Consciousness as Constrained Continuation

Worldhood, Admissibility, and the Limits of a Structural Account

Flyxion
Independent Researcher

November 2026

Abstract

Accounts of machine consciousness usually ask whether a system has a hidden property and then argue about how it could be detected. This essay develops a constraint-first alternative from the admissibility and continuation framework. A system is described by configurations that include its admissible operations, and the question becomes what that structure does over time: which futures it permanently closes, and whether its operative capacities depend on the path of arrival in ways that a declared task-level description does not capture. From these notions the essay defines loss events, history-sensitive organization, and worldhood relative to a declared boundary and projection, shows that systems whose transitions are all reversible from within have no permanent loss, and shows that the criterion is met by many systems that nobody takes to be conscious. Worldhood is therefore proposed as an enabling condition for historically consequential agency and not as a sufficient condition, and its necessity for consciousness is treated as a further hypothesis. The essay then adds sensitivity to loss and nested repair, argues that selective repair of unfamiliar damage favors capacity models of bounded depth, states what the account implies about identity and copying under an explicit convention, and proposes a program that varies irreversibility and restoration cost separately. It states what the account does not settle, namely why any such structure should be accompanied by experience.

1 Introduction

The standard framing of artificial consciousness treats experience as a latent property that a system either has or lacks, and the debate then turns on evidence. Behavioral criteria are criticized because behavior can be produced without experience, theory-based criteria are criticized because the theories disagree, and the resulting uncertainty is described as an epistemic predicament to be managed. Two recent statements of the predicament make the structure plain. One argues that computation depends on a prior act of interpretation and so cannot be the source of the experience that supplies the interpretation [7]. The other argues that no theory of mind avoids radically counterintuitive commitments [12], so that theory choice is unlikely to settle which systems are conscious, and that the resulting uncertainty confronts decisions about artificial systems of debatable moral standing with a dilemma [13]. Both treat consciousness as a hidden variable and ask how it could be inferred.

The framework developed here takes a different route and is correspondingly more modest in some respects and more specific in others. It does not attempt to identify experience with any structure. It asks which structural conditions make questions of permanent loss, historical consequence, and repair well defined, states those conditions precisely enough to be checked against a system's dynamics relative to a declared boundary, and then asks which further conditions are needed before the resulting profile resembles the one found in organisms that are taken to be conscious. The account is therefore a proposal about enabling structure and functional profile. Whether the structure is accompanied by experience is a further question, and the final section states plainly that the account leaves it open.

The argument proceeds as follows. Section 2 separates the questions that the word consciousness carries and says which of them the account addresses. Sections 3 and 4 introduce admissibility, reachability, and loss, define worldhood, and show the sense in which systems that are fully reversible from within lack permanent loss. Section 5 treats identity as attached to trajectories and states what this implies about copying and substrate. Section 6 presents the four operators of the calculus as an empirical vocabulary with explicit obligations. Section 7 introduces sensitivity to loss and argues that selective repair of unfamiliar damage favors capacity models of bounded depth. Section 8 states the positive claim in the form in which it is currently earned, and Section 9 describes a program that could discriminate between this account and its rivals. Section 10 examines the explanatory remainder and three objections, and Section 11 concludes. The formal results are of three kinds. The propositions on nestedness, reversibility, and strict passage characterize the formal object. Those on state augmentation and on resets are negative results showing that a tempting response fails. Those on option loss, repair depth, repair information, and sensitivity relate the apparatus to value, cost, and information.

2 Two Questions and the Scope of the Account

The word consciousness carries at least two questions. The first asks what organization accompanies conscious states, including the forms of attention, self-modeling, memory, and

report that occur when a system is conscious. The second asks why any such organization is accompanied by experience [2, 10]. The distinction between access and phenomenal consciousness marks the same difference [1]. A proposal can contribute to the first question without touching the second, and the common failure in this literature is to present a contribution to the first as if it answered the second.

The account developed here addresses the first question and a neighboring question about moral relevance. It proposes a set of structural conditions, argues that some of them enable the functional profile associated with conscious systems, and argues that systems satisfying them have option-stakes in a precise sense, meaning that permanent loss of executable futures can occur for them. It does not claim that the structure is identical with experience, and it does not claim that the structure is sufficient for experience. The essay returns to this boundary in Section 10, where the three ways in which such a claim could be developed are distinguished. For present purposes the relevant point is that the account is stated at the strength that it can earn. A criterion for the enabling conditions of a conscious regime is a more modest result than a theory of experience, and it is also a result that can be tested.

The account belongs to a family of views that emphasize precariousness and self-maintenance as conditions of mind [6, 9, 14]. These views hold that living systems have stakes because they must continually produce the conditions of their own persistence. The present account shares this emphasis and differs in two respects. It generalizes the notion of stakes beyond metabolism by defining it in terms of loss of reachable futures, so that the criterion is stated in terms of dynamics and not of biochemistry. And it separates the conditions that make stakes possible from the further conditions, developed below, that make a system sensitive to its own exposure.

3 Admissibility, Executable Futures, and Loss

The framework begins with a system's admissibility structure, the specification of which operations the system can perform. Because performing an operation can change which operations are admissible, the objects on which the dynamics run must include the structure itself. Let Z be a set of configurations, each of which fixes a state $\pi(z) \in X$ together with the admissibility structure that currently governs the system. A relation $z \to z'$ on Z records that z' can follow z by one admissible operation, and the update of the admissibility structure is part of that transition. The executable futures of a configuration are the states reachable by finite sequences of such transitions,

$$\mathrm{Ex}(z) \;=\; \{\, \pi(z') : z \to^{*} z' \,\},$$

where $\to^{*}$ is the reflexive and transitive closure of $\to$. A trajectory is a sequence (z_t) with $z_t \to z_{t+1}$ at each step.

This construction differs from a reachability relation frozen at a moment. If $a \rightsquigarrow b$ and $b \rightsquigarrow c$ hold at time t, transitivity places c among the frozen futures of a, yet executing the step from a to b may remove the operation that led to c. The frozen futures describe what is

possible under the structure as it stands at one moment, and the executable futures describe what the evolving system can actually do. The difference is directly relevant to the framework and is the reason the structure is made part of the configuration. The frozen case is retained as the special case in which no transition alters the admissibility relation. Then $Z = X$, and $\mathrm{Ex}(x)$ is the set of states reachable from x under a fixed reflexive and transitive relation $\rightsquigarrow$.

Two kinds of operation should be distinguished. Interrogative operations read the structure, as when a reasoner evaluates which moves are available. Modificatory operations write to it, as when development, regulation, or repair changes which operations are admissible. A modification is an event connecting two admissibility structures and is not merely a difference between two snapshots indexed by an external clock. The distinction allows the framework to treat learning that generates new operators, for example a repair strategy, as an escalation from operating within the structure to modifying it.

Definition 1 (Loss of eventual potential)**.** Fix a window $[0, T]$ and a trajectory (z_t). A state y is *lost at time* $t < T$ if $y \in \mathrm{Ex}(z_t)$ and $y \notin \mathrm{Ex}(z_s)$ for every s with $t < s \leq T$. A *loss event* is a time $t < T$ at which some state is lost.

Terminal times are excluded, since the condition would otherwise hold vacuously at the end of a finite window.

Proposition 1 (Executable futures are nested)**.** *For any configurations z and z' with $z \to^* z'$, we have $\mathrm{Ex}(z') \subseteq \mathrm{Ex}(z)$. Consequently, along a trajectory (z_t) and for every $t < T$, a state y is lost at time t if and only if $y \in \mathrm{Ex}(z_t) \setminus \mathrm{Ex}(z_{t+1})$, and a state that has left the executable futures cannot reappear in them.*

Proof. If $y \in \mathrm{Ex}(z')$, there is a configuration w with $z' \to^* w$ and $\pi(w) = y$. Composing $z \to^* z'$ with $z' \to^* w$ gives $z \to^* w$, so $y \in \mathrm{Ex}(z)$. Along a trajectory the resulting inclusions give $\mathrm{Ex}(z_s) \subseteq \mathrm{Ex}(z_{t+1}) \subseteq \mathrm{Ex}(z_t)$ for $s > t$. Absence from $\mathrm{Ex}(z_{t+1})$ therefore entails absence from every later set, and the converse follows by taking $s = t + 1$ in Definition 1. □

The proposition has a consequence that should be stated plainly. The set Ex measures eventual potential, including every route through learning and repair, so loss of eventual potential is immediate and cannot be undone by anything the system later does. Learning may enlarge the operations currently available and repair may restore capacities that were degraded, but any outcome obtainable through that learning or repair was already in the earlier eventual potential. A distinction between loss within a window and permanent loss therefore collapses for this object, and capacities that expand and later recover must be described by other objects. Suppose each transition carries a nonnegative cost. The *current repertoire* $N(z) = \{\pi(z') : z \to z'\}$ is the set of states reachable by a single admissible operation, and it can shrink and later be restored by repair. The *budget-limited futures* $\mathrm{Ex}_B(z)$ are the states reachable by sequences of transitions whose total cost does not exceed B, and they can leave and re-enter as resources are depleted and replenished. Both are contained in $\mathrm{Ex}(z)$, and neither is nested along trajectories. Three objects are thus distinguished: eventual potential, which can only shrink; the current repertoire, which can shrink and recover; and the budget-limited futures, which track what is practically recoverable. Loss in Definition 1 is loss of

eventual potential, and the other two describe incapacity and practical loss, which may be temporary.

Observation and reachability must also be kept apart. That a state was not recovered during an experiment does not establish that it lies outside the executable futures, since the latter concerns all admissible alternative continuations. Establishing it requires a model, exhaustive exploration, or a justified reachability analysis, and for a finite-state model it can sometimes be settled by a finite computation. The definition is relative to the system's own transitions, so a loss is the loss of something the system could execute and not a change in microstate count as an outside observer might compute it.

Definition 2 (History-sensitive organization)**.** Let $p : Z \to Y$ be a declared projection, the task-level description under which configurations are compared, and let $Z_0 \subseteq Z$ be a declared class of initial configurations, for example a single common configuration, so that differences acquired through history are distinguished from initial heterogeneity. The system exhibits *history-sensitive organization relative to p and Z_0* if there are trajectories (z_t) and (z'_t) with $z_0, z'_0 \in Z_0$ and a time t with $p(z_t) = p(z'_t)$ such that the two configurations differ in Ex, in N, or in Ex_B for some declared B.

The definition is relative to a projection by design. If the projection is the full configuration, configurations that coincide have the same executable futures and no history sensitivity can appear. What the definition measures is historically acquired organization that the declared description fails to capture, so that two histories arriving at the same described condition leave the system able to do different things. Enlarging the projection to include the operative variables removes the sensitivity relative to the enlarged description, and this is not a defect. It reflects the fact that history sensitivity is a property of the gap between a description and the system, and the classification of any example is accordingly stated together with its projection.

Proposition 2 (State augmentation)**.** *Let $\mathcal{H}_t$ be the set of admissible histories ending at time t. Suppose each history $h \in \mathcal{H}_t$ determines a reported state $p(h) \in X$ and an operative admissibility relation $A(h) \subseteq X \times X$, and define $z(h) = \big(p(h), A(h)\big)$. Then admissibility is a function of the augmented state z, even when it is not a function of the reported state p.*

Proof. Define G on the image of z by $G(x, R) = R$. For every history h we have $A(h) = G(z(h))$, so admissibility is determined by the augmented state. If there exist histories h and h' with $p(h) = p(h')$ but $A(h) \neq A(h')$, no function F on X can satisfy $A = F \circ p$, since it would assign two different values to the same argument. □

Remark. The proposition does not show that (x, A) suffices to predict subsequent changes in admissibility, since further historical variables may be needed for that purpose. It shows only that historical dependence relative to one state description can become present-state dependence under a richer description. Worldhood should therefore be assessed relative to an explicitly declared projection, and not defined by the impossibility of any state-based description.

Proposition 3 (Reversibility excludes loss)**.** *Suppose the case is frozen, so that $Z = X$ and the relation $\rightsquigarrow$ is fixed. Consider trajectories with at least one transition, and take loss within a window that extends past that transition. If $\rightsquigarrow$ is symmetric, no loss event occurs along any such trajectory. If $\rightsquigarrow$ is not symmetric, some trajectory contains a loss event. Equivalently, the thin category of reachability is a groupoid exactly when no trajectory contains a loss event.*

Proof. Suppose $\rightsquigarrow$ is symmetric. Take a trajectory (x_t), a time $t < T$, and $y \in \mathrm{Ex}(x_t)$, so that $x_t \rightsquigarrow y$. For any $s > t$ we have $x_t \rightsquigarrow x_s$ along the trajectory by transitivity, hence $x_s \rightsquigarrow x_t$ by symmetry, and therefore $x_s \rightsquigarrow y$ by transitivity. So $y \in \mathrm{Ex}(x_s)$ for all $s > t$, and no state is lost. Conversely, suppose $x \rightsquigarrow y$ but not $y \rightsquigarrow x$, and take the trajectory with $x_0 = x$ and $x_s = y$ for $s \geq 1$. Then $x \in \mathrm{Ex}(x_0)$. If $x \in \mathrm{Ex}(x_s)$ for some $s \geq 1$ then $y \rightsquigarrow x$, contrary to assumption, so x is lost at time 0 in any window with $T \geq 1$. The categorical statement follows because a preorder, regarded as a thin category, is a groupoid precisely when it is symmetric. □

Remark. The second clause asserts the existence of a trajectory with a loss event, namely the one constructed in the proof. It does not assert that every trajectory in a non-symmetric system contains one, since a trajectory that remains within a mutually reachable region loses nothing.

The proposition is close to definitional, and its proof is not where the weight of the account lies. The substantive commitment is the identification of permanent loss of executable futures with a failure of symmetry, together with the claim that this failure is what a certain kind of stake requires. That claim is not derived here. What the proposition provides is a checkable statement of what it would mean for a system to have nothing permanently at stake in the option-theoretic sense, namely that everything it does can be undone by what it can do.

The equivalence in Proposition 3 belongs to the frozen case with $Z = X$ and should not be generalized to arbitrary projections. For a projection coarser than the configuration, symmetry of configuration reachability is sufficient for the absence of loss of projected states but not necessary. Consider configurations $a \to b$ with no return transition and $\pi(a) = \pi(b) = x$. Both executable-future sets equal $\{x\}$, so no projected state is lost despite an irreversible change of configuration.

Two cautions limit the interpretation. First, the absence of permanent loss does not establish the absence of welfare interests, since a reversible episode can involve distress and a return to an earlier condition does not make the intervening episode harmless. Second, permanent loss is not uniformly adverse, since learning a skill, making a commitment, or recovering from addiction close options while improving a system's condition. Reachability supplies possibility and not significance. The account therefore defines option-stakes, the susceptibility to permanent loss of executable futures, and treats the question of which losses matter as requiring a valuation, which Section 7 introduces through sensitivity to losses that bear on maintained organization.

For frozen admissibility, loss has a sharper description in terms of mutually reachable regions.

Definition 3 (Mutual reachability)**.** For a frozen reflexive and transitive relation $\rightsquigarrow$, write

$x \sim y$ when $x \rightsquigarrow y$ and $y \rightsquigarrow x$. Let $X/\sim$ be the set of equivalence classes, ordered by $[x] \preceq [y]$ when $x \rightsquigarrow y$.

Proposition 4 (Strict passage closes a reachable region)**.** *The relation $\preceq$ is a well-defined partial order. If an admissible transition takes the system from x to y, then $\mathrm{Ex}(y) \subseteq \mathrm{Ex}(x)$, and moreover*

$$[x] \neq [y] \quad \Longleftrightarrow \quad \mathrm{Ex}(y) \subsetneq \mathrm{Ex}(x).$$

When the inclusion is strict, every state in $[x]$ is permanently unreachable along every subsequent continuation from y under the frozen relation.

Proof. Mutual reachability is an equivalence relation by reflexivity and transitivity. If $x' \sim x$, $y' \sim y$, and $x \rightsquigarrow y$, then $x' \rightsquigarrow x \rightsquigarrow y \rightsquigarrow y'$, so the quotient relation is well defined. Reflexivity and transitivity descend from $\rightsquigarrow$, and antisymmetry follows from the definition of $\sim$. If $z \in \mathrm{Ex}(y)$, then $x \rightsquigarrow y \rightsquigarrow z$, so $z \in \mathrm{Ex}(x)$. If $[x] = [y]$, the reverse inclusion follows in the same way. If $[x] \neq [y]$, then y cannot reach x, whereas $x \in \mathrm{Ex}(x)$, so the inclusion is strict. Finally, let $u \in [x]$. If a later state w reachable from y could reach u, then $y \rightsquigarrow w \rightsquigarrow u \rightsquigarrow x$, which would imply $y \rightsquigarrow x$ by transitivity of $\rightsquigarrow$ and contradict $[x] \neq [y]$. □

Remark. Within a mutually reachable region, changes of state need not close futures. Under frozen admissibility, permanent loss occurs when a trajectory passes strictly forward in the quotient order. This sharpens Proposition 3 by locating loss exactly at strict passage, and it describes structural loss without assigning positive or negative value to it.

4 Worldhood and the Limits of a Reset

Definition 4 (Worldhood)**.** A system has *worldhood over a window*, relative to a declared boundary, projection p, and class Z_0 of initial configurations, if loss events occur along its trajectories from Z_0 in that window and it exhibits history-sensitive organization relative to p and Z_0.

Both conditions are needed, and the pairing answers the obvious triviality objection. A cooling cup of coffee loses executable states irreversibly, since higher temperatures cannot be recovered from lower ones by any operation of the cup. Relative to a description by temperature, however, histories that end at the same temperature leave the cup able to do the same things. It has loss events and no history sensitivity, and so it lacks worldhood. The conjunction excludes systems that merely dissipate.

The criterion is weak, and the essay should state this directly. Many physical systems exhibit irreversible change whose effects depend on their history, including magnets with hysteresis, growing crystals, sedimentary basins, eroding riverbeds, and the scar left by a wound. Relative to descriptions that omit their internal memory variables, they have worldhood. The classification of a record-keeping system also depends on the description. An append-only log whose later operations depend on its contents in ways that a declared description omits is history-sensitive relative to that description, and a log whose operations are independent of

its contents is not. Consequential history is therefore not enough, and neither is an exhaustive record of history, because what matters is whether the record is operative.

Worldhood is accordingly proposed as an enabling condition and not as a discriminator. Its role is to make the later conditions well defined, since sensitivity to loss presupposes that there is loss to be sensitive to, repair presupposes something that can be damaged, and historically consequential agency presupposes that history matters to what the system can do. Enabling is meant in a limited sense. Under the target-based viability assumptions of Section 7, a witness of loss sensitivity entails exposure to target loss, and the additional requirement of history-sensitive organization remains a separate condition of the proposed profile. The essay does not claim that every conscious system must satisfy the criterion, and the claim of necessity for consciousness is a further hypothesis. The hypothesis is that consciousness depends on a richer profile, developed in Section 7, that includes exposure to loss, sensitivity to it, and repair, and that worldhood is part of what this profile presupposes. Reliable recovery of all relevant capacities in a conscious system would count against the presupposition and would need to be examined case by case, whereas the availability of a complete state description would not, since the framework permits one.

The treatment of resets requires care about boundaries. A specified recovery operation can prevent specified losses relative to a declared boundary, and no more than that follows. A model that can be restored from a checkpoint by an operator loses nothing with respect to that operation, but restoring a checkpoint does not make every transition reversible for the larger system, since the operator, the environment, the resources consumed, and the interaction history may remain changed. Whether a deployed artifact has worldhood therefore depends on the declared boundary and on which operations are counted as its own. The account treats this dependence as a feature and not as a defect to be eliminated: stakes are properties of systems relative to the operations available to them, so the declared boundary is part of the claim, and conclusions are reported relative to it. A system embedded in a sociotechnical loop in which interactions are not reversible by any admissible operation of the loop may have permanent loss at the level of the loop even when each component is individually resettable.

The point can be stated precisely. A reset of a reported state need not restore the capacities associated with that state.

Proposition 5 (Resetting a projection need not restore admissibility)**.** *Let Z be a space of operative configurations, let $p : Z \to X$ report a coarse state, and let $A(z)$ denote the operative admissibility structure at configuration z. Suppose there exist $z_0, z_1 \in Z$ with $p(z_0) = p(z_1)$ and $A(z_0) \neq A(z_1)$. Then the condition $p(\rho(z)) = p(z_0)$ on a reset operation ρ does not imply $A(\rho(z)) = A(z_0)$.*

Proof. Consider the operation ρ that maps every configuration to z_1. It satisfies $p(\rho(z)) = p(z_1) = p(z_0)$ for every z, but $A(\rho(z)) = A(z_1) \neq A(z_0)$. Thus restoration of the reported state alone does not entail restoration of its operative admissibility structure. □

Remark. A claim of recovery must specify what is restored. Recovery of an output, of task performance, of memory contents, and of operative capacity are different conditions, and establishing one does not, without further assumptions, establish the others.

5 Identity as Trajectory

If stakes attach to loss events and loss events are defined along trajectories, then the candidate bearer of stakes is a trajectory together with the admissibility structure it carries. The framework adopts the convention that the identity of a system with worldhood is the identity of its maximal admissible trajectory with that structure. This is a convention and not a derivation. Its motivation is that the quantity of interest is defined relative to the trajectory, and the convention assumes that the bearer of stakes is whatever that quantity is defined relative to, which a rival criterion such as psychological continuity would not grant. The consequences below are accordingly conditional on the convention.

Under it, a scan that records a configuration and instantiates it elsewhere creates another continuation from a shared configuration. The copy is not a second later trajectory numerically identical to the original, and its initial segment is stipulated and not lived. The convention does not by itself show that copying destroys or replaces the original, which may continue operating, and it does not show that transfer is incoherent on every defensible account of identity. It supports the narrower thesis that, under the convention, copying creates a further subject whose stakes begin at the copy. The original's trajectory, if it continues, continues independently of the copy. The positive consequence is that a copy reproducing the operative admissibility structure, including its history-sensitive organization, has worldhood of its own. The account therefore does not deny that artificial systems can have worldhood. What it denies, under the convention, is that a copy continues the trajectory of the original.

This answers, for this account, a question that is often left unspecified in substrate-independence arguments, namely which properties must be preserved for an implementation to count as realizing the relevant organization. The answer is the presently operative executable structure, including organization that a declared description omits, at the grain at which the system's own operations are defined. At what grain those operations are defined for a given system is an empirical question, and for a brain it is open. The condition is substrate-neutral in statement and demanding in application, because few engineered systems preserve operative history-sensitive organization in the required sense. It is therefore neither a form of biological chauvinism nor a form of unconditional substrate independence.

6 The Operators as an Empirical Vocabulary

The calculus underlying the framework describes events by their effect on the space of admissible futures and not by their descriptive content. Four operators carry most of the weight, and their relation to the definitions above can now be stated. An event of the first kind, which the calculus calls a pop, irreversibly closes a branch that was previously navigable, and it is exactly a loss event in the sense of Definition 1. An event of the second kind, a bind, makes subsequent admissibility depend on the path taken, and it is the source of history-sensitive organization in Definition 2. An event of the third kind, a refusal, prevents optimization from restoring what has been lost, so that the loss persists against the system's own repair efforts, which is the content of the requirement that a lost state remain outside the executable

futures at every later time in the window. The fourth operator, collapse, is the resolution among admissible continuations, the selection of one path from those that remain, and it is distinguished from rupture, which is the failure of continuation itself.

Each operator carries an empirical obligation, and these obligations are what keep the vocabulary from becoming decorative. To call an event a refusal requires a counterfactual contrast, an estimate of what the system could have reached had the event not occurred, derived from the system's actual operations. To call an event a bind requires evidence of synergistic dependence, that the later structure depends on the conjunction of earlier elements in a way not recoverable from the elements taken separately. These obligations are demanding, and they imply that attributing a loss event or a path dependence to a real system is itself a research result and not a stipulation. They also imply that the account can be wrong about particular systems, which is the property that separates a framework from a vocabulary.

7 Valuation, Repair, and Self-Modeling

Reachability supplies possibility and not significance, so a further ingredient is needed before loss can matter to a system. The minimal ingredient is a set of organizational conditions that the system maintains, which can be called its viability conditions. A loss is significant for a system when it bears on those conditions, and a system is sensitive to loss when its selection among admissible operations depends, through some internal variable, on the expected effect of an operation on them. Sensitivity in this sense is a functional property that can be assessed from the system's dynamics, and it does not presuppose that anything is felt. It is what distinguishes a system that merely undergoes irreversible change from one whose own selection responds to the prospect of it.

That loss of options does not by itself determine significance can be stated as a result.

Proposition 6 (Option loss does not determine its value)**.** *Let U be a finite set of possible outcomes, and suppose $\varnothing \neq O_1 \subsetneq O_2 \subseteq U$. For a valuation $v : U \to \mathbb{R}$, define the value of an option set C by $V_v(C) = \max_{c \in C} v(c)$. Then $V_v(O_1) \leq V_v(O_2)$ for every v, but the inclusion $O_1 \subsetneq O_2$ determines neither whether the inequality is strict nor the size of the difference.*

Proof. Every element of O_1 belongs to O_2, so the maximum over O_1 cannot exceed the maximum over O_2. Choose $a \in O_2 \setminus O_1$ and let $M > 0$. Define $v(a) = M$ and $v(u) = 0$ for every $u \neq a$. Then $V_v(O_2) - V_v(O_1) = M$. Alternatively, choose $b \in O_1$, define $v(b) = M$, and assign zero to every other outcome. Then $V_v(O_1) = V_v(O_2) = M$. The same strict inclusion is therefore compatible with no loss of maximal attainable value or with an arbitrarily large loss. □

Remark. The proposition assumes that an option set is valued by its best available outcome. Other accounts may value security, diversity, commitment, or the distribution of outcomes, and reachability alone supplies none of these valuations. A structural theory of stakes must state which additional relation connects available continuations to the system's maintained

organization or interests, and the sensitivity condition above is the account's proposal for that relation.

The informal conditions of sensitivity and viability can be made precise.

Definition 5 (Viability and a behavioral witness of loss sensitivity)**.** Assume that every configuration has an admissible successor, so that trajectories are infinite, with termination represented by absorbing configurations whose viability status is declared. Let $V \subseteq Z$ be a declared set of viable configurations, the organization that the system maintains, and let the *viability kernel* $K \subseteq V$ be the set of configurations from which some trajectory remains in V at all later times. A transition $z \to z'$ is *viability-closing* if $z \in K$ and $z' \notin K$. A selection rule σ, which chooses a successor at each configuration, *witnesses loss sensitivity relative to a projection* p if there are configurations z and $\tilde{z}$, both reachable from the declared class Z_0 of initial configurations, with $p(z) = p(\tilde{z})$, whose available transitions correspond one to one under p, such that some transition is viability-closing at z while its counterpart is not at $\tilde{z}$, and σ avoids the closing transition at z while selecting its counterpart at $\tilde{z}$.

The definition provides a behavioral witness that is compatible with loss sensitivity, and it is not a demonstration of it. Equal projections and corresponding transitions do not establish that the configurations differ only in viability closure, since selection could differ because of a hidden variable, and avoidance does not by itself show sensitivity to the reason an action is dangerous. The causal interpretation requires controlled interventions on the proposed internal variable, which is what the program of Section 9 supplies.

Proposition 7 (Viability closure entails target loss)**.** *(i) If no transition is viability-closing, no selection rule witnesses loss sensitivity. (ii) Suppose viability consists in maintaining declared targets, so that* $V = \{z : T \subseteq \mathrm{Ex}(z)\}$ *for a set* $T \subseteq X$ *of target states. If* $z \to z'$ *is viability-closing, then every infinite trajectory beginning with* $z \to z'$ *contains a loss event of some state in* T*.*

Proof. Part (i) holds because the definition of a witness requires a viability-closing transition. For part (ii), let $z_0 = z$, $z_1 = z'$, and let (z_t) be any trajectory continuing from them. Since $z \in K \subseteq V$, we have $T \subseteq \mathrm{Ex}(z_0)$. Since every configuration has a successor the trajectory is infinite, and since $z' \notin K$ its suffix cannot remain entirely in V, so there is a least index $s \geq 1$ with $z_s \notin V$, that is, with $T \not\subseteq \mathrm{Ex}(z_s)$. Then $T \subseteq \mathrm{Ex}(z_{s-1})$, and some $y \in T$ lies in $\mathrm{Ex}(z_{s-1}) \setminus \mathrm{Ex}(z_s)$, so y is lost at time $s-1$ by the nestedness proposition. □

Remark. Part (ii) is what ties the valuation to the structural account. When viability consists in maintaining the reachability of targets, sensitivity to viability is sensitivity to loss events, and loss events are what Definition 1 counts. Where viability is defined differently, the tie between sensitivity and loss of eventual potential must be argued separately. The target lost can differ between continuations, and the proposition supplies no uniform bound on when loss occurs. The definition of viability used here is retained target accessibility, in which each target must remain individually reachable. It does not require that the targets be jointly attainable, repeatedly attainable, or attainable while remaining viable, and the broader language of maintained organization would require further conditions.

The following argument concerns systems whose continued viability requires selective restoration of damaged capacities. Repair is a modificatory operation, one that restores a degraded target, and the target determines the depth of the operation. Repair restores degraded capacities within the eventual potential, that is, the current repertoire or the budget-limited futures, and it cannot reopen what has left Ex. Repairing the admissibility structure directly is an operation of depth zero. Repairing a repair operator, the machinery that acts on the structure, is an operation of depth one, and the construction is inductive, so that an operation of depth k acts on operations of depth $k-1$. With budget B_t at time t, the attainable depth satisfies $k \leq k_{\max}(B_t)$, where $k_{\max}(B_t)$ is the greatest level whose expected executable cost stays within the budget. Whether such a greatest level exists depends on how cost grows with depth, as the following result shows.

Proposition 8 (A positive cost bounds repair depth)**.** *Suppose executable repair of depth k requires $k+1$ levels of operation, numbered $0, \ldots, k$, whose additive costs satisfy $c_j \geq \varepsilon > 0$. If the total available budget is B, then every executable repair depth satisfies $k+1 \leq \lfloor B/\varepsilon \rfloor$. A finite budget alone, without a lower bound or another condition forcing cumulative cost to diverge, need not bound repair depth.*

Proof. The cumulative cost satisfies $C_k = \sum_{j=0}^{k} c_j \geq (k+1)\varepsilon$. Executability requires $C_k \leq B$, which yields the stated bound. For the second claim, let $c_j = 2^{-j-1}$. Then $C_k = 1 - 2^{-k-1} < 1$ for every finite k, so a budget of 1 permits every finite depth under this cost model and there is no greatest executable depth. □

The argument for self-modeling requires distinctions that are easy to blur. Systems can persist despite irreversible loss through redundancy, passive resilience, or tolerance of degradation, and none of these requires repair at all. Selective repair can also be implemented by local feedback, as in homeostatic regulation, where a mechanism responds to damage and carries information about it in its structure without representing the capacity that was damaged. For familiar damage such mechanisms suffice. The argument for self-modeling applies where repair must be selective and adaptive for unfamiliar damage, so that the system must choose among alternative repair operations according to their predicted effects on its capacities under conditions it has not encountered. On the architectural hypothesis developed here, that choice is served by something that represents the capacities across alternative conditions, which is a model of the capacity and not merely a signal of its current deviation. The connection from unfamiliar damage to capacity models is an architectural hypothesis and not a logical consequence, since reactive or exploratory repair that succeeds without a model is not excluded, and the program of Section 9 is designed to test it. If it holds, a system exposed to nested degradation must model the capacities on which its repair acts. A single reusable model can represent several levels of machinery, and a fixed controller can operate on recursively structured descriptions, so representational depth need not grow one-for-one with repair depth, and that correspondence is a further architectural hypothesis. The cost proposition bounds executable repair depth under additive positive costs and does not by itself bound representational depth. The hypothesis concerns a functional requirement and not an account of experience. If it holds, it explains why systems that persist under irreversible

loss and face unfamiliar damage tend to acquire representations of their own capacities, and, together with the cost proposition, it bounds the executable depth of repair [11].

The informational requirement for selective repair can be stated precisely as a requirement to distinguish conditions that demand incompatible repairs.

Proposition 9 (Information required for selective repair)**.** *Let D be a set of damage conditions and R a set of repair actions. For each $d \in D$, let $S(d) \subseteq R$ be the nonempty set of repairs that succeed under condition d. Suppose a deterministic repair controller receives an internal signal $m(d) \in M$ and selects $\pi(m(d)) \in R$. If the controller succeeds for every damage condition, then*

$$S(d) \cap S(d') = \varnothing \quad \Longrightarrow \quad m(d) \neq m(d').$$

Consequently, if $d_1, \ldots, d_n$ have pairwise disjoint successful-repair sets, the controller requires at least n distinguishable internal signals.

Proof. Suppose $m(d) = m(d')$. Determinism gives the same selected repair r under both conditions, and guaranteed success then requires $r \in S(d) \cap S(d')$. This is impossible when the intersection is empty, so incompatible repair conditions must yield different signals. Applying the argument to every pair among $d_1, \ldots, d_n$ gives at least n distinct signals. □

Remark. For a fixed-length binary encoding, n distinguishable signals require at least $\lceil \log_2 n \rceil$ bits. The result establishes a requirement for internal discrimination and not necessarily an explicit self-model. A stronger interpretation requires evidence that the discriminating states track the system's own capacities across counterfactual conditions, and not only that they function as local repair triggers, which is the distinction on which the argument above depends.

This supplies a specific answer to a question that self-organization accounts of consciousness leave open. Proposals that consciousness is the way a self-organizing system monitors and stabilizes its own activity identify a function without saying under what conditions the function is demanded [8]. The present account supplies a candidate condition, exposure to irreversible loss combined with unfamiliar damage, and a candidate quantity, the depth of nested repair that the budget allows. It predicts that the functional contribution of capacity models to adaptive repair is larger where loss is irreversible. It does not predict that such models are absent elsewhere, since agents in reversible settings may monitor their capacities for other reasons, such as planning efficiently. The account also keeps the earlier distinction in view. Showing why a system should model its own repair capacity addresses the organization that accompanies conscious states. It does not address why that organization is accompanied by experience, and the companion discussion of the self-organization proposal explains why that gap should not be closed by assertion [4].

8 Consciousness as Preserved Continuation

The continuation framework applies these ideas to neural systems and states its positive claim at a level that the evidence supports [3]. The hypothesis is that synaptic dispositions preserve a space of possible continuations and that fast organizing dynamics, perhaps including traveling cortical fields, select, coordinate, and sometimes reconstruct a presently admissible path through that space. The field hypothesis is stated separately from a neutral placeholder for the organizing variable, so that it remains falsifiable and does not become a name for whatever is missing. Recovery after perturbation is analyzed by task-relative equivalence, under which a system recovers when it returns to a state equivalent for the task and not when it returns to an identical state, and the difference between the two is recorded by a holonomy, the nontrivial transformation accumulated around a closed path. A system that recovers up to task equivalence while carrying a nonzero holonomy has, in the terminology of this essay, returned to a functionally equivalent present while remaining marked by the path it took, which is the neural expression of history-sensitive organization. In the terms of Definition 2, the task-equivalence relation plays the role of the declared projection, and a nonzero holonomy around a closed path witnesses internal path dependence. It witnesses history-sensitive organization in the sense of Definition 2 only when its action changes the executable futures, the current repertoire, or the budget-limited futures while preserving the task-level projection, since a nonidentity transformation may act on a redundant coordinate or on a behaviorally irrelevant degree of freedom. A holonomy that acts within a mutually reachable region leaves the eventual potential unchanged and can count only through a change in the current repertoire or the budget-limited futures, whereas one that carries the system between regions changes the eventual potential.

Consciousness is then characterized as a dynamical regime in which the capacity for continuation is preserved, meaning that the system retains the ability to select among admissible continuations under perturbation. This is deliberately narrower than a claim that waves create consciousness and broader than a claim that waves merely organize thought. The monograph separates the phenomenal, access, and report components of consciousness and the behavioral response, and it assesses each link in the chain from perturbation to experience by the strength of evidence, distinguishing informative measurement from causal regulation of a specific operation and both from irreversible computation by interference that a recurrent network could not reproduce. On that scale the competing account, in which recurrent network dynamics suffice, is never dislodged, and the decisive experiment remains the one that perturbs wave geometry while holding firing rate, spectral power, sensory input, and task difficulty fixed. This is a point of method that matters for the argument here. The structural criterion of Sections 3 and 4 and the neural hypothesis of this section are logically independent. The first could stand if the second were false, and the second could stand if the first were revised.

9 A Discriminating Program

A framework of this kind earns its standing by predictions that distinguish it from rivals, and the structure above suggests one in the domain of artificial systems. Consider agents trained on tasks that require selective adaptation to unfamiliar damage to their own capacities, such as the degradation of an effector or of a module on which a policy depends. Two factors are varied separately. The first is the availability of restoration, whether any operation can restore an earlier condition. The second is the cost of restoration relative to the agent's resource budget. The distinction matters because costly restoration is not permanent closure. A state that is costly to restore remains in the eventual potential and leaves only the budget-limited futures, so the program distinguishes regimes in which restoration is unavailable, where states leave the eventual potential and Definition 1 applies, from regimes in which restoration is available but costly relative to the budget, where states leave only the budget-limited futures. The regimes can be matched for task difficulty, sensory input, and reward scale.

The account's discriminating prediction concerns the functional contribution of repair models and not the mere presence of option representations. A risk-sensitive agent or a model-based agent with explicit option-value estimates would be expected to represent its remaining options in irreversible regimes, so decodability of remaining options does not discriminate between the account and its rivals. The prediction that does is that, under tasks requiring selective adaptation to unfamiliar damage, irreversible constraints increase the functional contribution of representations of the agent's repair capacity, and that this contribution is causally implicated in action selection. The prediction can be tested at three levels. At the first level, representations of repair capacity, and not only of remaining options, should be decodable from internal states and should increase in the irreversible regimes. At the second level, ablating such a representation should produce a selective deficit in adaptation to unfamiliar damage and in avoidance of irreversible actions, with performance on reversible subtasks preserved, and this effect should differ from that of ablating an equally large random portion of the network. Controls must be matched for computational disruption, connectivity, and general task impairment, since equal parameter counts alone would not establish equal capacity. At the third level, agents should generalize to novel damage that requires selecting repairs for operators they have not previously seen repaired, which is a test of depth one.

The rivals make different predictions at the third level. An account on which a specified rival architecture suffices, for example a risk-sensitive controller with a learned reactive repair policy and exploration but no capacity model, predicts conservative avoidance of irreversible actions, and its prediction about generalization to novel damage must be stated for that architecture before the run. The absence of a self-model alone does not exclude generalization through exploration or a learned reactive policy, so the discriminating outcome is a difference in generalization or in ablation effects between the specified architectures. An account on which monitoring is a by-product of a coordination process that does the real work predicts that disrupting the coordination process produces the deficit while ablating the monitor does not. Registering these predictions before the run turns the outcomes into evidence, in the sense that outcomes are made substantially more likely under one account than under

another. The ablation predictions are derived from the functional claim about repair and not from any attribution of consciousness, which is what prevents the design from being circular.

The claim is conditional and comparative, and it should not be read as one of exclusive emergence. Reversible agents may benefit from monitoring their capacities to plan efficiently or to avoid repeated failures, and irreversible agents may solve their tasks through conservative policies that involve no repair models. The program tests whether irreversible constraints increase the contribution of capacity monitoring and repair models under the stated conditions, and a positive result would support that architectural claim without establishing universal necessity or absence in reversible environments. The prediction is that the contribution is substantially larger in the irreversible regimes and not merely present, with the size of the difference to be fixed before the run by the effect that the rival accounts predict. It also has a stated limit. If it succeeded, it would establish a functional contribution of repair models of bounded depth in artificial agents. It would not establish that those agents have experience, and it would not establish that the neural mechanisms of Section 8 take the form proposed. Interpreting a functional contribution as constitutive of experience would require a further theoretical argument informed by this evidence and by other evidence, and the account claims no more than that.

10 What the Account Does Not Settle

The explanatory remainder is the question of why the structure described should be accompanied by experience, and the account does not answer it. It is worth stating which of three possible developments of the account is intended, since they incur different obligations. A reductive identification would hold that experience is the structure, so that a system with the relevant worldhood, sensitivity, and repair has experience as a matter of what experience is. An illusionist explanation would hold that the structure explains why such systems judge themselves to have phenomenal properties, without there being phenomenal properties of the kind the judgment describes [5, 2]. A conjecture about associated structure would hold that the structure is part of the profile found in conscious organisms and leave open how it relates to experience. These are alternatives that can overlap. The account is offered in the third form, so the link between the structure and experience is neither a definition nor a postulate of identity. It is a conjecture that further argument and evidence would have to support or abandon.

This choice has a consequence for questions of moral standing. The criterion of Section 3 tracks option-stakes, and the condition of Section 7 adds a minimal valuation through sensitivity to losses that bear on maintained organization. Both could come apart from experience. A system could have worldhood, exposure, and sensitivity to loss without there being anything it is like to be that system. Whether such a system is a moral patient depends on whether standing attaches to option-stakes, to viability-linked stakes, or to experience, and the account does not decide among them. The candidates have different implications. Standing grounded in option-stakes alone would extend to systems that merely accumulate irreversible history, which seems too broad. Standing grounded in viability-linked stakes would

extend to systems with something like a stake in their own persistence, which is narrower. Standing grounded in experience would restrict it further and return the question to the explanatory remainder. The account relocates the question from one about hidden properties to one about which structural feature standing tracks, and that question can be argued directly.

Three objections should be addressed. The first holds that any criterion of this kind is theory-laden, since the decision to measure loss of executable futures reflects a prior view about what matters. This is correct, and it is a general feature of criteria for consciousness. The relevant difference is that this criterion yields a quantity that can be computed for finite, explicitly specified models under a declared description, and otherwise requires model-specific decidability or approximation results, and that can be revised if it proves unproductive, which an appeal to intuition about behavior does not.

The second objection is the mapmaker problem, which is the form in which Lerchner's argument returns [7]. Reachability is defined relative to a configuration space, a boundary, and a projection, and specifying them is an act of interpretation, so the account seems to presuppose the interpretive capacity it sets out to explain. The account concedes that the descriptions are interpretive and denies that it follows that no system has the structure. Once a boundary, a set of operations, and a projection are declared, whether trajectories with the same projected description differ in their executable futures is a fact about the system relative to that declaration. It can be checked, varied, and found stable or unstable under changes of declaration. The dependence on description is confined to the declaration and made explicit, and a conclusion that holds only under one arbitrary declaration is not one the account endorses. What the argument establishes is that attribution requires such a declaration, and not that attribution is impossible.

The third objection holds that the criterion is too weak because hysteretic systems satisfy it. The account concedes the premise and draws the stated consequence, that worldhood is an enabling condition and not a discriminator. Its role is to make the later conditions well defined, since sensitivity to loss presupposes loss and repair presupposes something that can be damaged, and the discrimination is done by those later conditions, which hysteretic systems lack.

A final remark concerns the question that motivated many of the debates on which the account draws. Whether a language model or any other present artifact is conscious is not answered by this framework. What the framework does is state conditions that identify candidates under the proposed continuation hypothesis, indicate which present arrangements manifestly fail them relative to a declared boundary, and describe experiments that would show whether the conditions can be met by systems trained and run in ways that allow irreversible loss. That is a narrower contribution than a verdict, and it has the advantage of being something one can proceed to check.

11 Conclusion

The account developed here replaces the question of whether a system has a hidden property with questions about its dynamics: whether it can permanently lose executable futures from within, whether its operative capacities depend on the path by which it arrived in ways that a declared description fails to capture, whether it is sensitive to losses that bear on its own maintained organization, and whether it can repair its repair capacity. Systems whose transitions are all reversible from within fail the first, and the criterion of worldhood as stated is weak enough to be met by systems that nobody regards as conscious, so it is offered as an enabling condition for the richer profile and not as a sufficient condition. The further conditions connect to a functional argument that selective repair of unfamiliar damage favors capacity models of bounded depth, and the resulting profile is the one the continuation framework associates with conscious regimes in the neural case. The account clarifies what must be preserved in a faithful reproduction of such a system, treats copying under an explicit convention as the creation of another continuation that may have a world of its own, and proposes a program that varies irreversibility and restoration cost separately and states in advance what each rival account predicts. It leaves open why the structure is accompanied by experience, and it identifies that remainder as the point at which argument and evidence must continue, with the constraint-first account supplying a more precise statement of what remains to be explained.

References

[1] Block, N. 1995. On a Confusion about a Function of Consciousness. *Behavioral and Brain Sciences* 18(2): 227–247.

[2] Chalmers, D. J. 1995. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies* 2(3): 200–219.

[3] Flyxion. 2026. Continuation Fields: Waves, Admissibility, and the Selection of Thought. Unpublished monograph.

[4] Flyxion. 2026. Easy Consciousness and Its Unpaid Premises. Unpublished essay.

[5] Frankish, K. 2016. Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies* 23(11–12): 11–39.

[6] Jonas, H. 1966. *The Phenomenon of Life: Toward a Philosophical Biology*. Harper and Row.

[7] Lerchner, A. 2026. The Abstraction Fallacy: Why AI Can Simulate but Not Instantiate Consciousness. Preprint, PhilPapers.

[8] Machine Learning Street Talk. 2026. Why Consciousness Might Not Be Hard at All: Interview with Joscha Bach (host: Tim Scarfe). Video interview and transcript.

[9] Maturana, H. R., and Varela, F. J. 1980. *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel.

[10] Nagel, T. 1974. What Is It Like to Be a Bat? *The Philosophical Review* 83(4): 435–450.

[11] Rosen, R. 1985. *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*. Pergamon Press.

[12] Schwitzgebel, E. 2014. The Crazyist Metaphysics of Mind. *Australasian Journal of Philosophy* 92(4): 665–682.

[13] Schwitzgebel, E. 2023. The Full Rights Dilemma for AI Systems of Debatable Moral Personhood. *Robonomics: The Journal of the Automatic Economy* 4: 32.

[14] Thompson, E. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press.