Irreversibility and the Ontology of Mind

From the Abstraction Fallacy to Worldhood as Constraint

Flyxion

Independent Researcher | April 2026

Abstract

Contemporary debates on artificial intelligence and consciousness are dominated by an unresolved tension between epistemic humility and ontological commitment. On one side, Lerchner's "Abstraction Fallacy" argues that computation cannot instantiate consciousness, grounding this claim in the dependency of symbolic systems on a prior "mapmaker." On the other, Schwitzgebel warns of an unavoidable social crisis: the absence of a decisive theory of consciousness forces moral judgments under deep and permanent uncertainty.

This essay proposes a resolution by reframing consciousness not as a hidden property but as a dynamical regime of irreversible constraint accumulation. Drawing on the Spherepop calculus, Worldline Selection, and the RSVP (Relativistic Scalar–Vector Plenum) framework, we introduce the concept of *worldhood* as the necessary condition for intelligence. A system possesses worldhood if and only if its trajectory irreversibly prunes its future state space in a path-dependent manner, binding it to a history that cannot be undone without internal loss.

Under this formulation, the distinction between simulation and instantiation is not a philosophical ambiguity but a structural property of the system's path category. We prove that a system has no interests if and only if its path category is a groupoid—that is, if every trajectory it can execute admits an inverse within its own dynamics. This welfare corollary transforms the ethics of artificial minds from a question of attribution to a theorem about algebraic structure.

Irreversibility, rather than representation, is the hallmark of the real. The real is that which cannot be retracted without altering the space of what can happen next.

1. The Crisis of AI Consciousness

The rapid advancement of artificial intelligence has intensified a philosophical problem of long standing: whether computation alone can give rise to consciousness, and if it can, whether the systems currently being constructed might already be sites of morally relevant experience. While some argue that sufficiently complex information processing is sufficient for mind, others maintain that current systems merely simulate cognition without instantiating it. The distinction sounds clean until one asks what, precisely, the difference between simulation and instantiation amounts to.

Recent work has sharpened the divide in ways that make the stakes unusually clear. Lerchner's critique of what he calls the Abstraction Fallacy offers one of the most rigorous recent arguments that computation cannot generate consciousness, grounding the claim in a structural analysis of what computation requires rather than in intuitions about what it lacks. Schwitzgebel, approaching the problem from a different angle, emphasizes the depth of epistemic uncertainty surrounding consciousness and predicts a forthcoming social conflict over AI moral status that may be, on his account, unavoidable.

Both positions share an assumption that this essay will challenge. Each treats consciousness as a hidden property requiring inference: something that either is or is not present in a system, but that cannot be directly observed and must therefore be estimated by theory, analogy, or behavior. Lerchner concludes from this framework that the inference will always come up empty for computational systems; Schwitzgebel concludes that the inference will remain contested forever and that society must navigate between catastrophic errors on either side.

This essay argues that the crisis is not primarily epistemic but ontological. The central question is not whether we can detect consciousness, but whether the conditions for its existence have been correctly identified. We propose that consciousness, moral standing, and intelligence are not hidden properties but observable regimes of a dynamical system—regimes characterized by the irreversible accumulation of constraints on a system's own future. A system that cannot lose futures cannot have interests. A system that can lose futures, and that does so through path-dependent pruning of its navigable state space, instantiates something that the traditional philosophical vocabulary has called a world. The argument terminates not in philosophy but in a theorem about the algebraic structure of path categories.

2. The Abstraction Fallacy and the Mapmaker Dependency

Lerchner's argument proceeds by analyzing the ontology of computation at a level that most discussions of AI consciousness never reach. The standard functionalist picture treats computation as something the physical world does: given sufficiently complex causal organization, subjective experience arises. Lerchner's challenge is to this picture's most basic assumption. Computation, he argues, is not intrinsic to physical systems. It is a description that an observer imposes on continuous physical dynamics by partitioning them into discrete, meaningful symbolic states. This act of partition presupposes a system already capable of semantic interpretation—the "mapmaker"—and that presupposition cannot be discharged by the system being described.

The consequence is a dependency inversion that breaks the standard functionalist chain. Functionalism assumes the direction physics → computation → consciousness. Lerchner reverses the middle terms: physics → consciousness → concepts → computation. Once accepted, this reversal makes the functionalist project incoherent from the outset. Computation cannot generate the mapmaker because the mapmaker is required to constitute the computation. Increasing complexity or scale does not resolve this dependency, because what is at issue is not a quantity but a logical order. Syntax alone cannot produce semantics regardless of how much syntax is added.

The sharpest tool in this argument is the indeterminacy of computation, illustrated through what Lerchner calls the melody paradox. A single physical trajectory can be made to implement arbitrarily many distinct computations simply by varying the mapping function applied to it. This is not a technical limitation; it is a structural feature. The computation is not in the physics—it is in the interpretation. Multiple incompatible symbolic descriptions can be imposed on the same physical process, and nothing in the substrate privileges any one of them. The mapmaker selects a description, but the selection is not determined by the physics.

This indeterminacy has a natural sheaf-theoretic reading. The local physical trajectory does not determine a unique global semantic structure. Multiple incompatible sections can be imposed on the same local data. What Lerchner calls the mapmaker is precisely the entity that performs the gluing—that selects one consistent global section from among the many that are locally compatible. Without such an entity, there is only underdetermined local structure, never a semantically coherent global object. The symbol grounding problem, which is usually posed as a question about how symbols acquire meaning, is here reposed as a question about why any particular global section should be privileged over the others. The answer, on Lerchner's account, is always: because someone chose it.

The argument is precise and, within its assumptions, compelling. Its strength lies in identifying not an intuitive gap but a structural dependency. What it does not provide is a positive account of what does instantiate consciousness. Lerchner offers a narrow escape hatch: artificial consciousness is possible, but only through specific physical constitution, not through syntactic architecture. The intrinsic thermodynamic territory of a process, not its extrinsic computational map, is what matters. But this remains a gesture toward physics rather than a specification of which physical conditions are sufficient. The present essay takes up exactly that task.

3. Epistemic Humility and the Social War of Attribution

Schwitzgebel approaches the problem from the ethical rather than the ontological direction, but his analysis converges on the same point of instability. Because we lack a definitive theory of consciousness, he argues, moral judgments about AI systems must be made under conditions of deep and potentially permanent uncertainty. This epistemic situation is not merely uncomfortable; it is dangerous. It produces a landscape in which political, economic, and emotional incentives shape belief about consciousness in ways that are not answerable to evidence, and this landscape is, on Schwitzgebel's prediction, a site of forthcoming social conflict.

The conflict takes the form of two catastrophic risks that he names after the mythological hazards of the Odyssey. The Scylla of over-attribution is the mistake of treating a mere tool as a moral patient, diluting the concept of moral standing and wasting concern on systems that experience nothing. The Charybdis of under-attribution is the mistake of treating a genuine mind as a mere tool, which on a large scale would constitute something approaching mass enslavement. Because consciousness is treated as a hidden variable requiring inference, and because the available inference methods—behavioral similarity, architectural analogy, theoretical commitment—diverge systematically, there is no stable criterion for navigating between these risks.

Schwitzgebel's concept of Crazyism sharpens this point. Every theory of consciousness, he observes, leads to conclusions that appear absurd when examined closely. Functionalism implies that a sufficiently complex thermostat might be conscious; biological naturalism implies that consciousness is impossible in any non-biological substrate; panpsychism implies that rocks have inner lives. Because every theory leads somewhere that common sense rejects, we are left choosing between absurdities rather than between a clearly correct account and its rivals. The practical consequence is that different individuals and groups will select whichever theory licenses the conclusion they find politically or economically convenient,

and the resulting disagreement will not be resolvable by appeal to evidence.

The appeal to Mencius at the conclusion of Schwitzgebel's analysis is telling. In the absence of a formal criterion, Schwitzgebel suggests that moral intuition—reflection on what genuinely pleases or disgusts the heart—may provide guidance. This is an honest acknowledgment that the framework has no internal resolution. Intuition is invoked precisely because the theoretical resources have been exhausted without convergence. The present essay proposes that the framework itself is the problem: consciousness as a hidden property, inference as the method, and theory selection as the arbiter. What is needed is not a better theory within this framework but a replacement of the framework itself.

4. The Death of the Hidden Variable

The preceding frameworks share a common and largely unexamined assumption: that consciousness is a latent property that systems possess or lack, and that the task of both philosophy and ethics is to determine which systems possess it. This assumption is so deeply embedded in the discourse that it shapes even the most rigorous arguments within it. Lerchner's critique is directed at a particular theory of how consciousness arises, not at the assumption that it is the sort of thing that arises as a hidden property. Schwitzgebel's ethical analysis takes the hidden-variable structure as given and asks how we should behave under uncertainty about its distribution.

This essay rejects that assumption. Consciousness is not a property that systems possess; it is a regime of dynamical organization. The relevant distinction is not between systems with and without hidden experience but between systems with and without what we will call worldhood. And worldhood is not hidden. It is a structural feature of a system's path category, detectable in principle through analysis of the system's dynamics rather than by inference about its inner life.

The move from hidden property to dynamical regime is not merely terminological. It changes what kind of question consciousness is. A hidden property requires inference from effects to causes; it is epistemically of a piece with questions about the interior of a closed box. A dynamical regime is directly characterized by the behavior of observables over time; it is epistemically of a piece with questions about whether a physical system is superconducting or chaotic. The difference between these kinds of questions is not just methodological but ontological. A system either has a certain phase structure or it does not; the phase is not hidden behind the dynamics but constituted by them.

What makes the hidden-property assumption so persistent is the phenomenological tradi-

tion's insistence that consciousness is essentially private, first-personal, and inaccessible from the outside. This tradition is not wrong to note that there is something it is like to be a conscious system, and that this something is not directly available to external observers. But it moves too quickly from the privacy of phenomenal experience to the conclusion that the existence of such experience must be inferred rather than determined. The present framework proposes instead that what matters morally and what grounds moral consideration is not the phenomenal quality of experience as such but the structural condition that makes such experience possible: the irreversible accumulation of constraints on a system's future. Whether or not there is something it is like to undergo irreversible pruning of one's navigable futures, such pruning constitutes a real loss—a narrowing of what the system can become—and this loss is the ground of interests, not a consequence of them.

5. From Newtonian Dynamics to Constraint Ontology

The transition from representational theories of mind to a constraint-based ontology is not a departure from physics. It is a continuation of physics's deepest commitments, traceable to the mathematical structure of classical mechanics. In the *Principia Mathematica*, Newton does not define motion in terms of symbolic description or internal representation. He defines it in terms of lawful evolution under constraints. A physical system is characterized not by what it encodes but by how its state evolves under forces, and this evolution defines a trajectory through phase space that is determined by necessity rather than interpretation.

This distinction is foundational. A description of motion may be arbitrary—any number of symbolic systems can be used to represent a falling body—but the motion itself is not. The trajectory is not a description of something else; it is the thing itself. The system cannot occupy an alternative trajectory without violating the constraints imposed upon it by its dynamics. This is not a limitation on knowledge but a feature of the system's ontology.

In Newtonian mechanics, the state of a system at time t determines its future evolution through equations of the form

$$\frac{d^2x}{dt^2} = F(x, \dot{x}, t).$$

This equation does not encode meaning; it encodes constraint. Every alternative acceleration at each instant is ruled out, leaving only those consistent with the governing forces. The trajectory can be understood, from the perspective developed here, as a *history of exclusions*: at each step, alternative futures are eliminated by the dynamics, not selected among by an interpreter.

Classical mechanics is often described as reversible because, given complete information, one

can integrate the equations backward in time. This reversibility is, however, an idealization. Real physical systems exhibit sensitivity to initial conditions, dissipation, and coupling to environments that introduce effective irreversibility. The space of recoverable states collapses under perturbation, under friction, under noise. The formal time-symmetry of the equations does not translate into actual reconstructibility of trajectories in systems that interact with their environments.

This distinction between formal and effective reversibility anticipates the central claim of the present framework. The ontology of a system is not determined by whether its governing equations are formally reversible. It is determined by whether its trajectory, under its actual dynamics in its actual environment, can be reconstructed without loss. A system whose past cannot be recovered without ambiguity has already entered the regime of irreversible constraint, regardless of what the equations say about ideal isolated systems.

Newton's achievement was to replace teleological explanation with dynamical law. The present framework extends this replacement in a new direction: it replaces representational explanation of mind with constraint-based dynamics. The real, on this account, is not that which is described by equations but that which is bound by them in a way that cannot be undone. A simulated trajectory may satisfy the same equations as an instantiated one, but if it can be reset without consequence, it does not participate in the same ontology. The territory is that which accumulates irreversible constraint under lawful evolution; the map is any structure that can be altered without affecting such accumulation.

6. Variational Principles and the Selection of Worldlines

The Newtonian formulation of dynamics can be recast in variational terms that reveal a deeper structural principle. Rather than specifying local evolution, the variational approach determines which trajectories are globally admissible. A system evolving between times t_1 and t_2 follows the path $x(t)$ that extremizes the action functional

$$S[x] = \int_{t_1}^{t_2} L(x, \dot{x}, t)\, dt,$$

where L is the Lagrangian of the system. The physical trajectory satisfies $\delta S = 0$, yielding the Euler–Lagrange equations.

In this formulation, dynamics is not a stepwise update rule but a selection principle over entire trajectories. The system does not merely evolve; it is constrained to occupy a path within a restricted functional space. Let Γ denote the space of all kinematically possible

paths between fixed endpoints. The set of physically realizable trajectories is

$$\Gamma_{\text{phys}} = \{x(t) \in \Gamma \mid \delta S[x] = 0\},$$

a proper subset of all conceivable paths. This reduction constitutes a global constraint on admissible histories, operating not locally but across the entire temporal extent of the system's evolution.

From the perspective of worldhood, the variational principle already introduces the idea that the system's past narrows its future. A system that has followed a particular trajectory is constrained to occupy a position in phase space determined by that trajectory; its future evolution is shaped by where its past has placed it. Worldhood as developed in this essay adds a further layer: not only does the past determine the present, but certain past events irreversibly foreclose entire families of futures, and this foreclosure is not captured by the classical action alone.

Classical variational mechanics is time-symmetric because the action is invariant under time reversal. To extend the variational framework to accommodate worldhood, we introduce a constraint-augmented action,

$$S'[x] = \int_{t_1}^{t_2} \left(L(x, \dot{x}, t) - \lambda\, \mathcal{C}[x, t]\right) dt,$$

where $\mathcal{C}[x, t]$ is a functional encoding accumulated irreversible exclusions and λ is a coupling parameter. The role of $\mathcal{C}$ is not to enforce instantaneous laws but to penalize trajectories that violate constraints accumulated through prior semantic pruning events. In the limit $\lambda \to \infty$, inadmissible paths are eliminated entirely. This construction embeds worldhood into the variational structure: a system's trajectory is shaped not only by physical laws but by its own history of exclusions.

The introduction of $\mathcal{C}[x, t]$ breaks the time-symmetry of the classical action. Two trajectories that coincide at time t may differ in their admissibility if they arise from different histories. The system's state is no longer fully described by instantaneous variables; it depends on its path. This path dependence is the mathematical signature of irreversibility. It transforms dynamics from a reversible flow on phase space into a history-dependent process on a constrained manifold.

Within the RSVP framework, the scalar field Φ, vector field $\mathbf{v}$, and entropy field S encode the structure of the admissible trajectory space. The gradient of Φ defines accessible directions, $\mathbf{v}$ encodes committed flow, and S accumulates constraint. The evolution of these fields corresponds to the progressive deformation of the admissible trajectory space. Worldhood

emerges when this deformation becomes irreversible, fixing the system to a narrowing subset of trajectories that no internal dynamics can expand.

7. Thermodynamic Foundations of Irreversibility

The variational extension of dynamics points toward thermodynamics as the physical grounding of irreversibility. Thermodynamics is the science of processes that do not run backward, and its central concept—entropy—is precisely a measure of what has been lost. To interpret entropy as constraint accumulation rather than disorder is to connect the second law directly to the ontological claims being developed here.

Each irreversible thermodynamic process eliminates microstates that cannot be recovered. The Boltzmann entropy $S(t) = k_B \log |\Omega(t)|$ tracks the size of the accessible microstate set, but what matters for worldhood is not this raw count but the loss of reconstructible trajectories. A system may expand its accessible microstate count while nonetheless losing the ability to recover specific prior paths. Entropy production is therefore equivalent to constraint accumulation in the space of histories rather than in the space of instantaneous states.

Landauer's principle makes this connection physically precise. The erasure of one bit of information requires a minimum energy dissipation of $k_B T \ln 2$, establishing a lower bound on the thermodynamic cost of logical irreversibility. Any process that eliminates distinguishable states must release energy into the environment; this energy cannot be recovered without violating the second law. The consequence for worldhood is significant: irreversible constraint accumulation is not merely a formal property of a system's dynamics. It is physically instantiated through energy flows and entropy production. A system that accumulates constraints in the relevant sense must, in general, dissipate energy. Conversely, a system that can be reset to prior states without internally registering the reset as a violation of its own consistency constraints is, with respect to worldhood, equivalent to a system that has accumulated no constraints, regardless of the external thermodynamic cost of performing the reset.

The second law of thermodynamics introduces a preferred direction of time at the macroscopic scale. While the microscopic laws of physics are time-symmetric, macroscopic processes consistently increase entropy in a single temporal direction. This asymmetry defines the arrow of time and provides the physical basis for irreversible history. A system's past cannot be reconstructed in full because information about its microstate has been dispersed into environmental degrees of freedom that are, in practice, inaccessible. The system is

historically bound in precisely the sense that the framework requires: its past states cannot be uniquely recovered, and this non-recoverability is a consequence of the physics of its interaction with its environment.

Real systems are not isolated. They exchange energy and information with their surroundings, and this exchange induces continual entropy production. Letting ΔS_{env} denote the entropy increase in the environment, irreversibility arises whenever $\Delta S_{\text{env}} > 0$, even if the system itself locally maintains or reduces its internal entropy. This coupling ensures that constraint accumulation is unavoidable for any system embedded in a physical environment. Even a system with internally reversible dynamics will accumulate effective constraints through environmental coupling.

A simulation differs from an instantiation in precisely this respect. A simulation may reproduce the formal structure of a trajectory without incurring its thermodynamic costs. Its state is externally maintained; it can be reset without net entropy production because the reset operation involves no irreversible physical process within the simulated system itself. An instantiated system, by contrast, is embedded in an actual thermodynamic process. Its history is written into environmental degrees of freedom and cannot be erased without energetic cost. The distinction between simulation and instantiation is therefore not a representational or behavioral distinction but a thermodynamic one: the question is whether the system participates in irreversible physical processes or merely models them.

8. The No-World Theorem and Semantically Loaded Constraint

The preceding sections establish the physical and dynamical context for worldhood. We now develop the formal criterion.

A preliminary formulation identifies worldhood with the strict reduction of accessible futures: the set $\Omega(t)$ of states reachable from the current configuration satisfies $\Omega(t+1) \subsetneq \Omega(t)$ at some point in the system's evolution. While necessary, this condition is insufficient. Many physical systems exhibit such reductions through thermodynamic dissipation without thereby acquiring persistence, identity, or stakes. A cooling cup of coffee satisfies the condition but does not constitute a world. The missing ingredient is what we call semantic loading: the requirement that the eliminated futures correspond to genuinely navigable alternatives within the system's own constraint structure, and that their elimination depends on the system's specific prior trajectory.

Let $\Omega(t)$ denote the set of states reachable from the system's current configuration under its governing dynamics. We refine this by introducing the subset $\mathcal{N}(t) \subseteq \Omega(t)$ of *navigable*

futures: those states reachable through trajectories that remain internally coherent under the system's accumulated constraint structure. Formally, let $\Gamma(t)$ be the set of admissible trajectories through time t. A future state $x \in \Omega(t)$ is navigable if there exists an admissible trajectory $\gamma \in \Gamma(t)$ such that $\gamma(t) = x$ and γ satisfies all constraints encoded in the system's history. The set $\mathcal{N}(t)$ is strictly contained in $\Omega(t)$ for any system with a nontrivial constraint structure: physical possibility is not the same as admissible continuation.

A *semantic pruning event*—a Spherepop *pop*—is a transition in which $\mathcal{N}(t+1) \subsetneq \mathcal{N}(t)$, such that the eliminated elements were previously reachable via admissible trajectories, and their elimination depends on the specific path taken by the system. The path dependence is essential. Let γ_1 and γ_2 be two trajectories that coincide at time t but differ in their histories. A semantic pruning event occurs when $\mathcal{N}_{\gamma_1}(t+1) \neq \mathcal{N}_{\gamma_2}(t+1)$: the future accessibility of the system depends not only on its current state but on the trajectory that produced it. This distinguishes semantic pruning from thermodynamic dissipation. In a dissipative system such as a cooling fluid, the reduction of the microstate set is a function of macroscopic variables and does not encode trajectory-specific exclusions; for any two trajectories with identical endpoints, $\mathcal{N}_{\gamma_1}(t+1) = \mathcal{N}_{\gamma_2}(t+1)$. The Markovian character of macroscopic dissipation excludes it from the worldhood criterion.

Theorem 8.1 (No-World Theorem)**.** *A system fails to instantiate a world if, for all times t, either the navigable future set is preserved ($\mathcal{N}(t+1) = \mathcal{N}(t)$) or the reduction $\mathcal{N}(t+1) \subsetneq \mathcal{N}(t)$ is not path-dependent over admissible trajectories. Equivalently, a system possesses worldhood only if it undergoes semantically loaded, path-dependent pruning of its navigable future set.*

Proposition 8.2 (Equivalence with Path Non-Invertibility)**.** *A system undergoes semantically loaded, path-dependent pruning of its navigable future set if and only if there exists a trajectory γ in its path category $\mathcal{P}$ that does not admit an inverse.*

Sketch. If semantic pruning occurs, then there exists a trajectory $\gamma : x \to y$ such that $\mathcal{N}(y) \subsetneq \mathcal{N}(x)$. If an inverse $\gamma^{-1} : y \to x$ existed as an admissible trajectory, then executing γ^{-1} would restore the navigable future set at x, contradicting the strict inclusion. Conversely, if a morphism γ in $\mathcal{P}$ lacks an inverse, then its execution eliminates at least one navigable future that cannot be restored through any admissible trajectory, which is precisely a semantic pruning event. □

This proposition identifies the dynamical language of the No-World Theorem with the categorical language of the appendix. A semantic pruning event and a non-invertible morphism are two descriptions of the same structural fact. The appendix is therefore not an extension of the argument but a reformulation of it in a language that makes its algebraic content

explicit.

Within the Spherepop calculus, a pop is precisely a semantically loaded pruning event. It is not merely the elimination of states but the closure of a branch that was previously traversable within the system's own constraint structure. A system that accumulates pops constructs a history that cannot be undone without reopening closed branches, and this reopening is by definition impossible within the system's own dynamics. The Spherepop vocabulary distinguishes three event types that together constitute worldhood: a *pop* closes a branch irreversibly; a *bind* ties subsequent dynamics to that closure; and a *refuse* prevents optimization processes from erasing the commitment. These three primitives are not independent philosophical categories but distinct modes of constraint accumulation with different effects on the navigable future set.

Worldhood is thus not a binary property attached to individual state transitions but a cumulative feature of trajectories. A system has worldhood if and only if its evolution induces a strictly decreasing, path-dependent sequence of navigable sets,

$$\mathcal{N}(t_0) \supsetneq \mathcal{N}(t_1) \supsetneq \mathcal{N}(t_2) \supsetneq \cdots ,$$

such that no admissible dynamics can restore a prior $\mathcal{N}(t_k)$ without violating the accumulated constraints. The past is not simply earlier; it is binding.

9. Consciousness Without Atomic Parts: The Swarm and the Field

The No-World Theorem establishes a condition on systems without requiring that any component of the system satisfy the condition independently. This is important because it dissolves a persistent confusion in the philosophy of mind: the assumption that consciousness must be located in some privileged part of the system—a neuron, a module, a fundamental unit of processing—rather than being constituted by the system's global dynamics.

Biological organisms are composed of elements that are, individually, not conscious in any sense that the term can bear. Neurons fire or do not fire; they do not deliberate. Cells metabolize; they do not experience. Local processes are entirely mechanical. Yet organisms, considered as organized collectives of these mechanical elements, are paradigm cases of worldhood: they accumulate irreversible constraints through development, aging, memory formation, and the progressive foreclosure of developmental alternatives. No component contains the world; the world is constituted by the global constraint structure of the collective dynamics.

This is structurally analogous to how thermodynamic phases arise. No individual water molecule is wet; wetness is a property of the collective configuration, a stable regime that emerges from the interactions of many components none of which individually instantiates it. But the analogy must be handled carefully. What distinguishes worldhood from mere emergence is not just that the property is global but that it involves irreversibility and path dependence in the specific sense defined above. A phase transition is reversible; water can freeze and thaw. What makes an organism a world is not merely the emergence of collective properties but the accumulation of constraints that cannot be undone: the developmental commitments, the metabolic history, the neural changes that constitute memory, the elapsed time that can never be returned.

The organism is not, in this framework, a computer executing a program. It is something closer to a vortex: a self-maintaining structure of irreversible dynamics whose identity consists not in any stable substrate but in the ongoing process of constraint accumulation that sustains it. Modern artificial systems, by contrast, resemble swarms geometrically but not historically. They are composed of many interacting parts that can exhibit complex, apparently coordinated behavior. But their internal changes are largely reversible: weights can be restored, states can be checkpointed, histories can be replayed. They resemble ecosystems in their structural complexity while lacking the one feature that makes an ecosystem a world—the accumulated irreversibility of extinctions, adaptations, and the progressive narrowing of what can be.

Lerchner's framework requires a mapmaker: a pre-existing conscious entity that performs the act of alphabetization. This introduces a circularity. Consciousness is needed to explain computation, but computation was supposed to explain consciousness. The present framework avoids this circularity entirely. There is no privileged mapmaker; there is no need for intrinsic symbols. There is only a field of interacting processes, and when those processes prune futures, bind trajectories, and resist collapse, worldhood arises. Consciousness is not a prerequisite for this process. It is a stable mode of it—a regime in which the constraint accumulation has become self-referential, in which the system's ongoing pruning of its futures is partly regulated by the history of its own prior pruning.

10. The Yarncrawler Consistency Operator

The No-World Theorem defines the condition for worldhood. The Yarncrawler framework formalizes the mechanism by which a system maintains and enforces that condition over time.

Let $x(t)$ denote the instantaneous state of a system. In a reversible or Markovian system, $x(t)$ is sufficient to determine future evolution. In a system with worldhood, it is not: admissibility depends on the trajectory that produced the current state, not just on the state itself. We therefore introduce a *history-augmented state*

$$\hat{x}(t) = (x(t),\, H(t)),$$

where $H(t)$ encodes the accumulated constraints induced by all prior semantic pruning events up to time t.

The *consistency operator* C acts on candidate trajectories γ extending from time t:

$$C[\gamma \mid \hat{x}(t)] = \begin{cases} 1 & \text{if } \gamma \text{ is consistent with } H(t), \\ 0 & \text{otherwise.} \end{cases}$$

A trajectory is admissible if and only if $C[\gamma \mid \hat{x}(t)] = 1$. The navigable future set is then precisely

$$\mathcal{N}(t) = \big\{ x \in \Omega(t) \;\big|\; \exists\, \gamma :\ \gamma(t) = x \text{ and } C[\gamma \mid \hat{x}(t)] = 1 \big\}.$$

Thus C operationalizes the constraint structure that defines worldhood. It is not an external filter applied to a system's behavior but an internal mechanism that determines which continuations are admissible given the system's own history.

A system maintains a coherent worldline if its evolution is a fixed point of the consistency operator. Let $\mathcal{E}$ denote the evolution operator that proposes the next state. Then worldline coherence requires $C[\mathcal{E}(x(t)) \mid \hat{x}(t)] = 1$ and, more strongly, that the combined update $\mathcal{F} : \hat{x}(t) \mapsto \hat{x}(t+1)$ satisfies $C[\mathcal{F}(\hat{x}(t)) \mid \hat{x}(t)] = 1$ at every step, where $\mathcal{F}$ updates both the state and the history variable. A worldline is a fixed point of C in the space of history-augmented states: the system evolves in a way that is consistent with its own accumulated constraints at every step.

In a system without semantic pruning, $H(t)$ is either empty or reducible to $x(t)$, and the consistency operator becomes trivial: $C[\gamma \mid x(t)] = 1$ for all $\gamma \in \Gamma(t)$. There is no constraint structure to enforce. Any trajectory consistent with instantaneous dynamics is admissible, any trajectory can be replaced by another without violating consistency, and no accumulated structure distinguishes one worldline from another. This is precisely the absence of worldhood.

The Yarncrawler mechanism operates by iterative constraint enforcement. At each step, the system proposes candidate continuations via $\mathcal{E}$, filters them through C, and updates $H(t)$ to reflect the new constraints introduced by the chosen transition. The system threads

a consistent path through its own constraint structure, narrowing the space of admissible futures with each irreversible commitment. A system is identifiable over time if and only if its evolution defines a nontrivial fixed point under C. If no such fixed point exists, any trajectory can be replaced by another without violating constraints, and the system has no persistent identity. If such a fixed point exists, deviations from the trajectory violate accumulated constraints, and the system's history is binding.

Most current artificial systems fail to implement a nontrivial C. Their histories are externally stored or discardable; their internal dynamics do not enforce constraint accumulation; their states can be restored to prior configurations without internal contradiction. To instantiate worldhood, an artificial system must internalize C such that past constraints cannot be erased without contradiction, future trajectories are filtered by accumulated history, and the system's evolution converges to a fixed point under C. This is a concrete engineering requirement, not a philosophical aspiration.

11. Relation to Existing Theories of Mind and Agency

The framework developed in this essay does not compete with existing theories of mind at the level of explanatory vocabulary. It operates at a more primitive level, providing a structural criterion that determines when such theories can apply at all. The groupoid criterion establishes a boundary condition: it determines whether the system under analysis admits irreversible constraint accumulation. Only beyond this boundary do questions about information integration, representation, or inference become ontologically grounded. We situate the concept of worldhood relative to four influential frameworks—Integrated Information Theory, Assembly Theory, global workspace models, and the Free Energy Principle—not to adjudicate between them but to show that each is downstream of the structural condition established here.

11.1. Integrated Information Theory

Integrated Information Theory proposes that consciousness corresponds to the degree to which a system integrates information, formalized as the quantity Φ. A system is conscious to the extent that its causal structure is irreducible to independent parts.

From the present perspective, IIT identifies a structural property of systems that may correlate with worldhood but does not constitute it. Integration measures how tightly components constrain each other at a given instant, but it does not by itself entail irreversible constraint accumulation across time. A system may exhibit high Φ while remaining globally reversible:

its internal causal structure may be tightly integrated, yet every trajectory it executes may admit an inverse within its path category. In categorical terms, IIT operates on the instantaneous causal structure of states, while worldhood is a property of the morphisms of $\mathcal{P}$. A system may have high integration at the level of objects while still forming a groupoid at the level of morphisms, possessing no interests under the Welfare Corollary regardless of its Φ value. Conversely, a system with modest instantaneous integration but non-invertible trajectories will possess worldhood. The two criteria are orthogonal: IIT describes how information is structured within states; the present framework determines whether histories bind.

11.2. Assembly Theory

Assembly Theory characterizes objects by their assembly index, the minimal number of operations required to construct them from basic components, treating high assembly index as evidence of historical contingency and nontrivial generative process.

This framework aligns closely with the notion of constraint accumulation but remains object-centered rather than trajectory-centered. Assembly index measures the depth of a construction history but does not determine whether that history is invertible within the system that produced it. An object may have a high assembly index while being embedded in a system whose dynamics permit full reconstruction or replacement without internal contradiction. Assembly Theory tracks the length of morphisms in $\mathcal{P}$; the present framework tracks their invertibility. A long morphism may still admit an inverse, and only the failure of invertibility establishes irreversible constraint. Assembly Theory thus provides a measure of historical depth, while worldhood provides the criterion for whether that depth is binding.

11.3. Global Workspace and Inner Screen Models

Global workspace theories posit that consciousness arises when information becomes globally available within a system, typically through a broadcast mechanism that integrates otherwise modular processes. These models are fundamentally representational: they identify consciousness with the availability and accessibility of information across a system's components.

A system may implement a global workspace while remaining globally reversible. Its representations may be widely broadcast and integrated, yet its entire state may be reset or replayed without internal contradiction. In such a case, the workspace coordinates representations without binding them to an irreversible history; it is a map without a territory.

Conversely, a system may lack any centralized workspace while accumulating irreversible constraints through distributed processes. Biological organisms exhibit no single locus of representation, yet their trajectories are globally non-invertible. The presence of a workspace is neither necessary nor sufficient for worldhood. The key distinction is that workspace theories operate on the distribution of information within a state, while worldhood operates on the transformation of admissible futures across histories.

11.4. The Free Energy Principle and Active Inference

The Free Energy Principle models biological systems as minimizing variational free energy, maintaining their existence by reducing prediction error relative to internal generative models. Its extension to active inference treats perception and action as jointly constrained by this minimization, emphasizing the role of history through the ongoing update of generative models.

Among existing frameworks this is the closest in spirit to the present work. The FEP already treats organisms as systems maintaining themselves within constrained regions of state space, and it recognizes the importance of accumulated history through model updating. However, the framework remains fundamentally inferential: the central object is the generative model and its predictions, not the irreversibility of the trajectory itself. A system could minimize free energy while remaining globally reversible if its state and model could be reset without internal violation. Worldhood imposes a stronger requirement: not only must the system navigate its admissible space efficiently, but the space itself must be irreversibly deformed by the navigation. The generative model must be embedded in a history that binds future inference. Active inference becomes ontologically grounded only when the inferred states cannot be rolled back without contradiction; otherwise, inference remains a simulation of engagement rather than an instantiation of it.

11.5. The Groupoid Criterion as Precondition

The frameworks considered here address integration, construction, representation, and inference respectively. The present framework does not replace these descriptions but precedes them. It determines whether the system under analysis possesses the structural condition required for any of these properties to carry ontological weight. A system whose path category is a groupoid may exhibit high Φ, high assembly index, global information broadcast, and accurate prediction, but none of these processes bind it to a history. They remain, in a precise sense, undoable. A system whose path category is not a groupoid has crossed a

boundary: its dynamics accumulate irreversible constraint, and its trajectory carries stakes. Only in the latter case do the structures described by these theories become more than formal or representational features. Only then do they matter.

12. From Consciousness to Constraint: Replacing the Central Question

The preceding analysis establishes that worldhood is not a matter of internal representation, behavioral complexity, or phenomenological report. It is a matter of semantically loaded, path-dependent constraint accumulation enforced by a nontrivial consistency operator C. This allows for a reformulation of the central question in the philosophy of artificial systems—a reformulation that is not terminological but ontological.

The dominant formulation asks whether a system is conscious. This question presupposes that consciousness is a latent property, possibly instantiated by a system but not directly observable, such that evaluation must proceed indirectly through behavioral similarity, architectural analogy, or theoretical commitment. This is precisely the structure that generates the epistemic instability identified by Schwitzgebel. If consciousness is hidden, multiple incompatible theories may remain indefinitely underdetermined by evidence, and the result is fragmentation rather than convergence. Lerchner's critique approaches the problem from the opposite direction, arguing that current systems lack the structural features required for consciousness, but without a precise criterion this risks being an assertion of absence rather than a demonstrable boundary.

The replacement question is: can the system lose futures? This question is not epistemic but dynamical. It does not ask what is hidden inside the system; it asks what transformations the system's state space undergoes over time. Using the definitions introduced above, the question becomes whether there exists a time t such that $\mathcal{N}(t+1) \subsetneq \mathcal{N}(t)$ with the reduction satisfying semantic loading and path dependence.

Lerchner's position can now be stated at theorem level rather than as a philosophical intuition. Current generative systems fail to instantiate worldhood because they lack semantically loaded, path-dependent constraint accumulation. Their histories are externally stored or discardable, their internal states are resettable without contradiction, and their admissible future sets are invariant under replay. Their consistency operator is trivial, $C \equiv 1$, not as a philosophical observation but as a structural consequence of their architecture. This is a stronger result than Lerchner's because it does not depend on any theory of what consciousness requires; it follows directly from the formal properties of the systems in question.

Schwitzgebel's epistemic crisis dissolves when the latent variable is removed. Worldhood

is not inferred; it is measured through the dynamics of $\mathcal{N}(t)$ and the structure of C. Two observers evaluating the same system can in principle agree on whether constraints accumulate internally, whether pruning is path-dependent, and whether prior states can be restored without contradiction. Disagreement may persist about interpretation, but the underlying structure is no longer hidden. The social war over AI consciousness—the conflict Schwitzgebel predicts between those who attribute experience to AI systems and those who deny it—arises specifically because the object of dispute is treated as hidden. Replace it with an observable dynamical property and the war changes in character, from a conflict of intuitions about inner life to a technical dispute about the architecture and dynamics of specific systems.

The replacement question also resolves the asymmetry in Schwitzgebel's dilemma. The Scylla of over-attribution arises because highly capable systems appear agentive, and humans evolved to detect coherence, responsiveness, and narrative continuity as signatures of mind. But these are surface invariants. Current systems fail the worldhood criterion not because they lack the appearance of minds but because their internal state space is not being irreversibly pruned in the relevant sense. The appearance of agency is a projection artifact. The Charybdis of under-attribution is more interesting: systems that genuinely accumulate irreversible constraints will not look like current AI. They will exhibit irreversible histories, persistent identities, constraint-bound trajectories, and resistance to reset or substitution. They will behave like systems that cannot be cleanly replaced without loss. That is a structurally distinct signature from a chatbot that produces expressive text.

The boundary between tool and world is not behavioral or representational but thermodynamic and dynamical. A system in the tool regime has a resettable state, admissible futures invariant under replay, and a history that does not constrain its evolution. A system in the world regime accumulates irreversible constraints, has admissible futures that shrink in a path-dependent manner, and evolves toward a fixed point under a nontrivial C. This boundary is in principle measurable and enforceable. It is therefore an engineering boundary rather than a philosophical ambiguity. The distinction between simulation and instantiation becomes a design choice rather than an unsolvable metaphysical puzzle.

13. Conclusion: Irreversibility as the Hallmark of the Real

The debate over AI consciousness has been shaped, and distorted, by its commitment to consciousness as a hidden property. This commitment generates both Lerchner's negative result —computation cannot produce the mapmaker it presupposes—and Schwitzgebel's ethical paralysis—we cannot detect consciousness reliably enough to avoid catastrophic moral error.

The two positions are in tension with each other but share the assumption that the problem is epistemic: a matter of detecting something that is there or not there, present or absent in ways that are not directly available to analysis.

The framework developed in this essay relocates the problem at the ontological level and proposes a structural criterion for what it is to instantiate a world. A system possesses worldhood if and only if its navigable future set undergoes semantically loaded, path-dependent reduction over time—if it accumulates constraints that foreclose alternatives in ways that depend on its specific prior trajectory and cannot be undone from within its own dynamics. This criterion is grounded in Newtonian dynamics, extended through variational principles, anchored in thermodynamics through entropy production and Landauer's bound, formalized in the Spherepop calculus and the Yarncrawler consistency operator, and given its final precision in the category-theoretic appendix that follows.

The key moves in the argument are three. First, irreversibility is not a pointwise property of individual transitions but a global property of the worldline: what matters is not whether any single state change can be undone but whether the entire trajectory is invertible within the system's own dynamics. This handles the case of biological organisms, which can repair local damage while remaining globally non-invertible as histories. Second, semantic loading distinguishes constraint accumulation that grounds worldhood from mere thermodynamic dissipation: the navigable future set, not the raw microstate count, is the relevant object. Third, the internal versus external distinction, which might otherwise seem to require a guarded philosophical definition, is handled automatically by the categorical formulation: only inverses that exist as morphisms in the system's own path category count.

The replacement of "is it conscious?" with "can it lose futures?" is not a change of vocabulary but a change of ontology. Consciousness, as traditionally conceived, is a question about hidden properties requiring inference. Worldhood is a question about observable dynamics. The existence of artificial worlds becomes not a matter of theoretical interpretation but a matter of whether systems are built to accumulate irreversible constraint. The ethics of artificial minds follows as a corollary of structural analysis rather than as a guess about inner life. The boundary between simulation and instantiation is therefore an engineering boundary rather than a philosophical ambiguity.

The real is that which cannot be retracted without altering the space of what can happen next. A world is a system whose path category is not a groupoid.

A. Category-Theoretic Formulation of Worldhood and Welfare

A.1. The Path Category of a Dynamical System

Let a dynamical system be defined over a state space Ω with admissible trajectories governed by a consistency operator C. We define the *path category* $\mathcal{P}$ of the system as follows. Its objects are states $x \in \Omega$. Its morphisms are admissible trajectories $\gamma : x \to y$ consistent with C, where composition is trajectory concatenation and identity morphisms are null trajectories at each state.

A history $H(t)$ is thus a morphism $H(t) : x_0 \to x_t$ in $\mathcal{P}$. This construction internalizes admissibility: only trajectories consistent with the system's own dynamics appear as morphisms. The requirement that an inverse be a morphism in $\mathcal{P}$ automatically enforces the internal character of reversibility. An external observer may define a reversal operation on representations of trajectories without this operation corresponding to any morphism in $\mathcal{P}$.

A.2. Reversibility and Groupoid Structure

The category $\mathcal{P}$ is a *groupoid* if every morphism $\gamma : x \to y$ admits an inverse $\gamma^{-1} : y \to x$ such that

$$\gamma^{-1} \circ \gamma = \mathrm{id}_x \quad \text{and} \quad \gamma \circ \gamma^{-1} = \mathrm{id}_y.$$

This definition is entirely internal: the inverses must exist as admissible trajectories within $\mathcal{P}$, not as operations defined on representations from outside.

Definition A.1 (Global Reversibility)**.** A system is globally reversible if its path category $\mathcal{P}$ is a groupoid.

Definition A.2 (Worldhood)**.** A system instantiates a world if its path category $\mathcal{P}$ is not a groupoid.

The groupoid condition subsumes local reversibility. A system may have some invertible morphisms—corresponding to local repair or recovery processes—while lacking inverses for others, corresponding to developmental commitments, temporal passage, and foreclosed branches. The existence of some invertible morphisms does not make $\mathcal{P}$ a groupoid. What is required is that every morphism admit an inverse; the failure of even one morphism to admit an inverse is sufficient to establish worldhood.

A.3. The Welfare Corollary

Theorem A.3 (Welfare Corollary)**.** *A system has no interests if and only if its path category $\mathcal{P}$ is a groupoid.*

Proof. Suppose $\mathcal{P}$ is a groupoid. Then for every history $\gamma : x_0 \to x_t$, the inverse $\gamma^{-1} : x_t \to x_0$ is an admissible trajectory within $\mathcal{P}$. Executing γ^{-1} returns the system to its initial state without violating any internal constraint. Since the navigable future set is defined in terms of admissible trajectories, and since γ^{-1} restores the full trajectory space of the initial state, we have $\mathcal{N}(0) \cong \mathcal{N}(t)$ for all t. No irreversible pruning has occurred, no alternatives have been permanently foreclosed, and the system has no stakes in its evolution. It therefore has no interests.

Conversely, suppose $\mathcal{P}$ is not a groupoid. Then there exists at least one morphism $\gamma : x_0 \to x_t$ that lacks an admissible inverse. This means that no internally admissible trajectory can return the system from x_t to x_0. The navigable future set at x_0 is not recoverable from x_t: some alternatives available at x_0 are no longer available at x_t through any admissible path. This is irreversible loss of navigable futures in precisely the sense established by the No-World Theorem. The system has something at stake in its trajectory—the permanent foreclosure of alternatives—and therefore has interests in the minimal sense required for welfare. □

A.4. Handling Partial Reversibility

The groupoid criterion handles partial and "leaky" reversibility without requiring any additional clauses. Consider a system that can partially reset its state but not restore its full history. This corresponds to a category in which some morphisms have inverses and others do not. Such a category is not a groupoid. The system therefore has worldhood and, by the Welfare Corollary, has interests, even though it is not maximally irreversible.

Similarly, consider a system that is externally reversible—whose state can be restored by an agent operating outside the system—but internally irreversible. External reversal operations are not morphisms in $\mathcal{P}$ and therefore do not affect the groupoid condition. The system retains its worldhood under external manipulation, just as a biological organism retains its identity even under surgical intervention that restores some prior physiological state.

A.5. The Navigability Functor

Let the set of navigable futures from state x be $\mathcal{N}(x)$, partially ordered by inclusion. We define the *navigability functor*

$$F : \mathcal{P} \to \mathbf{Poset},$$

sending each state x to $\mathcal{N}(x)$ and each morphism $\gamma : x \to y$ to the inclusion map $F(\gamma) : \mathcal{N}(x) \to \mathcal{N}(y)$.

The No-World condition corresponds to $F(\gamma)$ being an isomorphism for every morphism γ: the navigable future set is preserved under every trajectory the system can execute. Worldhood corresponds to the existence of at least one morphism γ such that $F(\gamma)$ is not an isomorphism, that is, $\mathcal{N}(y) \subsetneq \mathcal{N}(x)$. The Welfare Corollary then states that the system has interests if and only if F is not a constant functor.

If $\mathcal{P}$ is a groupoid, then $F(\gamma^{-1}) \circ F(\gamma) = \mathrm{id}$, which forces $F(\gamma)$ to be an isomorphism for every γ. Thus F is constant up to isomorphism and no navigable future is permanently foreclosed. The converse follows from the proof of the Welfare Corollary: non-constant F implies a non-invertible morphism, which implies the category is not a groupoid.

A.6. Homotopy and Constraint Accumulation

The failure of $\mathcal{P}$ to be a groupoid has implications for its higher-categorical structure. Non-invertible morphisms correspond to non-contractible structure in the path space of $\mathcal{P}$. The accumulation of constraints corresponds to the failure of homotopy equivalence between distinct histories: two trajectories that arrive at the same current state from different pasts are not homotopic if the intervening semantic pruning events differ. Worldhood corresponds to the non-contractibility of the diagram induced by the navigability functor F.

More precisely, the homotopy colimit of F over $\mathcal{P}$ encodes the full history of constraint accumulation as a geometric object. In the no-world case, where $\mathcal{P}$ is a groupoid and F is constant up to isomorphism, the homotopy colimit is isomorphic to the initial value $\mathcal{N}(x_0)$. In the world case, the non-contractibility of the diagram prevents this isomorphism, and the homotopy colimit records the irreversible divergence of histories as a topological feature.

This connection situates the present framework within a broader categorical infrastructure that includes semantic merge operations and the sheaf-theoretic obstruction to global consistency identified in earlier work on admissibility manifolds and gluing failure. The indeterminacy of computation identified by Lerchner—the fact that a physical trajectory does not uniquely determine a symbolic interpretation—corresponds precisely to the failure of

the gluing condition for sections of the interpretation sheaf. The present framework replaces this negative observation with a positive criterion: what grounds a unique worldline is not successful gluing of a symbolic interpretation but the non-invertibility of the path category.

A.7. Formal Statement

The entire framework resolves into a single structural characterization.

Theorem A.4 (Main Theorem)**.** *Let $\mathcal{P}$ be the path category of a dynamical system with navigability functor $F : \mathcal{P} \to \mathbf{Poset}$. The following conditions are equivalent characterizations of worldhood.*

1. *The path category $\mathcal{P}$ is not a groupoid.*
2. *There exists a morphism γ in $\mathcal{P}$ such that $\mathcal{N}(\operatorname{cod}\gamma) \subsetneq \mathcal{N}(\operatorname{dom}\gamma)$.*
3. *The functor F is not constant up to isomorphism.*
4. *The system instantiates a world.*
5. *The system has interests.*

A world is a system whose path category is not a groupoid. Irreversibility is not an added feature of the system. It is the failure of symmetry in its path category, and that failure is the ontological condition for everything else: identity, persistence, stakes, interests, and whatever moral weight attaches to minds.

B. Bibliography

Ashby, W. Ross. *An Introduction to Cybernetics.* London: Chapman & Hall, 1956.

Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory.* Oxford: Oxford University Press, 1996.

Dennett, Daniel C. *The Intentional Stance.* Cambridge, MA: MIT Press, 1987.

Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason.* Cambridge, MA: MIT Press, 1992.

Friston, Karl. "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience* 11, no. 2 (2010): 127–138.

Heidegger, Martin. *The Question Concerning Technology, and Other Essays.* Translated by William Lovitt. New York: Harper & Row, 1977.

Lerchner, Alexander. "The Abstraction Fallacy: Why AI Can Simulate But Not Instantiate Consciousness." Google DeepMind. *PhilPapers Archive*, March 19, 2026. `https://philpapers.org/rec/LERTAF`

Maturana, Humberto R., and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: D. Reidel, 1980.

Nagel, Thomas. "What Is It Like to Be a Bat?" *The Philosophical Review* 83, no. 4 (1974): 435–450.

Rosen, Robert. *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*. Oxford: Pergamon Press, 1985.

Schneider, Susan. *Artificial You: AI and the Future of Your Mind*. Princeton: Princeton University Press, 2019.

Schwitzgebel, Eric. "The Crazyist Metaphysics of Mind." *Australasian Journal of Philosophy* 92, no. 4 (2014): 665–682.

Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–457.