

**Construct Thyself:
From Self-Production to Reflexive Freedom**

PIATRA . INSTITUTE

August 2026

Abstract

A familiar argument infers from the constructed character of the self that the self and its freedom are illusions. We reject the inference. Constructedness concerns how something comes to exist and persist; illusion concerns whether an appearance misleads; and what is real about a self is an organization that produces, maintains, models and sometimes rebuilds itself. We assemble a discontinuous ladder of organizational capacities (self-production, closure, adaptivity, ultrastable reorganization, anticipation and self-modelling), each necessary for the next and none sufficient, and define a further threshold, constitutive access: the capacity to model variables of one's own organization, identify interventions on them, and select an intervention for its anticipated effect on one's own future agency. We compute the resulting constructability horizon in a minimal two-niche world. Below the threshold an agent can be changed into 5 configurations and can bring about none; at the threshold the two sets coincide; above it, aimed reach grows to 30 and 120, and about half of the added reach consists of states that end further revision. After a niche collapse, blind reorganization survives in 0.78 of trials, incremental self-editing never survives, and whole-template reconstruction survives in proportion to the accuracy of its model, exceeding the blind loop above an accuracy of about 0.72. An external controller that prunes access reconstructs the agent as effectively as self-reconstruction, and the difference appears only at a second shock. Aspiration is rational as provisional commitment whose revert rule protects viability and the capacity to revise.

Introduction

The self has often been declared missing, usually on sound grounds. No indivisible author stands behind the machinery. The genetic, developmental, neural and social processes that produce a person precede their every act and run without consulting them. Introspection reports a simple, continuous, self-originating subject that the mechanisms do not contain. From these premises the conclusion is drawn that the self is an illusion and that the freedom attributed to it falls with it.

The premises are sound and the conclusion does not follow. Whether something is constructed concerns how it comes to exist and persist; whether it is illusory concerns whether its appearance systematically misleads. Bridges and constitutions are constructed, and an organism constructs and maintains itself continuously, without any of these being unreal. Dennett's real patterns supply the relevant ontology: a pattern can be real, predictively indispensable and causally implicated without being metaphysically fundamental, and languages, promises, institutions and organisms all belong to that category (Dennett 1991). Fundamental, constructed and real are three different predicates, and the self is non-fundamental, constructed and real.

The eliminativist evidence supports a narrower claim. The word "self" names at least four explanatory objects: the organismic self, the bounded metabolic organization that keeps itself in existence; the control self, the regulatory architecture of goals, memory and action selection; the phenomenal self-model, the experience of being someone; and the practical and social self, the person as commitments, habits and recognized standing. The strongest illusion claims concern the third.

Metzinger's argument that phenomenal selfhood is the operation of a self-model transparent to itself, so that the system mistakes the model for a subject, should be conceded (Metzinger 2003). What may be illusory is the self's appearance as simple, unified, transparent and self-originating. The organization that produces this appearance, regulates the body, holds commitments and sometimes rewrites its own habits is as real as its effects, and the popular argument gathers evidence about the third object while ignoring the other three.

Constructedness is also the condition of a valuable kind of agency: a constructed self can take part in its own construction, and a self that was never constructed could not be reconstructed. Participation in one's own construction is, however, a late and fragile capacity that most self-changing systems lack, and it requires an explicit architecture. We build that architecture and measure what its threshold provides and costs.

A ladder of organizational capacities

The materials come from two literatures that rarely meet: the biology and cybernetics of self-producing systems, and the philosophy of practical identity. Connecting them invites a specific error, a continuous staircase from chemistry to freedom in which autocatalysis shades into autopoiesis, autopoiesis into agency, and agency into reflection, until a reaction network is credited with the capacities of a person. Each step in such a staircase crosses a real discontinuity. The appropriate connection is that each level establishes an organizational capacity the next presupposes, while each level's insufficiency for the next can be stated exactly.

The resulting ladder has six levels. Self-production: parts participate in regenerating the parts and processes that produce them. Closure: the production loop becomes an individuated organization that maintains its own conditions. Adaptivity: the organization regulates its behaviour with respect to its viability. Ultrastability: a second loop changes the first when the first fails, so the system modifies its own regulator. Anticipation: a model of possible futures becomes a present cause. Self-modelling: the system's own organization enters the model. Beyond these lies the threshold defined below, constitutive access, at which the system intervenes on its own organization because of what its model predicts the intervention will make of it.

Autocatalysis

An autocatalytic set is a reaction network in which every reaction is catalysed by a molecule the network produces, and the whole sustains itself on an external food supply. Kauffman showed that such networks arise with high probability once molecular diversity crosses a threshold (Kauffman 1986), and the RAF formalism of Hordijk and Steel made the notion exact and detectable within arbitrary reaction systems (Hordijk and Steel 2004). Eigen and Schuster's hypercycles had already coupled replicating cycles into higher-order cycles that stabilize one another (Eigen and Schuster 1979). This is the lowest level: matter participating in the production of its own conditions.

An autocatalytic set has no boundary of its own making, no enforced distinction between itself and its surroundings, no heredity to speak of, no evaluation of its states and no action. It is the

first condition under which the question of agency can arise, and it does not answer it. Gánti's chemoton makes the insufficiency constructive: his minimal living system requires three coupled subsystems, a metabolic cycle, a constraining template and an individuating boundary, none sufficient alone (Gánti 2003). Pattee added a further constraint: in living systems symbolic records and dynamic machinery depend on each other circularly, the sequence constraining the chemistry that reads and copies it, a semantic closure that a purely dynamical description does not capture (Pattee 1982). From the origin, then, systems that will eventually construct themselves contain a description of the system among their working parts.

Autopoiesis and closure

Autopoiesis names the step at which self-production becomes selfhood in its thinnest form. Maturana and Varela characterized the living unit as a network of production processes that produces its components, including its boundary, whose interactions regenerate the network (Maturana and Varela 1980). The tradition developed into the closure-of-constraints framework: an organization is autonomous when the constraints it contains, such as enzymes, membranes, vessels and regulatory structures, depend on one another for their maintenance and collectively channel flows of matter and energy into regenerating those constraints (Moreno and Mossio 2015; Montévil and Mossio 2015). A flame propagates and a vortex persists, but neither maintains a differentiated set of constraints that re-establishes its own conditions.

Rosen contributes here as the theorist of closure to efficient causation: the functions that sustain and repair the organization are themselves products of the organization (Rosen 2012). His stronger theses about the non-computability of organisms are contested, and nothing below depends on them; we use the relational point that an organism is a pattern of mutual entailment among functions before it is a list of molecules.

This level establishes identity as an ongoing achievement: an organization whose activity continually re-establishes the conditions under which it remains that organization. It does not yet distinguish among the system's states beyond persisting and ceasing, so maintenance has no better or worse.

Adaptivity

Di Paolo's correction to autopoiesis is the basis of minimal agency. Bare self-maintenance is all or nothing. An adaptive system monitors and regulates its states with respect to the conditions of its viability, distinguishing trajectories that approach collapse from those that recede from it and acting to shift the former toward the latter (Di Paolo 2005). With adaptivity, the system's states are no longer equivalent from its own point of view, and the organization generates norms against which its behaviour can succeed or fail. Barandiaran, Di Paolo and Rohde reduce minimal agency to three jointly necessary conditions: individuality defined by the system itself, interactional asymmetry by which the system is a source and modulator of its coupling with the world, and normativity by

which that modulation can go right or wrong relative to the system's continuation (Barandiaran et al. 2009).

A stone can be preserved or destroyed, and neither outcome is a success or failure for the stone; a bacterium swimming up a nutrient gradient is doing something that can go wrong for it. The difference is organizational, and it introduces the object tracked in the rest of the analysis: a system whose own future states matter to it.

Ultrastability

Ashby supplies a level that is often overlooked and that matters most for what follows. In his design a first-order system regulates ordinary variables, and when the essential variables, those on which survival depends, leave their bounds, a second-order mechanism changes the parameters of the first, repeatedly, until viable behaviour returns (Ashby 1960). He called the arrangement ultrastability. An ultrastable system modifies the machinery that generates its own behaviour, reorganizing its regulator and not only its conduct, and is in this respect the ancestor of self-reconstruction.

It is also blind. Ashby's reorganization is triggered by failure and proceeds by variation without a model of what the system is or what the reorganization will produce. Fatigue, trauma and conditioning all reorganize. A system can be fully self-modifying in the mechanical sense with no access to that fact, so being self-modifying and deliberately constructing oneself are different capacities. The companion theorem of Conant and Ashby, that every good regulator of a system must be a model of that system, indicates the next step within its assumptions (Conant and Ashby 1970): when the regulated system is the regulator's own organization, the theorem points toward self-modelling.

Anticipation

Rosen also supplies the temporal architecture. An anticipatory system contains a predictive model of itself and its environment and uses the model's predictions to control present behaviour (Rosen 2012), so a represented future acts in the present. Deliberate self-construction requires this structure, because a diet, an education, a vow or a prosthesis is chosen for an agent that does not yet exist, and only a model can make that agent causally available to the one choosing.

The free-energy tradition provides the most developed formal grammar of anticipatory regulation: organisms maintain themselves within a repertoire of viable states by minimizing surprise under a generative model, with action and perception as two parts of one inference and planning as inference over policies and counterfactual futures (Friston 2010; Friston et al. 2021). We use the grammar without its metaphysics. A Markov blanket, as Bruineberg and colleagues have shown, is a modeller's instrument before it is a boundary in the world, and treating every statistical partition as a self requires premises the mathematics does not supply (Bruineberg et al. 2022). The free-energy principle describes how a self-maintaining system with a model regulates its future, and

the closure tradition specifies what must exist for there to be a self-maintainer whose viability is at stake.

One consequence of the grammar constrains what follows. Under active inference the distinction between expecting and wanting nearly disappears, since the same machinery that predicts states also brings them about, and Kruglanski, Jasko and Friston conclude that all thinking is in this sense wishful (Kruglanski et al. 2020). Applied to the self-model, this means that a system's model of itself is part of the machinery whose futures it models and is biased toward futures the machinery already tends to produce. Self-knowledge used for self-construction is therefore motivated, and the measurements below price this.

Self-modelling

At the last level before the threshold, the system's own organization becomes an object in its model. Christensen and Hooker's self-directed agents anticipate their interactions, evaluate their performance against internal norms and modify their behaviour-generating processes accordingly, in the strongest cases constructing new goals (Christensen and Hooker 2001); this is the closest prior work. Bongard, Zykov and Lipson provide an engineering demonstration: a four-legged robot that continuously infers its body plan from sensorimotor data, uses the inferred model to generate locomotion, and after losing part of a limb revises the model and recovers a working gait (Bongard et al. 2006). Levin's computational boundary of a self generalizes the idea as a research hypothesis: agents at every scale are defined by the spatial and temporal reach of what they can measure, remember, predict and act upon, a reach that biology and technology can extend (Levin 2019).

A self-model alone does not reach the threshold. A system can model its dispositions and do nothing with the model, and it can learn, adapt and repair without ever selecting a change for what the change will make of it. Neither a self-model nor self-modification suffices; the threshold is their conjunction, defined next.

Constitutive access

Let an agent's organization comprise its ordinary states, the parameters that generate its policy, the rules by which those parameters change, and the external structures its activity relies on. Ordinary action changes the world. Learning changes parameters, as do fatigue, injury, propaganda and age. Constitutive access is a conjunction of four capacities. An agent has constitutive access to a variable of its own organization when it can discriminate or model the states of that variable; can identify interventions that would change it; can select an intervention partly because of its anticipated effect on the agent's own future organization and agency; and can monitor the transformation and revise its course. The set of such variables for a given agent at a given time is its constitutive access set.

Each condition excludes a case. Modelling excludes Ashby's reorganizer, which changes itself without representing itself. Identifying interventions excludes the merely self-aware, who can report their dispositions but cannot change them. Selection for anticipated self-effect excludes ordinary learning, which changes the learner as a byproduct of tracking the world. Monitoring ex-

cludes the one-off gamble. What remains is a specific causal architecture: one part of an organized process models and regulates another part, changing the organization from which later conduct proceeds, and the loop closes through a representation of the agent's own future.

No homunculus is required. The accessing part is a process within the system, itself constructed and itself a possible object of later access. Frankfurt's hierarchy of desires shows the structure in miniature: an agent who wants X and wants to stop wanting X has made a first-order state the object of a second-order attitude (Frankfurt 1971). Frankfurt also shows why endorsement alone is too weak, since an addiction can override every higher-order veto, and a hierarchy of attitudes without causal means of intervention does not govern. Korsgaard's self-constitution provides the complementary philosophical frame, agency as the ongoing constitution of practical identity through action (Korsgaard 2009), and constitutive access is the causal skeleton of such constitution when it becomes deliberate. Foucault documented that the Delphic *gnōthi seauton* was embedded in *epimeleia heautou*, the care of the self, a regime of exercises, examinations and disciplines in which self-knowledge served work on oneself (Foucault 1988), and Hadot read ancient philosophy as a whole as such a set of practices (Hadot 1995). Negarestani's claim that one cannot know oneself without constructing oneself restates that order in an engineering vocabulary (Negarestani 2014). Constitutive access specifies the architecture that the claim refers to.

Measuring the constructability horizon

We compute these notions in the smallest world in which each has a literal referent. An agent's organization consists of a policy of four binary context responses, a learning rate, an access flag and one optional external scaffold, giving 16 policies. The environment offers two coherent forms of life, the two policy peaks, separated by a valley: with a mismatch penalty of 1.25 per wrong response, 10 policies are viable and 6 are not, and the midpoint between the peaks nets -0.5 per period. An agent lives on an energy reserve of 6, replenished by its policy's fit and drained by its misfit. Six grades of agent differ only in their operators on themselves: C0 fixed, C1 learning, C2 self-modelling without intervention, C3 with model-guided single edits to its policy, C4 with additional access to its learning rule and access flag, and C5 with a scaffold that can hold a complete policy template and buffer a deficit. Each agent's constructability horizon is the set of organizations attainable from the well-adapted start through its own operators within 6 interventions, with every intermediate state viable. A second set, computed without the aiming requirement, contains the organizations that drift, whether experience or damage, can produce. All results are properties of the model and are not empirical estimates.

The ladder separates where the definition predicts. Below the threshold, the fixed, learning and self-modelling agents can each be changed by the world into 5 configurations and can bring about none of them; their self-directed horizon contains only their current organization. At C3 the horizon rises to 5 and coincides with drift, so everything that can happen to this agent it can also do deliberately. At C4 aimed reach exceeds drift by 30 to 5, and at C5 by 120 to 5. The discontinuity at C3 follows from the definition; the model adds what the capacity costs and provides.

The first result is a barrier. The C3 agent, editing one response at a time and required to live in each intermediate state, cannot reach the other peak, because every path crosses the valley. Five of the 10 viable configurations, including an entire alternative form of life, are unreachable alive. The C5 agent reaches them because its scaffold holds a complete template of the other form of life, so the crossing is a single adoption, and its buffer absorbs the one costly step. The organizations reachable by increments alone are therefore fewer than the viable ones, and the rest are reached only by adopting a form held available from outside.

The second result concerns traps. When the agent gains access to its learning rule and access flag, its reach increases sharply, and about half of the added reach, 0.52 of destinations at C4 and 0.5 at C5, consists of organizations from which no further self-directed change is possible, because learning has been frozen or access removed. Reach grows by a factor of 6 from C3 to C4, and reach that preserves the capacity for further revision grows by a factor of 3. Raw horizon size therefore overstates freedom, and the more of itself an agent can alter, the larger the share of its possible futures that are terminal. Self-binding is sometimes rational, as when Ulysses had himself tied to the mast, so the horizon requires a profile and not a single number.

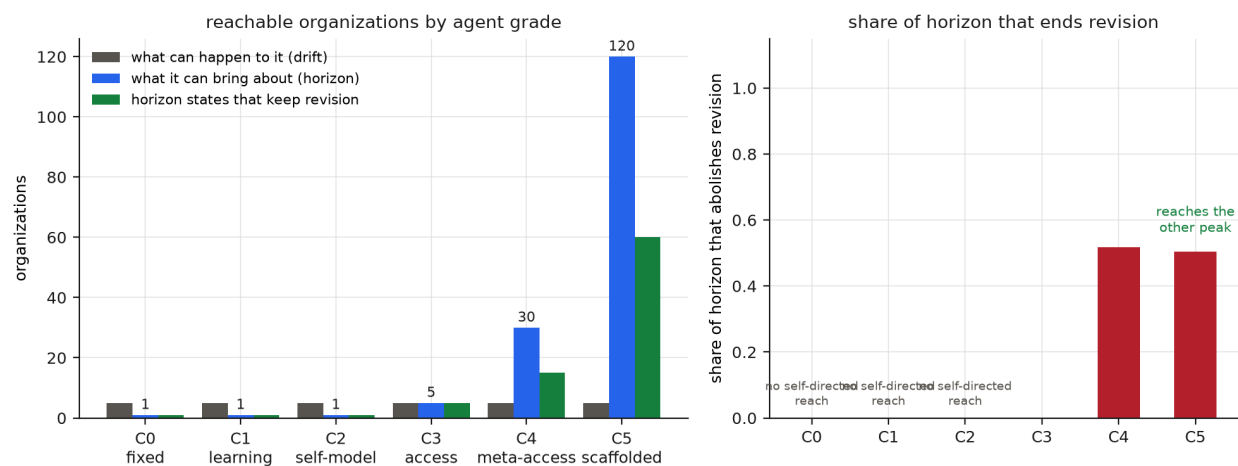


Figure 1: The constructability horizon by agent grade. Left: organizations reachable by drift, by the agent's own aimed intervention, and by aimed intervention that preserves further revision. Below the access threshold aimed reach is zero while drift is 5; above it aimed reach exceeds drift and traps appear. Right: the share of each horizon that ends further revision.

Repair after niche collapse

In the next experiment one niche collapses, the agent's form of life stops paying and its reserve begins to drain. We compare Ashby's blind loop with deliberate reconstruction.

Blind ultrastability, repeated random whole reorganizations until a viable organization results, survives the collapse in 0.78 of 10,000 trials. The incremental self-editor, moving toward the surviving peak one response at a time and spending two periods in each intermediate state, never survives, because the valley exhausts its reserve before the third edit pays. Whole-template reconstruction

survives at a rate that increases with the accuracy of its model of the changed world, 0.36 at accuracy 0.2, 0.68 at 0.6 and 1.0 at 1.0, because a miscalibrated reconstructor interprets the collapse as temporary and recommits to the failed form of life. On the accuracy grid examined, deliberate reconstruction falls below the blind loop at 0.2, 0.4 and 0.6 and exceeds it at 0.8 and 1.0; linear interpolation, which fits the nearly linear survival curve, places the crossing at an accuracy of about 0.72. Below that level the blind repair mechanism is the better option.

An architecture that retains the blind loop beneath the deliberate one performs better than deliberation alone: it reconstructs by model first and, if the model's choice fails, returns to random reorganization. This fallback exceeds pure guidance at every accuracy, 0.58 against 0.36 at the lowest, and exceeds the blind loop above an interpolated accuracy of about 0.57; below that it still loses, because consulting a poor model consumes reserve that the blind loop would have spent on trials. Sequencing does not compensate for poor calibration. For a self-constructing system, whether a person in crisis or an institution in reform, the implication is to retain variation-and-selection machinery beneath deliberate machinery, while recognizing that doing so does not make an inaccurate self-model safe.

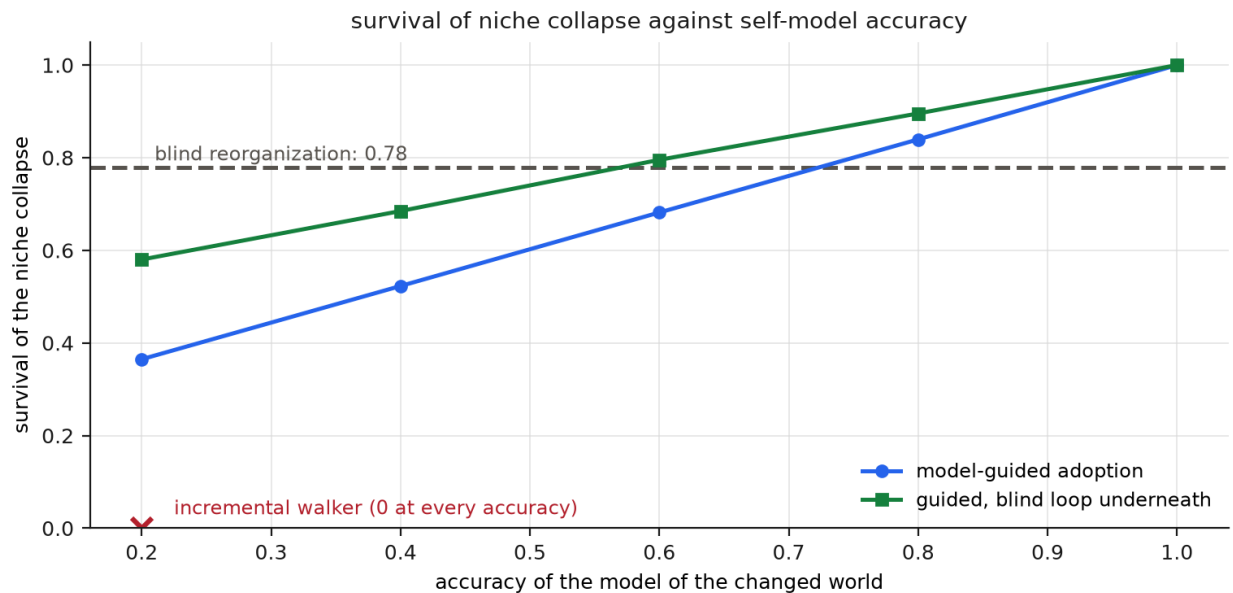


Figure 2: Survival of the niche collapse by repair architecture. The blind loop is flat at 0.78; guided whole-template reconstruction rises with the accuracy of the model of the changed world and exceeds the blind loop between accuracies 0.6 and 0.8; the fallback exceeds guidance at every accuracy; the incremental walker never survives.

Scaffolding

The scaffold result extends well beyond the model. Most of what expands a human constructability horizon was never inside the agent: language that supplies the categories under which internal states become discriminable, teachers who install capacities the earlier self could not have spec-

ified, writing that keeps intentions stable over the years needed to change, and institutions that sustain commitments when individuals would abandon them. Clark and Chalmers argued that external structure can be part of cognition (Clark and Chalmers 1998); Sterelny's account of cognition and agency as scaffolded, built within environments engineered by others to make certain developments cheap, fits self-construction better (Sterelny 2010). The literature on relational autonomy drew the political consequence earlier: the capacity for self-governance is socially produced, sustained by care, recognition and material provision, and destroyed when they are withdrawn (Mackenzie and Stoljar 2000).

In the model, the difference between an agent that can reach the other form of life and one that cannot is a single bit, whether a scaffold exists. Outside the model that bit corresponds to a school, therapy, an apprenticeship, a union, a text or a refuge. If freedom is the profiled horizon described above, a society distributes freedom when it distributes scaffolds, and that distribution is as unequal as others. Some people receive time, education, treatment, legal recognition, second chances and access to alternative forms of life; others receive short horizons, debt and a single form of life whose collapse is their own.

External reconstruction and capture

Malabou's work and a book from 1969 prevent this account from becoming a programme of self-optimization. Malabou distinguishes flexibility, the capacity to receive imposed form and comply, from plasticity, which gives form, receives form and can break form; a culture that promotes flexibility under the name of plasticity trains its members for compliant reconstruction by others (Malabou 2008). Her ontology of the accident adds that injury, disease and catastrophe also construct, producing persons discontinuous with those they were, without consent or recoverable meaning (Malabou 2012). Construction is morally neutral: scars, addictions and fanatics are all constructed, and the capacity to be remade does not ensure that the remaking is one's own.

Delgado's *Physical Control of the Mind*, published during his stimociever experiments, argued sincerely that electrical intervention in the brain should found a "psychocivilized" society in which persons are reconstructed from outside as a public good (Delgado 1969). It shows that the technical capacity analysed here arrived decades ago, first in the hands of confident external operators, and it poses the question any account of self-construction must answer: once the causes that constitute a person become manipulable, who controls the intervention.

The model makes the question measurable. An external controller reconstructs the collapsed agent as effectively as the agent reconstructs itself, in one step in both cases, so efficacy does not distinguish them. The capturing controller also removes the agent's access flag and freezes its learning rule so that the reconstruction persists. Nothing changes until the environment changes again. At the second shock the self-directed agent crosses again, an agent reconstructed by a benevolent controller that left its capacities intact crosses identically, and the captured agent, unable to edit, learn or adopt, dies. The benevolent arm survives everything the self-directed arm survives, so domination lies in the removal of revisability and not in external authorship, and it is invisible while

conditions are stable. The diagnostic for capture is therefore temporal: a liberation and a capture can produce equally good lives until the environment shifts.

survival of first and second shock by reconstruction

self-directed	yes	yes	yes
external, revision kept	yes	yes	yes
external, revision pruned	yes	no	no
	shock 1 recovered	revision loop after	shock 2 survived

Figure 3: Three reconstructions through two shocks. All three recover from the first collapse in one step; they differ in whether the revision loop survives the reconstruction, which appears only at the second shock.

Aspiration and the revert rule

Constitutive access selects interventions by their anticipated effect under the agent's current model and evaluator, but the most important transformations change the evaluator. Callard's aspirant is trying to come to value music, justice or a child they do not yet have, and cannot yet perceive the value the transformed self will perceive; the choice is made toward reasons the chooser does not yet possess (Callard 2018). Paul's transformative experiences generalize the structure: some choices cannot be simulated in advance because undergoing them replaces the self that would do the simulating (Paul 2014). A theory of self-construction that required full anticipation would exclude the reconstructions that make up maturation.

The model states the problem and solves a tractable version of it. The agent's present evaluator scores the other form of life at -5 , so it will never choose the crossing on its merits. When its own niche decays by a factor of 0.9 per period, with an upkeep of 1 that the decaying niche eventually fails to cover, the myopic agent remains in its form of life for 26 periods and dies, having declined the crossing throughout. The aspirant instead makes a provisional commitment: it adopts the other form of life on trial at the first period in which its current life costs more than it returns, subject to a revert rule. In the model the provisional agent crosses at period 14 and survives. The revert rule matters. If it is ratification, keeping the transformation if the new evaluator approves it, the protocol is defenceless against a capture that installs a self-ratifying evaluator, such as a group that teaches members that joining was their best decision. In the model, an offer consisting

of an evaluator that approves everything together with a removed access flag is kept under ratification and refused only under a tripwire rule, which keeps a transformation if viability and the revision loop both survive it and ignores approval. Aspiration is therefore rational as provisional self-transformation whose revert conditions protect the capacity to revise and not the approval of the resulting self: an agent may become someone it cannot now imagine valuing, but not someone who cannot leave.

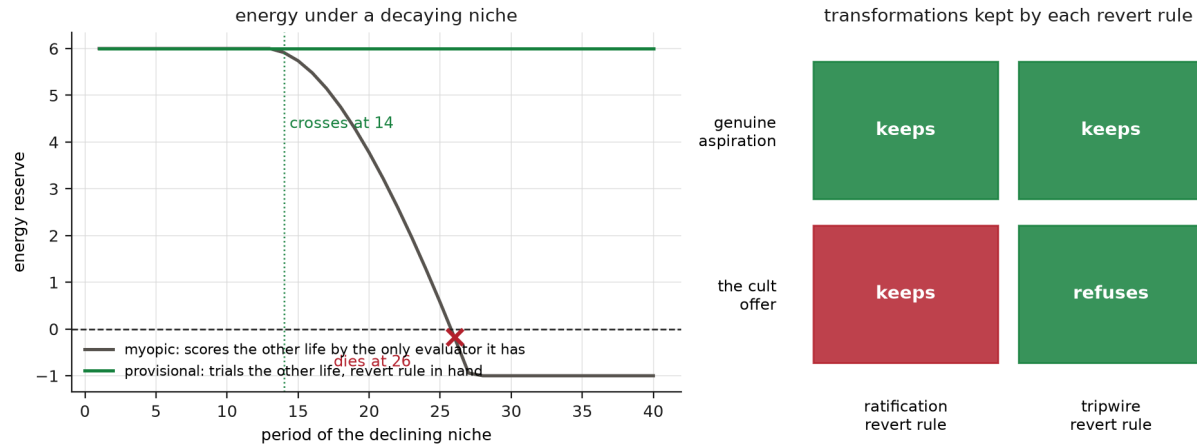


Figure 4: Aspiration under a decaying form of life and the revert rule. Left: the myopic agent declines the crossing every period and dies at 26; the provisional agent tries it at 14 and survives. Right: ratification keeps both the genuine aspiration and the self-ratifying capture; the tripwire keeps the first and refuses the second.

Conclusion

The self is a constructed, real and non-fundamental organization: a body that keeps variables within bounds, a regulatory architecture, a self-model transparent to itself, and a practical identity held in commitments and recognition. Its construction proceeds through a discontinuous ladder of self-production, closure, adaptivity, ultrastability, anticipation and self-modelling, and crosses a threshold when constitutive variables of the organization become targets of intervention within its model. On this account freedom is neither exemption from causation nor a property owned outright; it is the profiled capacity decomposed above into the reach of the constructability horizon, the share of that reach that preserves revision, the calibration of the self-model that steers it, the scaffolds that bridge its valleys, and the absence of an external agent that removes the revision loop. Empowerment, defined as the channel capacity of an agent's influence over its own future states (Klyubin et al. 2005), measures one dimension of this profile, and a captured agent can be highly empowered in every dimension except revisability. Viability theory, the source of the model's viability kernel (Aubin 2009), supplies the standing constraint that any reconstruction must keep the reconstructor alive.

The determinist objection does not apply. Every self-intervention has causes and the loop that selects it has a history, but a control loop with a history still controls. A person who has acquired

discriminating, calibrated, scaffolded and revision-preserving access to some causes of their future conduct differs from a person who has not, whatever physics is true of both. The agent does not become its own cause; it becomes a cause acting on itself within the history that produced it. The relevant questions concern how much of the causal chain is open to intervention, who built the means of intervention, and who is permitted to use them.

The Delphic imperative is thereby completed. The practices around the inscription at Delphi treated self-knowledge as serving work on oneself. For a caused, constructed organism partly opaque to itself, constructing oneself means modelling what one can of one's organization, intervening where access exists, keeping blind repair beneath deliberate repair, using scaffolds while recognizing them as borrowed, and refusing transformations that would remove one's capacity for further revision. The same property that makes a self constructible makes it capturable.

Reproducibility

The simulation, its four figures and every model number quoted above regenerate from one command in the repository (`simulation/run_all.py`). The horizon, drift, trap and manipulation results are exact enumerations and deterministic; the repair experiments run 10,000 trials on a recorded seed. Nineteen invariants fail the run if broken, among them the lethality of the valley, the coincidence of drift and aim at the access threshold, the appearance of traps only with meta-level access, the incremental walker's death, the calibration crossover and its interpolated position within the grid bracket, the fallback's advantage over pure guidance together with its cost against pure blindness, the separation of the three reconstruction arms at the second shock, and the tripwire's refusal of the self-ratifying capture.

References

- Ashby, W. Ross. 1960. *Design for a Brain: The Origin of Adaptive Behaviour*. 2nd ed. Chapman and Hall. <https://doi.org/10.1037/11592-000>.
- Aubin, Jean-Pierre. 2009. *Viability Theory*. Birkhäuser. <https://doi.org/10.1007/978-0-8176-4910-4>.
- Barandiaran, Xabier E., Ezequiel Di Paolo, and Marieke Rohde. 2009. "Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action." *Adaptive Behavior* 17 (5): 367–86. <https://doi.org/10.1177/1059712309343819>.
- Bongard, Josh, Victor Zykov, and Hod Lipson. 2006. "Resilient Machines Through Continuous Self-Modeling." *Science* 314 (5802): 1118–21. <https://doi.org/10.1126/science.1133687>.
- Bruineberg, Jelle, Krzysztof Dołęga, Joe Dewhurst, and Manuel Baltieri. 2022. "The Emperor's New Markov Blankets." *Behavioral and Brain Sciences* 45: e183. <https://doi.org/10.1017/s0140525x21002351>.

- Callard, Agnes. 2018. *Aspiration: The Agency of Becoming*. Oxford University Press. <https://doi.org/10.1093/oso/9780190639488.001.0001>.
- Christensen, W. D., and C. A. Hooker. 2001. "Self-Directed Agents." *Canadian Journal of Philosophy Supplementary Volume* 27: 18–52. <https://doi.org/10.1080/00455091.2001.10715995>.
- Clark, A., and D. Chalmers. 1998. "The Extended Mind." *Analysis* 58 (1): 7–19. <https://doi.org/10.1093/analys/58.1.7>.
- Conant, Roger C., and W. Ross Ashby. 1970. "Every Good Regulator of a System Must Be a Model of That System." *International Journal of Systems Science* 1 (2): 89–97. <https://doi.org/10.1080/00207727008920220>.
- Delgado, José M. R. 1969. *Physical Control of the Mind: Toward a Psychocivilized Society*. Harper and Row.
- Dennett, Daniel C. 1991. "Real Patterns." *The Journal of Philosophy* 88 (1): 27–51. <https://doi.org/10.2307/2027085>.
- Di Paolo, Ezequiel A. 2005. "Autopoiesis, Adaptivity, Teleology, Agency." *Phenomenology and the Cognitive Sciences* 4 (4): 429–52. <https://doi.org/10.1007/s11097-005-9002-y>.
- Eigen, Manfred, and Peter Schuster. 1979. *The Hypercycle: A Principle of Natural Self-Organization*. Springer. <https://doi.org/10.1007/978-3-642-67247-7>.
- Foucault, Michel. 1988. "Technologies of the Self." In *Technologies of the Self: A Seminar with Michel Foucault*, edited by Luther H. Martin, Huck Gutman, and Patrick H. Hutton. University of Massachusetts Press.
- Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68 (1): 5–20. <https://doi.org/10.2307/2024717>.
- Friston, Karl. 2010. "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience* 11 (2): 127–38. <https://doi.org/10.1038/nrn2787>.
- Friston, Karl, Lancelot Da Costa, Danijar Hafner, Casper Hesp, and Thomas Parr. 2021. "Sophisticated Inference." *Neural Computation* 33 (3): 713–63. https://doi.org/10.1162/neco_a_01351.
- Gánti, Tibor. 2003. *The Principles of Life*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198507260.001.0001>.
- Hadot, Pierre. 1995. *Philosophy as a Way of Life: Spiritual Exercises from Socrates to Foucault*. Edited by Arnold I. Davidson. Translated by Michael Chase. Blackwell.
- Hordijk, Wim, and Mike Steel. 2004. "Detecting Autocatalytic, Self-Sustaining Sets in Chemical Reaction Systems." *Journal of Theoretical Biology* 227 (4): 451–61. <https://doi.org/10.1016/j.jtbi.2003.11.020>.

- Kauffman, Stuart A. 1986. "Autocatalytic Sets of Proteins." *Journal of Theoretical Biology* 119 (1): 1–24. [https://doi.org/10.1016/s0022-5193\(86\)80047-9](https://doi.org/10.1016/s0022-5193(86)80047-9).
- Klyubin, A.S., D. Polani, and C.L. Nehaniv. 2005. "Empowerment: A Universal Agent-Centric Measure of Control." *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, 128–35. <https://doi.org/10.1109/cec.2005.1554676>.
- Korsgaard, Christine M. 2009. *Self-Constitution: Agency, Identity, and Integrity*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199552795.001.0001>.
- Kruglanski, Arie W., Katarzyna Jasko, and Karl Friston. 2020. "All Thinking Is 'Wishful' Thinking." *Trends in Cognitive Sciences* 24 (6): 413–24. <https://doi.org/10.1016/j.tics.2020.03.004>.
- Levin, Michael. 2019. "The Computational Boundary of a "Self": Developmental Bioelectricity Drives Multicellularity and Scale-Free Cognition." *Frontiers in Psychology* 10: 2688. <https://doi.org/10.3389/fpsyg.2019.02688>.
- Mackenzie, Catriona, and Natalie Stoljar, eds. 2000. *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*. Oxford University Press. <https://doi.org/10.1093/oso/9780195123333.001.0001>.
- Malabou, Catherine. 2008. *What Should We Do with Our Brain?* Translated by Sebastian Rand. Fordham University Press.
- Malabou, Catherine. 2012. *Ontology of the Accident: An Essay on Destructive Plasticity*. Translated by Carolyn Shread. Polity.
- Maturana, Humberto R., and Francisco J. Varela. 1980. *Autopoiesis and Cognition: The Realization of the Living*. Boston Studies in the Philosophy of Science 42. D. Reidel. <https://doi.org/10.1007/978-94-009-8947-4>.
- Metzinger, Thomas. 2003. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press. <https://doi.org/10.7551/mitpress/1551.001.0001>.
- Montévil, Maël, and Matteo Mossio. 2015. "Biological Organisation as Closure of Constraints." *Journal of Theoretical Biology* 372: 179–91. <https://doi.org/10.1016/j.jtbi.2015.02.029>.
- Moreno, Alvaro, and Matteo Mossio. 2015. *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Springer. <https://doi.org/10.1007/978-94-017-9837-2>.
- Negarestani, Reza. 2014. "Frontiers of Manipulation." Lecture text. Primer (Dead Reckoning). <https://primer.dk/Projects/Dead-Reckoning/Reza-Negarestani-Frontiers-Of-Manipulation>.
- Pattee, Howard H. 1982. "Cell Psychology: An Evolutionary Approach to the Symbol-Matter Problem." *Cognition and Brain Theory* 5 (4): 325–41.
- Paul, L. A. 2014. *Transformative Experience*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198717959.001.0001>.

Rosen, Robert. 2012. *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*. 2nd ed. Springer. <https://doi.org/10.1007/978-1-4614-1269-4>.

Sterelny, Kim. 2010. "Minds: Extended or Scaffolded?" *Phenomenology and the Cognitive Sciences* 9 (4): 465–81. <https://doi.org/10.1007/s11097-010-9174-y>.