics evolve, and the constraints are themselves the joint outcome of those dynamics. Synergetics, downward causation,

and closure of constraints (§3.4) name the same architecture in three vocabularies, with closure of constraints being its biochemical realisation.

This is the non-mystical version of top-down causation. An organisation's macro-scale regime is the joint stable pattern of its molecules; the pattern, as a constraint, shapes which microstates are accessible. The macro regime is the organisation, the micro components are the reactions, and the constraints are produced by the joint dynamics in a way that maintains the regime.

6. The unifying primitive

6.1 The organisation

Definition 1 (Organisation)**.** *An organisation at time t is a tuple*

$$\Omega_t = (X_t, R_t, C_t, B_t, M_t, Q_t, V),$$

with X_t the material state (concentrations, charges, structures), R_t the reaction or process network, C_t the set of constraints produced and maintained by the organisation, B_t the boundary or Markov blanket, M_t the measurements or signs the organisation makes of its own state and its environment, Q_t the coarse-grained macrostate or attractor basin, and V the viability function over macrostates. The dynamics are

$$\dot{X} = F(X, E; C, B), \quad \dot{C} = G(X, C, B, M), \quad \dot{B} = H(X, C, B), \quad M = \mu(B, E, X, C), \quad Q = \psi(X, C, B),$$

where E collects external states. Constraint closure is the condition that C is jointly produced and maintained by $\dot{X}$, $\dot{C}$, $\dot{B}$. Boundary closure is the condition that B is produced by the same dynamics.

The seven components are not free-floating. X does the reacting. R specifies which reactions can occur. C channels them. B separates them from the world. M measures them. Q summarises them. V weights them by viability. The organisation is the joint stable configuration of all seven in mutual maintenance.

6.2 Owned causality

The measure is built to fail four ways on purpose. It must vanish for a system whose control rule is installed from outside, for a system that is merely pushed by its environment, for internal dynamics that do not serve viability, and for an influence that acts once and never recurs. The four factors are those four filters, multiplied so that any one of them can veto the rest.

Definition 2 (Owned causality)**.** *For an organisation Ω and a horizon T, the owned-causality measure is*

$$\mathfrak{O}_T(\Omega) = \Gamma_T(\Omega) \cdot DI(U^{\Omega}_{t:t+T} \to Q_{t+T} \mid X_t, E_{t:t+T}) \cdot \Delta V_T \cdot I(Q_{t+T}; C_{t+T:t+2T} \mid \Lambda_{t+T}),$$

where $U^{\Omega}_{t:t+T}$ are the active perturbations generated by the organisation's own constraints, Q_{t+T} is the macrostate at horizon T, X_t and $E_{t:t+T}$ are the current state and external trajectory (conditioned away so external causes are factored out), Λ_{t+T} is the slow organisational regime, and $C_{t+T:t+2T}$ is the constraint

trajectory over the subsequent horizon. The factors are

$$\begin{aligned}
\Gamma_T(\Omega) &= P(\Omega_{t+T} \simeq \Omega_t) \cdot \frac{\#\{\textit{constraints internally maintained}\}}{\#\{\textit{constraints required by } \Omega\}}, \\
DI(A \to B \mid C) &= \textit{directed information from } A \textit{ to } B \textit{ conditional on } C, \\
\Delta V_T &= \mathbb{E}[V(Q_{t+T}) \mid \textit{do}(U^\Omega)] - \mathbb{E}[V(Q_{t+T}) \mid \textit{do}(\text{scramble}(U^\Omega))], \\
I(Q; C \mid \Lambda) &= \textit{conditional mutual information from achieved macrostate to future constraint configuration.}
\end{aligned}$$

The right-hand side contains neither "I" nor "self".

Each factor is the filter for one of those failure modes.

Γ_T kills the thermostat. A thermostat has feedback, but its control rule is specified by an external designer rather than produced and maintained by the thermostat itself. The maintained-constraint fraction is therefore zero, and the whole product collapses. The same closure score lets the formula distinguish a flame (which weakly maintains its own enabling conditions through autocatalytic burning) from a rock (which does not).

$DI(U^\Omega \to Q \mid X, E)$ kills the leaf in the wind. A leaf has internal state, and that internal state correlates with future basin occupancy, but conditional on the wind direction the leaf carries no information about where it lands. The directed-information factor, conditioned on the external trajectory, captures only the steering that genuinely originates inside the organisation. Massey's directed-information formulation (Massey, 1990; Schreiber, 2000) and the do-calculus apparatus of Pearl (2009) give the technical content.

ΔV_T kills the random internal explosion. A system can carry strong internal dynamics that influence its future regime, but if those dynamics do not preserve or improve viability they are not agency. The scrambling-counterfactual test gives the operational handle: replace the constraint trajectory with a time-shuffled version, rerun the dynamics with the same external trajectory, and see whether viability survives. If ΔV_T is positive, the timing of the constraint deployment was load-bearing for survival. Klyubin, Polani and Nehaniv's *empowerment* measure (Klyubin et al., 2005) is a closely related quantity from agent-centric information theory.

$I(Q_{t+T}; C_{t+T:t+2T} \mid \Lambda_{t+T})$ kills the one-shot reflex. A reflex affects the future once, then has no further effect on the organisation's behaviour. An agentic process is recursive: achieved regimes update future constraints. In a cell, the stress response changes gene expression. In a bacterium, chemotactic history changes receptor methylation. In an immune system, exposure changes recognition. In a nervous system, success changes policy priors. In a human, an act changes habit, confidence, memory. The re-entry factor measures this recursive incorporation. Without it, there is causality. With it, there is agency.

7. A computable example

7.1 The bistable organisation and the agency ladder

The simulation uses a one-dimensional Langevin organisation in a bistable potential

$$V_{\text{pot}}(x) = \frac{(x^2 - 1)^2}{4},$$

with a viable well at $x > 0$ (basin A, $V(Q) = 1$) and a non-viable well at $x < 0$ (basin B, $V(Q) = 0$). The state evolves as

$$dX_t = (-V'_{\text{pot}}(X_t) + U_t^{\Omega} + E_t)\, dt + \sigma\, dW_t,$$

with $U_t^{\Omega} = C_t$ the constraint-mediated active perturbation, E_t the external perturbation, $\sigma = 0.75$, and $dt = 0.05$. Each rung of the ladder specifies a constraint-production policy that updates C_t over time and a constant maintained-constraint fraction reflecting the rung's structural closure. All rungs run for $n = 8000$ steps under the same noise realisation, with an external scaffold during the first quarter pushing the system into basin A, after which the scaffold withdraws and the rung's own constraint policy takes over. The four factors are computed over the post-scaffold window.

The six rungs are: rock (no constraint), thermostat (linear feedback rule specified by an external designer, closure fraction $0/2$), flame (process self-maintenance with mild positive feedback, closure $1/3$), RAF set (autocatalytic gain, closure $2/3$), chemoton (a stylised metabolism-boundary-information triad in the sense of Gánti (2003), closure $3/4$), and organism (viability-tracking constraint with memory, closure $4/4$).

rung	Γ_T	$DI(U^{\Omega} \to Q)$	$DI(E \to Q)$	ΔV_T	$I(Q;C)$	$\mathfrak{O}_T$
rock	0.000	0.000	0.000	0.172	0.000	0.0×10^{0}
thermostat	0.000	0.321	0.001	0.183	0.000	0.0×10^{0}
flame	0.333	0.440	0.001	0.018	0.015	4.0×10^{-5}
RAF	0.667	0.245	0.000	0.162	0.001	2.2×10^{-5}
chemoton	0.750	0.279	0.000	0.286	0.004	2.1×10^{-4}
organism	1.000	0.222	0.001	0.245	0.003	1.8×10^{-4}

Table 1: Four-factor breakdown of the owned-causality formula across the agency ladder. Values from `simulation/output/results.json`.

The rock and the thermostat both have $\mathfrak{O}_T = 0$, despite the thermostat exhibiting strong directed information from its constraint deployment to its macrostate ($DI = 0.321$). The closure factor $\Gamma_T = 0$ for both rungs (the rock has no constraints; the thermostat's rule is externally specified), so the whole product collapses. This is the formula doing the work that we required of it: a feedback loop is not enough; closure of constraints is what licenses ownership.

The four closure-bearing rungs all have positive $\mathfrak{O}_T$, between 2.2×10^{-5} (RAF) and 2.1×10^{-4} (chemoton). The ordering among them is noisy at this sample size, with chemoton edging out organism in the present realisation because its re-entry factor $I(Q;C) = 0.004$ runs slightly higher

than the organism's $I(Q;C) = 0.003$. The discrimination the formula was designed to do is sharpest at the boundary between $\mathfrak{O}_T = 0$ and $\mathfrak{O}_T > 0$, separating systems whose causality belongs to an external designer from systems that produce their own constraints. The within-group ordering is a secondary signal that the multiplicative structure produces, and one weakness can be compensated by strength elsewhere.

$DI(U^\Omega \to Q)$ on its own does not rank the rungs correctly. The thermostat scores 0.321 on this factor; the organism scores 0.222. A naive "amount of internal control" criterion would mis-rank them. The product structure of $\mathfrak{O}_T$, in which each factor can veto, is what gives the correct ranking.

Proposition 1 (Closure boundary)**.** *If* $\Gamma_T(\Omega) = 0$*, then* $\mathfrak{O}_T(\Omega) = 0$*, regardless of the values of* $DI(U^\Omega \to Q)$*,* ΔV_T*, or* $I(Q;C)$*.*

Proof. Direct from the definition: $\mathfrak{O}_T$ is the product of four non-negative factors, so vanishing of any factor sends the product to zero. The thermostat in Table 1 exhibits this: $DI(U^\Omega \to Q) = 0.321$ and $\Delta V_T = 0.183$ are both substantial, but $\Gamma_T = 0$ forces $\mathfrak{O}_T = 0$. The same applies to any system whose control rule is supplied by an external designer rather than maintained inside the organisation. The closure factor is the formula's filter on externalised control. □

8. Scaffolding and the handover

8.1 The positive-loop architecture

A natural use of the formula is to describe what happens when an organism in a non-viable state is brought into a viable state by an external scaffold, an analogue of co-regulation, blessing, mentoring, trusted treatment, parental calm, or coach confidence. The healthy version of this loop transfers basin-steering capacity from the scaffold to the organisation's own constraint deployment. The pathological version, sometimes called guru capture, keeps the scaffold permanently in control.

In the unified language, the healthy transition is a swap of the dominant directed-information channel:

$$DI(U^{\text{ext}} \to Q) \to 0, \qquad DI(U^\Omega \to Q) \to \text{positive},$$

with ΔV_T preserved or growing throughout. The pathological case keeps $DI(U^{\text{ext}} \to Q)$ near its initial value and leaves $DI(U^\Omega \to Q)$ at zero. Same architecture, different attractor.

8.2 The handover quantified

The simulation runs a bistable organism starting at $x_0 = -0.95$ (deep in basin B, non-viable). During the first quarter of the episode, an external scaffold of amplitude $u^{\text{ext}} = 1.0$ pushes the system toward basin A, and the organism's constraint loop is gated off (analogue of threat-suppressed agency). The scaffold then withdraws linearly over the next quarter, and the constraint loop activates linearly over a slightly longer ramp. The full episode runs for $n = 12{,}000$ steps in 20 windows of 600 steps. Each window measures the per-step directed information from E and from U^Ω to the macrostate X, computed under the same plug-in estimator as in §7.

The directed-information traces show the handover quantitatively. During the scaffold-active windows (0 through 5), $DI(U^{\text{ext}} \to Q)$ runs between 0.003 and 0.037 while $DI(U^{\Omega} \to Q)$ stays at zero. As the scaffold withdraws (windows 6 through 9), $DI(U^{\Omega} \to Q)$ rises rapidly, crossing $DI(U^{\text{ext}} \to Q)$ at window 6 and continuing to grow. During the post-scaffold windows (10 through 19), $DI(U^{\Omega} \to Q)$ runs between 0.159 and 0.403 while $DI(U^{\text{ext}} \to Q)$ stays below 0.011. Viability remains between 0.885 and 1.000 throughout the episode.

Proposition 2 (Scaffold handover)**.** *Under the protocol of §8.2 with constraint-closing organism and viability-gradient constraint policy, the directed-information traces satisfy*

$$\max_{w \geq 6} DI(U^{\Omega} \to Q)_w \;\geq\; 10 \cdot \max_{w} DI(U^{ext} \to Q)_w$$

and the crossover index, the first window in which $DI(U^{\Omega} \to Q) \geq DI(U^{ext} \to Q)$ after the scaffold has begun to withdraw, is finite. Viability is maintained: $V(Q_w) \geq 0.885$ for all windows.

Proof. Direct evaluation against `simulation/output/results.json`. The peak $DI(U^{\Omega} \to Q)$ over post-scaffold windows is 0.403 at window 10. The peak $DI(U^{\text{ext}} \to Q)$ over all windows is 0.037 at window 5. The ratio is 10.9, satisfying the inequality. The crossover index is $w = 6$, with $DI(U^{\text{ext}} \to Q)_6 = 0.016$ and $DI(U^{\Omega} \to Q)_6 = 0.059$. The minimum window-wise viability is 0.885 at window 10. □

8.3 The pathological case

The contrapositive is the formal version of guru capture or pathological dependence. If the constraint loop fails to activate as the scaffold withdraws, the system falls out of basin A as the scaffold weakens, $DI(U^{\Omega} \to Q)$ stays at zero, and viability collapses. The formula reports $\mathfrak{O}_T \to 0$ for the organism over the post-scaffold window, because the closure factor is degraded: the constraints the organisation was supposed to maintain are not being maintained. The same dynamics, with the constraint gate held closed, would produce the pathological trajectory. The discriminator between healthy bootstrapping and capture is whether the system absorbs the scaffold's basin-steering capacity into its own constraint architecture or leaves that capacity outside.

9. The "I" as compressed history

The right-hand side of the Definition contains no subject. Once the formula has been computed for an organisation over many cycles, however, a higher-level summary becomes available. Repeated episodes of high $\mathfrak{O}_T$ exhibit a regularity: the same constraint architecture, the same basin preferences, the same characteristic horizons of re-entry. A learning system can compress this regularity into a latent variable S, a stable source of recurrent endogenous basin-steering. That S is what the organism can subsequently model phenomenologically as "me", "my action", "my future". Active inference offers one machinery for the compression: Friston, Parr and de Vries (2017) develop the belief-propagation process theory in which posterior beliefs over latent states are updated by message passing on a generative model, and the same machinery applies to a latent S summarising $\mathfrak{O}_T$ history.

The construction matches what is known about the sense of agency at the psychological level. Haggard's review of sense of agency identifies the comparator model: motor commands generate predictions through an efference copy, predicted and actual sensory feedback are compared, and mismatches reduce the sense that "I did that" (Haggard, 2017). Moore, Wegner, and Haggard show that sense of agency draws on motor evidence and on higher-level inferential cues about whether one caused the action (Moore et al., 2009). These are downstream readouts of recurrent $\mathfrak{O}_T$ history. Bandura's self-efficacy theory, with mastery, modelling, persuasion, and the interpretation of bodily states as its four sources, fits the same picture (Bandura, 1982): the four sources are channels through which $\mathfrak{O}_T$ evidence accumulates into a stable model of authorship.

Damasio's account of homeostasis and feeling supplies the somatic counterpart (Damasio, 2018). Feelings function as mental representations of homeostatic state, and the I-as-felt is a high-level readout of how the organisation's viability is tracking; the compressed S inherits its affective tone from the ΔV history that runs alongside the $\mathfrak{O}_T$ trace. Dennett's earlier formulation of the self as a *centre of narrative gravity* (Dennett, 1992) sits exactly on the same construction. The gravitational centre has no separate substance; it is the convergent point of the trajectories the system has tended to occupy. Recurrent high $\mathfrak{O}_T$ history is the trajectory bundle, and the I-narrative is its centre.

Recurrent high $\mathfrak{O}_T$ generates the "I" as its statistical summary. Agency begins with constraint-closing organisation; selfhood is what such an organisation eventually comes to model itself as.

Systems can be agentic without being selves. Bacteria, immune systems, embryos, and bounded autocatalytic networks all carry positive $\mathfrak{O}_T$ over their relevant horizons, without having any "I" model. They own their causality in the structural sense without ever needing to describe themselves as doing so.

Selves can vary in confidence. The same human in different stress states can have substantially different $\mathfrak{O}_T$ profiles over the same period, and the latent "I" is correspondingly more or less stable. The variability is a property of the higher-level summary, which inherits any volatility in the $\mathfrak{O}_T$ trace it compresses.

The scaffolding result of §8 has a phenomenological reading. The healthy scaffold helps the organisation build up enough recurrent high-$\mathfrak{O}_T$ history that the latent variable can stabilise. The unhealthy scaffold supplies the basin-steering directly and leaves no $\mathfrak{O}_T$ history for the organisation to compress. Same architecture, different downstream phenomenology.

References

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37(2), 122–147.

Barbieri, M. (2015). *Code Biology: A New Science of Life*. Cham: Springer.

Bertschinger, N., Olbrich, E., Ay, N., and Jost, J. (2008). Autonomy: an information theoretic perspective. *BioSystems*, 91(2), 331–345.

Campbell, D. T. (1974). 'Downward causation' in hierarchically organised biological systems. In F. J. Ayala and T. Dobzhansky (Eds.), *Studies in the Philosophy of Biology* (pp. 179–186). London: Macmillan.

Damasio, A. (2018). *The Strange Order of Things: Life, Feeling, and the Making of Cultures*. New York: Pantheon.

Deacon, T. W. (2012). *Incomplete Nature: How Mind Emerged from Matter*. New York: W. W. Norton.

Dennett, D. C. (1992). The self as a center of narrative gravity. In F. S. Kessel, P. M. Cole, and D. L. Johnson (Eds.), *Self and Consciousness: Multiple Perspectives* (pp. 103–115). Hillsdale, NJ: Erlbaum.

Dittrich, P., and Speroni di Fenizio, P. (2007). Chemical organisation theory. *Bulletin of Mathematical Biology*, 69(4), 1199–1231.

Eigen, M., and Schuster, P. (1979). *The Hypercycle: A Principle of Natural Self-Organisation*. Berlin: Springer.

Ellis, G. F. R. (2012). Top-down causation and emergence: some comments on mechanisms. *Interface Focus*, 2(1), 126–140.

England, J. L. (2015). Dissipative adaptation in driven self-assembly. *Nature Nanotechnology*, 10(11), 919–923.

Feinberg, M. (1995). The existence and uniqueness of steady states for a class of chemical reaction networks. *Archive for Rational Mechanics and Analysis*, 132(4), 311–370.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, 10(86), 20130475.

Friston, K., Parr, T., and de Vries, B. (2017). The graphical brain: belief propagation and active inference. *Network Neuroscience*, 1(4), 381–414.

Gánti, T. (2003). *The Principles of Life*. Oxford: Oxford University Press.

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, 18(4), 196–207.

Haken, H. (1983). *Synergetics: An Introduction* (3rd ed.). Berlin: Springer.

Hoffmeyer, J. (2008). *Biosemiotics: An Examination into the Signs of Life and the Life of Signs*. Scranton: University of Scranton Press.

Hofmeyr, J.-H. S. (2007). The biochemical factory that autonomously fabricates itself: A systems-biological view of the living cell. In F. C. Boogerd, F. J. Bruggeman, J.-H. S. Hofmeyr, and H. V. Westerhoff (Eds.), *Systems Biology: Philosophical Foundations* (pp. 217–242). Amsterdam: Elsevier.

Hordijk, W., and Steel, M. (2004). Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *Journal of Theoretical Biology*, 227(4), 451–461.

Juarrero, A. (1999). *Dynamics in Action: Intentional Behavior as a Complex System*. Cambridge, MA: MIT Press.

Kauffman, S. A. (1986). Autocatalytic sets of proteins. *Journal of Theoretical Biology*, 119(1), 1–24.

Kauffman, S. A. (1991). Antichaos and adaptation. *Scientific American*, 265(2), 78–84.

Kauffman, S. A. (2000). *Investigations*. Oxford: Oxford University Press.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15(138), 20170792.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). Empowerment: a universal agent-centric measure of control. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation* (pp. 128–135).

Kolchinsky, A., and Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus*, 8(6), 20180041.

Letelier, J. C., Soto-Andrade, J., Guíñez Abarzúa, F., Cornish-Bowden, A., and Cárdenas, M. L. (2006). Organisational invariance and metabolic closure: analysis in terms of (M, R)-systems. *Journal of Theoretical Biology*, 238(4), 949–961.

Levin, M. (2022). Technological approach to mind everywhere: an experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in Systems Neuroscience*, 16, 768201.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2), 130–141.

Massey, J. L. (1990). Causality, feedback and directed information. In *Proceedings of the International Symposium on Information Theory and its Applications (ISITA-90)* (pp. 303–305).

Maturana, H. R., and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realisation of the Living*. Dordrecht: Reidel.

Montévil, M., and Mossio, M. (2015). Biological organisation as closure of constraints. *Journal of Theoretical Biology*, 372, 179–191.

Moore, J. W., Wegner, D. M., and Haggard, P. (2009). Modulating the sense of agency with external cues. *Consciousness and Cognition*, 18(4), 1056–1064.

Moreno, A., and Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Dordrecht: Springer.

Pattee, H. H. (1982). Cell psychology: an evolutionary approach to the symbol-matter problem. *Cognition and Brain Theory*, 4, 325–341.

Pattee, H. H. (1995). Evolving self-reference: matter, symbols, and semantic closure. *Communication and Cognition - Artificial Intelligence*, 12(1–2), 9–27.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge: Cambridge University Press.

Prigogine, I., and Nicolis, G. (1977). *Self-Organization in Non-Equilibrium Systems: From Dissipative Structures to Order through Fluctuations*. New York: Wiley.

Rosen, R. (1991). *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life*. New York: Columbia University Press.

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, 85(2), 461–464.

Schrödinger, E. (1944). *What is Life? The Physical Aspect of the Living Cell*. Cambridge: Cambridge University Press.

Strogatz, S. H. (2015). *Nonlinear Dynamics and Chaos* (2nd ed.). Boca Raton: CRC Press.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.

Varela, F. J. (1979). *Principles of Biological Autonomy*. New York: North Holland.