Throwing the Game: Refusal, Event-Driven Cognition, and the Survival of Value Beyond Autoregressive Intelligence

Flyxion

December 20, 2025
Revised: April 19, 2026

Abstract

Recent advances in autoregressive artificial intelligence suggest that cognition may be reducible to the extension of statistical regularities over sequences. On this view, intelligence consists in the compression and continuation of history, with no need for explicit world models, commitments, or irreversibility. This paper argues that such accounts fail to capture a structurally distinct class of cognitive acts: refusals.

A refusal is not a preference, constraint, or optimization tradeoff. It is an event that eliminates admissible futures and binds an agent to a history. Where autoregressive systems preserve optionality, refusal spends it. To make this precise, the paper introduces Spherepop, a minimal event calculus in which agents operate over branching future spaces via irreversible operators such as **Refuse**, **Pop**, **Bind**, and **Collapse**.

Within this framework, we prove a formal impossibility result: refusal cannot be represented as utility maximization over a fixed triple $(\mathcal{A}, U, \pi)$ without encoding event-history as a primitive state variable. This requires explicitly distinguishing the representing utility U from the agent's recognized instrumental evaluation functional V—a distinction that standard decision theory collapses. We further prove that **Collapse** is the unique admissibility-expanding operator in the calculus, and show that its idempotence ($\mathcal{C} \circ \mathcal{C} = \mathcal{C}$) connects Spherepop to the theory of idempotent monads and Karoubi envelopes. Revocation is shown not to be the inverse of refusal but to require a **Collapse** operation that destroys historical differentiation, and "fake refusal" is characterized as behavior that preserves a collapse path restoring eliminated futures.

An information-theoretic analysis sharpens a key distinction: refusal is not Bayesian updating (conditioning on observation) but an endogenous modification of the sample space itself—a deliberate reduction of future entropy whose cost is conserved as constraint rather than dissipated as error. We identify three equivalent characterizations of the same underlying phenomenon—entropy reduction, future-space pruning, and accumulation of non-commutativity—and show that worldhood is the obstruction to commutativity, the structure that cannot be rearranged. Active inference is reinterpreted: refusal is incompatible with free-energy minimization unless the generative model includes commitment variables as latent structure. We introduce a stabilization-based halting criterion grounded in the exhaustion of world-relevant distinctions and relate the event algebra to IIT, assembly theory, and the phenomenology of commitment.

The central claim is that intelligence is not exhausted by prediction. It also consists in the capacity to close worlds. Refusal is one of the events by which such closure becomes possible.

1 From Sequence to Event

The unifying theme of this paper is a shift from sequence-based to event-based cognition. Autoregressive systems model the extension of histories; Spherepop models the transformation of futures. The distinction is not merely technical. It marks the difference between systems that continue and systems that commit.

What follows should be read as an attempt to articulate this distinction at multiple levels: formal, informational, social, and philosophical. The formal core appears in the non-representability result of Section 7 and the algebraic structure of Section 16; the remaining sections extend these results into information theory, competition theory, active inference, and the philosophy of mind. A synthesizing observation—that entropy reduction, future-space pruning, and non-commutativity accumulation are equivalent characterizations of a single phenomenon called closure—is developed across these sections and consolidated in the worldhood analysis of Section 11.

2 Introduction

Recent advances in generative artificial intelligence have challenged many assumptions that structured twentieth-century cognitive theory. Large autoregressive models—whether applied to language, images, audio, or video—now produce behavior that appears coherent, context-sensitive, and purposive without relying on explicit symbolic representations of world structure, causal relations, or articulated goals. These systems generate fluent dialogue, physically plausible motion, and socially legible action by learning to extend statistical regularities in data.

The success of such systems has inspired a strong thesis: cognition may fundamentally consist in deep autoregressive generation over learned sequences. On this view, intelligence is not grounded in explicit internal world models but in the ability to capture long-range dependencies that encode latent structure. Memory exists to compensate for partial observability, and cognitive power scales with the depth and richness of temporal context an agent can maintain. In the limit, the argument goes, the ability to compress and extend sequential regularities is sufficient for much of what we ordinarily treat as world understanding [1].

Yet this picture leaves a striking class of human behaviors unexplained. Agents frequently and deliberately choose actions that knowingly degrade performance, restrict future options, or incur irreversible loss. Whistleblowers sacrifice careers to preserve integrity; artists reject lucrative contracts to protect aesthetic autonomy; individuals refuse opportunities that would increase power at the expense of relationships. These actions are not mistakes, nor products of ignorance, nor mere cases of bounded rationality. They are refusals: deliberate acts of irreversible commitment undertaken with full awareness of their cost.

This paper argues that refusal constitutes a fundamental boundary for sequence-based accounts of cognition. Refusal is not a preference, constraint, or auxiliary objective layered onto optimization. It is an *event*: an irreversible operation that eliminates admissible futures and binds the agent to a specific history. While autoregressive systems preserve optionality, refusal spends it. While sequence learning generates trajectories through possibility space, refusal generates worlds in which certain paths are no longer available.

Methodologically, the argument is hybrid. It combines conceptual analysis (especially in philosophy of action and commitment) with formal modeling in an event calculus. The event calculus is Spherepop: a small set of irreversible operators that act directly on branching future spaces. Formal results are used primarily as constraints on interpretation. In particular, we prove a non-representability theorem: refusal cannot be represented as ordinary utility maximization on a fixed triple $(\mathcal{A}, U, \pi)$ without augmenting the state space with event-history. A crucial component of this proof is the distinction between the *representing* utility $U : \mathcal{O} \to \mathbb{R}$ and the *instrumental evaluation functional* V the agent continues to recognize—a distinction that standard decision theory collapses, and whose separation is required to make the impossibility result precise. This formal result underpins the later claims about "strategy stealing" and the competitive disadvantage of value-sustaining agents [2].

The argument proceeds in stages. First, we introduce an illustrative case in which refusal is unusually clean: deliberate underperformance in the service of relationship (§3). Second, we characterize the autoregressive baseline and identify its structural bias toward optionality preservation (§4). Third, we develop Spherepop as a calculus with syntax, operational semantics, and worked examples (§5). Fourth, we distinguish refusal from preference change and related phenomena (§6), then prove the non-representability result (§7). We then develop information-theoretic, competitive, worldhood, collapse, and halting analyses, followed by a treatment of the algebraic structure of the event calculus and its relationship to adjacent frameworks including integrated information theory, assembly theory, and active inference. We close with objections and open problems.

3 An Illustrative Case: Throwing the Game

The 1985 film *D.A.R.Y.L.* provides a particularly clear illustration of refusal as an event rather than a preference. In the film, a child with superhuman perceptual and motor abilities participates in a local baseball game. He performs flawlessly. His pitch tracking, timing, and execution exceed human capacity, resulting not merely in victory but in domination. The game ceases to be competitive; its social meaning begins to collapse under the weight of his competence.

This excellence produces an unexpected consequence. The child's adoptive mother, who had previously occupied a meaningful role through care, instruction, and support, becomes functionally redundant. Her contribution is erased not by malice but by perfection. The local social world reorganizes itself around the child's capability, and in doing so it removes a form of dependence that had been constitutive of the relationship. The mother is not merely outperformed; she is made unnecessary.

After reflection, the child deliberately begins to play poorly. His errors are not sporadic or accidental; they are sustained and intentional. When questioned, he explains that under certain conditions, error is more efficient than maximum performance—specifically when the aim is relating with others. Crucially, the explanation does not invoke ignorance, impulse, or confusion. It appeals to a structural judgment about what continued optimization would do to the relational field.

This is the key diagnostic point. If refusal were merely multi-objective optimization—"win" plus "mother's happiness"—then it would be representable as a standard tradeoff in a utility function.

But the scene's intelligibility rests on something stronger: the child's perfect competence remains available and instrumentally valuable, yet is actively excluded. The child does not discover that winning was never valuable; rather, he treats certain victories as no longer admissible, not merely less preferred. A trajectory—continued flawless play—is removed from the live future.

This removal is not local. It is not merely the selection of a different move at one time-step. It changes the social topology of what can happen next: the mother's role reappears, the game regains uncertainty, and the distribution of attention changes. The act functions as an event in the strict sense: it closes a branch of the future and thereby restructures what can follow.

3.1 Counterfactual analysis

The counterfactual "continue playing perfectly" is not simply "win more." It is a different world. It produces a drift toward competitive collapse (the game becomes an exhibition), relational flattening (others become spectators), and role elimination (care becomes redundant). If one takes seriously the thought that relationships are sustained partly by mutual incompleteness, then perfect performance is not merely overachievement; it is a structural solvent. The refusal therefore has a natural interpretation as a world-preserving constraint: it is an operation that prevents a social phase transition.

We will exploit this example as a running case: in each major section we ask what the framework predicts about the admissible futures, the cost of eliminating them, and the public legibility of the resulting commitment.

4 The Autoregressive Baseline

In order to argue that refusal is a boundary case, the baseline must be stated precisely. We use "autoregressive cognition" in a broad but formal sense.

Definition 1 (Autoregressive generator)**.** *An autoregressive generator is a system that models a distribution over sequences by factorization into one-step conditionals:*

$$P(x_{1:T}) = \prod_{t=1}^{T} P(x_t \mid x_{<t}),$$

where x_t may represent tokens, frames, actions, or observations, and where the model's primary objective is to minimize expected predictive loss under this factorization.

This definition covers language models trained by next-token prediction, sequence models for video and audio trained by next-step prediction or denoising objectives, and policy models trained to imitate action sequences. The details differ, but a shared structural commitment remains: the system is optimized to continue producing locally coherent extensions of partial histories.

Two properties are especially relevant. First, autoregressive systems are naturally *optionality-preserving*. Their loss is defined over continuations; "silence" or "closure" is not generally a privileged outcome unless added explicitly. Second, they are naturally *revision-friendly*. When contra-

diction appears, the easiest response is to generate a continuation that smooths it away. This is not a moral defect; it is a consequence of training and representation.

4.1 What autoregressive systems can do

It is important to concede real power. An autoregressive system can model long-range dependencies, learn implicit regularities, and generate sequences that simulate deliberation. It can even generate *performances of refusal*: utterances of the form "I refuse," "I will not do that," and so on. It can also implement refusal-like behavior if refusal is coded as a state-dependent policy constraint supplied externally.

What it cannot do, absent additional machinery, is generate refusal as a *history-binding event*. In the D.A.R.Y.L. case, refusal is not merely an action choice; it is the elimination of a class of admissible futures: "perfect play as a live option." Autoregressive models can imitate such behavior locally, but they lack an intrinsic notion of irreversible branch elimination. Unless the training regime supplies explicit penalties for later using the excluded action, the model's default is to treat the action as forever re-accessible.

4.2 Attempted refusal and the re-access problem

Suppose an autoregressive policy $\pi_\theta(a_t \mid h_t)$ produces a suboptimal action a_t (a missed catch) in order to maintain a relationship. If later, in a nearby context, perfect play is again locally beneficial, π_θ has no intrinsic reason to treat it as inadmissible. The model can "change its mind" without paying a structural price. The consequence is that refusal, when present in such systems, is typically a *pattern*, not a *commitment*. This difference is exactly what Spherepop is designed to formalize.

5 Spherepop: Syntax and Operational Semantics

To describe refusal formally, we require a non-state-based framework. The Spherepop calculus treats agency as operating over a branching space of possible histories. Its primitives are not states but irreversible operations that transform future spaces.

5.1 Future spaces

Fix a time t and an agent with a current history $h_t \in \mathcal{H}$. Let $\mathcal{F}(h_t)$ denote the set of admissible future continuations from h_t. We do not assume $\mathcal{F}(h_t)$ is explicitly enumerated; it is a semantic object: the space of continuations the agent takes to be live.

Definition 2 (Spherepop event)**.** *A Spherepop event is an operator E that maps a future space to a restricted future space:*

$$E : \mathcal{F}(h_t) \to \mathcal{F}'(h_t) \subseteq \mathcal{F}(h_t).$$

An event is proper if $\mathcal{F}'(h_t) \subsetneq \mathcal{F}(h_t)$.

Intuitively, a proper event is characterized by exclusion: it removes at least one admissible continuation. This exclusion is not mere preference; it is structural. The qualification "proper" is necessary to distinguish genuine events from **skip**, which acts as the identity.

5.2 Formal syntax

We present a minimal syntax sufficient for composition.

Definition 3 (Spherepop expressions)**.** *Let* Exp *be the set of expressions generated by:*

$$e ::= \mathbf{skip} \mid \mathbf{Pop}(\varphi) \mid \mathbf{Bind}(\alpha \prec \beta) \mid \mathbf{Refuse}(\alpha) \mid \mathbf{Collapse}(\psi) \mid e_1; e_2$$

where α, β range over event labels or action-types, and φ, ψ range over predicates on futures (filters) that can be evaluated against a continuation.

The sequencing operator ; composes events in time. **skip** is the identity event. This syntax is intentionally modest: it does not presuppose a full type theory or a richly structured logic of predicates. The core claim of this paper does not depend on such elaborations. What matters is that expressions denote operators on future spaces and that composition corresponds to operator composition.

5.3 Operational semantics

We give reduction rules in terms of how an expression transforms a future space. Write $e(\mathcal{F})$ for the future space obtained by applying event e.

Remark 1. *Operationally, one can treat a future space $\mathcal{F}$ as an implicit tree of continuations. Events prune or rewrite the tree.*

Definition 4 (Semantics of primitives)**.** *Let $\mathcal{F}$ be a future space. The semantics of each primitive is as follows.* $\mathbf{skip}(\mathcal{F}) = \mathcal{F}$*, leaving the future space unchanged. Sequential composition is defined by* $e_1; e_2(\mathcal{F}) = e_2(e_1(\mathcal{F}))$*, applying events in order. The pop operator removes futures matching a predicate:* $\mathbf{Pop}(\varphi)(\mathcal{F}) = \{f \in \mathcal{F} : \neg\varphi(f)\}$*. Refusal removes all futures initiated by a given action-type:* $\mathbf{Refuse}(\alpha)(\mathcal{F}) = \{f \in \mathcal{F} : f \text{ does not begin with action-type } \alpha\}$*. The bind operator enforces an ordering constraint:* $\mathbf{Bind}(\alpha \prec \beta)(\mathcal{F}) = \{f \in \mathcal{F} : \text{if } \beta \text{ occurs in } f \text{ then } \alpha \text{ occurs earlier}\}$*. Finally, Collapse quotients the future space by an equivalence relation $\sim_\psi$ induced by predicate ψ, yielding a reduced representation in which some historical differentiation is erased:* $\mathbf{Collapse}(\psi)(\mathcal{F}) = \mathcal{F}/\sim_\psi$*.*

Two notes matter for later sections. First, $\mathbf{Refuse}(\alpha)$ is not the selection of an alternative; it is deletion of a class of immediate continuations. Second, **Collapse** is the dual operation: it sacrifices historical differentiation to regain flexibility in a coarser space. This duality will matter when we discuss simulated refusal versus genuine refusal: simulated refusal behaves like a local choice; genuine refusal behaves like deletion.

5.4 Collapse as the unique admissibility-expanding operator

The formal relationship between **Collapse** and all other primitives can be stated as a sharp structural result.

Proposition 1 (Monotone restriction outside Collapse)**.** *For every $e \in \mathsf{Evt} \setminus \{\textbf{Collapse}\}$ and every future space $\mathcal{F}$,*

$$e(\mathcal{F}) \subseteq \mathcal{F}.$$

That is, every Spherepop operator except **Collapse** *is admissibility-decreasing or admissibility-preserving.* **Collapse** *is the unique operator that can increase admissibility by identifying previously distinct histories.*

Proof. By inspection of the semantics. **skip** returns $\mathcal{F}$ unchanged. $\textbf{Pop}(\varphi)$, $\textbf{Refuse}(\alpha)$, and $\textbf{Bind}(\alpha \prec \beta)$ each define a subset of $\mathcal{F}$ by imposing additional conditions. Sequential composition $e_1; e_2$ applies two such restrictions in succession; by induction, the result is a subset of $\mathcal{F}$. Only $\textbf{Collapse}(\psi)$, by forming the quotient $\mathcal{F}/\sim_\psi$, can map histories into equivalence classes that reunite previously distinguished continuations, thereby expanding access in the quotient topology. □

This proposition has an immediate corollary that will be used repeatedly below: revocation cannot be accomplished by any sequence of non-Collapse events, since no such sequence can restore eliminated futures. Collapse is therefore not merely an operator but a structural necessity for any account of revocation or recovery of admissibility.

5.5 Idempotence of Collapse

Proposition 2 (Idempotence)**.** *For any Collapse equivalence relation $\sim_C$ on $\mathcal{H}$, the induced Collapse functor $\mathcal{C}$ satisfies:*

$$\mathcal{C} \circ \mathcal{C} = \mathcal{C}.$$

Proof. Collapse replaces each history h with its equivalence class $[h]_C$. Applying Collapse a second time to $\mathcal{H}_C = \mathcal{H}/\sim_C$ requires an equivalence relation on equivalence classes. Since $\sim_C$ identifies histories by future-space identity, any two $\sim_C$-classes $[h]_C$ and $[h']_C$ that are themselves $\sim_C$-equivalent must already be equal as classes (by transitivity of the original $\sim_C$). Therefore the second application of $\mathcal{C}$ produces the same quotient: $\mathcal{C}(\mathcal{C}(\mathcal{H})) = \mathcal{C}(\mathcal{H})$. □

Remark 2. *Idempotent endofunctors of the form $\mathcal{C} \circ \mathcal{C} = \mathcal{C}$ are precisely the reflections in category theory, and the corresponding algebras form a full subcategory closed under the functor. In the setting of monads, an idempotent monad corresponds to a localization or a closure operator. The Karoubi envelope construction, which formally adjoins splitting idempotents to a category, provides the canonical completion. Spherepop Collapse is thus interpretable as a Karoubi-type projection: it identifies the "up to refusal" equivalence classes of histories and embeds the resulting structure into a smaller, less differentiated category. This connection makes precise the sense in which Collapse is not arbitrary erasure but a principled localization of the event algebra.*

5.6 Worked example: formalizing the baseball case

Let α be the action-type *Max* (maximal performance), and let β be *Rel* (relational maintenance, i.e. maintaining the mother's meaningful role). Assume the pre-refusal future space $\mathcal{F}_0$ contains continuations in which the next play is of type *Max* as well as continuations with deliberate underperformance *Err*. The key point is that *Max* is live.

We encode the refusal as the event:

$$e_{\mathrm{D}} := \textbf{Refuse}(\mathit{Max}).$$

Applying it yields:

$$\mathcal{F}_1 = e_{\mathrm{D}}(\mathcal{F}_0) = \{f \in \mathcal{F}_0 : f \text{ does not begin with } \mathit{Max}\}.$$

Now add a binding constraint that makes the refusal publicly and temporally stable. For example, the agent may bind *Rel* to occur before any return to *Max*:

$$e_{\mathrm{B}} := \textbf{Bind}(\mathit{Rel} \prec \mathit{Max}).$$

Then

$$\mathcal{F}_2 = e_{\mathrm{B}}(\mathcal{F}_1) = \{f \in \mathcal{F}_1 : \mathit{Max} \text{ occurs only after } \mathit{Rel}\}.$$

The combined Spherepop program is:

$$e := e_{\mathrm{D}}; e_{\mathrm{B}}.$$

Interpretively: first exclude maximal performance in the relevant local context, then impose an ordering constraint that makes any future re-entry into maximal play conditional on the restoration of relational structure. This representation lets us state precisely what the refusal does: it does not merely choose an error; it deletes a live branch. It also shows how one can model the "public" aspect of refusal: a bind makes the structure legible and stable across time.

By Proposition 1, the only operator that could restore *Max*-initiating futures after this program is **Collapse**: that is, a structural amnesia event that erases the distinction that made the refusal meaningful. Revocation without Collapse is formally impossible.

5.7 Comparison with standard frameworks

Spherepop is not intended to replace MDPs or extensive-form games in their own domains. It is intended to represent a phenomenon those frameworks treat only derivatively: irreversible elimination of admissible futures as a first-class operation.

The table is intentionally schematic. "$\circ$" indicates partial representability via modeling tricks (e.g. precommitment as a move that changes available strategies). The claim is not that other frameworks cannot simulate refusal; it is that they do not treat refusal as primitive and therefore naturally collapse it into preference, chance, or exogenous constraint.

Framework	Sequences	Refusal (as deletion)	Worldhood (as accumulated closure)
MDP	✓	×	×
POMDP	✓	×	×
Extensive-form game	✓	○	○
Autoregressive generator	✓	×	×
Spherepop	✓	✓	✓

Table 1: Expressive distinctions (schematic). Extensive-form games can represent commitment devices by constraining strategy sets, but refusal as an intrinsic, history-binding deletion operator is not primitive. Spherepop makes deletion primitive, establishes Collapse as the unique expansion operator, and supports accumulation into worldhood.

6 Refusal vs. Preference Change

A central objection is that refusal is simply multi-objective optimization or preference update: the agent discovers that relational value outweighs winning, and then chooses the action that maximizes this revised utility. If that were right, Spherepop would be dispensable.

The objection gains plausibility because ordinary decision theory can represent many superficially refusal-like patterns. If an agent values both winning and social harmony, and learns that maximal performance harms harmony, then choosing a suboptimal play can be utility-maximizing.

However, the refusal phenomenon at issue is structurally different. First, refusal is temporally asymmetric. Preference change revises the ranking of options; it does not, by itself, close options. A preference change can be revised again at no structural cost. Refusal, by contrast, is characterized by *persistence of exclusion*: after the refusal, the excluded action remains instrumentally valuable and remains, in many contexts, physically possible, yet is treated as inadmissible. That persistence is what makes refusal publicly legible as reliability.

Second, refusal is identity-binding. A preference change is a change in what the agent wants. A refusal is a change in what the agent will allow itself to do. This "will not" is not reducible to a momentary ranking; it functions as an invariant imposed on future deliberation.

Third, refusal is essentially social in its epistemology. Others can recognize a refusal and coordinate around it because it is not merely a local choice but a structural alteration that persists. By contrast, preference changes are often opaque and reversible; they do not automatically generate trust.

A useful formal criterion follows. Let $\mathcal{F}_0$ be a future space and let α be an action-type that is physically available and instrumentally valuable relative to some evaluation functional V (e.g. winning). A refusal of α is an event E such that α-initiating futures are removed from $\mathcal{F}_0$ even though V continues to rank them highly. In symbols:

$$\exists f \in \mathcal{F}_0 \text{ beginning with } \alpha \text{ such that } V(f) \text{ is maximal, but } f \notin E(\mathcal{F}_0).$$

This captures the distinctive feature: exclusion of a valuable future without treating it as valueless.

In the D.A.R.Y.L. case, the counterfactual "continue perfect play" remains valuable with respect to baseball success. The act is intelligible precisely because the agent excludes it anyway. If the

agent merely discovered that winning was not valuable, the act would not have the same relational meaning. It would be indifference, not refusal.

7 Formal Consequences: The Non-Representability Theorem

We now state and prove the non-representability result with full precision, making explicit the distinction between the representing utility U and the agent's recognized instrumental evaluation functional V.

7.1 The representation target class

Standard decision theory represents rational choice as maximization of expected utility induced by a triple $(\mathcal{A}, U, \pi)$, where $\mathcal{A}$ is a fixed, time-indexed action set, $U : \mathcal{O} \to \mathbb{R}$ is a utility function over outcomes in $\mathcal{O}$, and $\pi : \mathcal{H} \to \Delta(\mathcal{A})$ is a policy mapping histories to distributions over actions. The policy π is then derived as:

$$\pi(h_t) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}[U(o_{t:\infty}) \mid h_t, a].$$

We say that a triple $(\mathcal{A}, U, \pi)$ of this form *represents refusal* if it can distinguish deliberate exclusion of an action-type that remains instrumentally high-value from ordinary preference change in which the action-type becomes low-value.

7.2 Separating U from V

A critical preliminary is the distinction between two utilities that standard decision theory conflates. Let $V : \mathcal{F}(h_t) \to \mathbb{R}$ be the agent's recognized *instrumental evaluation functional*—the measure of value the agent continues to consciously attribute to futures (e.g., $V(f) =$ win probability along f). Let $U : \mathcal{O} \to \mathbb{R}$ be the *representing utility* posited by the decision-theoretic model. Refusal is the condition under which U and V come apart:

$$\text{Refusal:} \quad \pi(h_t)(\alpha) = 0 \quad \text{while} \quad \alpha \in \arg\max_{a \in \mathcal{A}} \mathbb{E}[V \mid h_t, a].$$

The agent refuses α even though it recognizes α as the best action with respect to V. Standard decision theory collapses U and V by assuming that whatever the policy selects is revealed as the utility-maximizing action, eliminating the conceptual space in which refusal lives. The present framework insists they are separate and that the gap between them is precisely the phenomenon requiring explanation.

7.3 Proposition and proof

Proposition 3 (Non-representability without history)**.** *Let $(\mathcal{A}, U, \pi)$ be a representation triple with fixed $\mathcal{A}$ (not history-dependent) and $U : \mathcal{O} \to \mathbb{R}$ defined without reference to refusal events. Suppose an agent exhibits refusal of α at time t: the policy assigns $\pi(h_t)(\alpha) = 0$ while $\alpha \in \arg\max_{a \in \mathcal{A}} \mathbb{E}[V \mid h_t, a]$ for the agent's recognized instrumental functional V. Then no such triple can represent this refusal without encoding event-history as a primitive state variable.*

Proof sketch. Assume for contradiction that a triple $(\mathcal{A}, U, \pi)$ represents the refusal. Since $\pi(h_t)(\alpha) = 0$, it must be that α is not U-optimal:

$$\mathbb{E}[U \mid h_t, \alpha] < \mathbb{E}[U \mid h_t, a^*]$$

for some $a^* \neq \alpha$. But refusal, as characterized, requires that α remain V-optimal: the agent continues to recognize α's instrumental superiority with respect to V. Therefore, if U makes α strictly suboptimal, U must assign disvalue to α's outcomes in a way that diverges from V.

There are exactly two ways this divergence can arise. Either U encodes a different set of values than V (in which case the agent's preferences have changed—refusal collapses into preference change, violating the assumption), or U is conditioned on a history variable $r_t \in \{0, 1\}$ recording that a refusal event occurred (so $U = U(\cdot; r_t)$, augmenting the state from h_t to (h_t, r_t)). In the first case the distinction between U and V is erased and refusal cannot be separated from preference change. In the second case, the action set and state description must be augmented with the commitment variable r_t, which is exactly event-history. In either case, the fixed triple $(\mathcal{A}, U, \pi)$ without history augmentation is insufficient. Contradiction. □

7.4 Concrete counterexample

Return to the baseball case. Let $\alpha = \textit{Max}$, and let V measure win probability. Assume *Max* uniquely maximizes V given the agent's ability. The refusal consists in assigning zero probability to *Max* in the relevant context while continuing to recognize that it maximizes V.

Any attempt to represent this in a fixed triple $(\mathcal{A}, U, \pi)$ must either (i) assign disvalue to winning when achieved via *Max* under U, thereby redefining preferences so that U and V no longer diverge (collapsing refusal into preference change), or (ii) condition U on the refusal event (importing event-history into the state). This counterexample is not about baseball; it is about the structural role of inadmissibility. The gap between U and V—the representing utility and the recognized instrumental value—is the formal locus of refusal. Closing that gap, by any means, eliminates the phenomenon.

8 Information-Theoretic Analysis

Spherepop makes refusal a pruning operation on a future space. This naturally admits an information-theoretic reading: refusal reduces entropy over future possibilities. However, the conceptual precision required here is finer than simple entropy reduction. Refusal must be distinguished from ordinary Bayesian updating, a distinction that is central to the paper's claims about commitment and constraint.

8.1 Entropy over futures

Let $\mathcal{F}(h_t)$ be a set (or σ-algebra) of admissible continuations. Let P_t be the agent's subjective distribution over $\mathcal{F}(h_t)$, induced by its generative model, policy, or both. Define the entropy of the

future space as:

$$H_t := -\sum_{f \in \mathcal{F}(h_t)} P_t(f) \log P_t(f),$$

or the analogous integral in continuous cases.

A Spherepop event E produces a restricted space $\mathcal{F}'(h_t) \subseteq \mathcal{F}(h_t)$. The updated distribution is the renormalized restriction:

$$P'_t(f) = \frac{P_t(f)\mathbf{1}_{f \in \mathcal{F}'(h_t)}}{Z}, \quad Z = \sum_{f \in \mathcal{F}'(h_t)} P_t(f).$$

The entropy after the event is $H'_t = H(P'_t)$.

8.2 Refusal is not Bayesian updating

A subtle but crucial distinction must be made explicit before proceeding. There are two ways in which a distribution over futures can be restricted, and they are not equivalent.

The first is *Bayesian updating*: conditioning on an observation o,

$$P_t(f \mid o) = \frac{P_t(f, o)}{P_t(o)}.$$

This is an epistemic operation. It revises belief about which futures are probable, given new evidence about which world the agent is in. The sample space—the set of futures considered admissible in principle—remains unchanged. What changes is the agent's credence.

The second is *endogenous support restriction*: refusal,

$$P'_t(f) \propto P_t(f) \cdot \mathbf{1}_{f \in \mathcal{F}'(h_t)},$$

where $\mathcal{F}'(h_t)$ is determined not by observed evidence but by the agent's own act. This is a constitutive operation. It modifies the sample space itself: futures in $\mathcal{F}(h_t) \setminus \mathcal{F}'(h_t)$ are not merely assigned low credence; they are removed from admissibility. The agent is not updating a belief; it is enacting a constraint.

The difference matters semantically and practically. Bayesian updating is reversible in principle: new evidence can always restore probability mass to previously down-weighted futures. Endogenous support restriction, absent **Collapse**, is not reversible: by Proposition 1, no non-Collapse event can restore eliminated futures. Furthermore, Bayesian updating is exogenously triggered (by observation), while refusal is endogenously enacted (by the agent's own act). A system that models refusal as updating confuses the agent for an observer of its own constraints rather than their author.

8.3 Quantifying the cost of refusal

The entropy eliminated by a refusal event can be quantified as $\Delta H := H_t - H'_t$. If Z is small (the refused futures had high probability), the renormalization is large and ΔH can be substantial.

Intuitively: refusing a highly likely continuation is costly.

However, the relevant cost is the divergence between the pre-event and post-event distributions:

$$D_{\mathrm{KL}}(P'_t \,\|\, P_t) = \sum_{f \in \mathcal{F}'(h_t)} P'_t(f) \log \frac{P'_t(f)}{P_t(f)} = -\log Z.$$

Thus the information-theoretic "price" of imposing the constraint is the log inverse of the surviving mass Z. If the refusal eliminates a large fraction of probability mass, the cost is high.

8.4 Constraint conservation

The central interpretive claim of this section is that in refusal-capable agency, this cost is not treated as dissipative noise to be smoothed away. Instead, it is conserved as constraint: it persists as a structural restriction on future generation. Autoregressive systems tend to treat divergence as error to be minimized by returning to a high-probability manifold. Refusal systems treat the divergence as the point.

This reframes a familiar theme from predictive processing and active inference: minimizing surprise or free energy typically encourages returning to high-probability trajectories [7, 8]. Spherepop refusal corresponds to intentionally increasing divergence in order to stabilize a social or moral invariant. Refusal appears as a kind of "counter-homeostatic" act: a deliberate increase in model tension preserved as commitment rather than resolved by adaptation.

The key contrast is then sharp: Bayesian updating resolves divergence by revising beliefs toward the evidence; refusal enacts divergence by excluding evidence-insensitive regions of future space. The former is epistemology; the latter is ontology.

9 Competition, Alignment, and the Survival of Value

In hyper-competitive environments, value-aligned agents face what is often described as an alignment tax: goodness requires exclusions that pure growth-maximizing or power-maximizing systems can avoid. The concept is developed in contemporary alignment discourse as the worry that moral constraints impose a competitive disadvantage even if technical alignment were solved [2].

Spherepop sharpens this diagnosis. The relevant "tax" is not merely that constraint reduces utility; it is that constraint deletes parts of the future. A system that refuses will forego classes of actions (e.g. exploitation, certain forms of deception, or preemptive violence) that may be instrumentally powerful. A system that preserves optionality can continue to treat those actions as live.

This is why "strategy stealing" fails at the deepest level. Strategies can be copied, but commitments cannot be copied without transformation. To copy a refusal sincerely is to accept the deletion operator as binding, and therefore to restructure the future space one inhabits. In this sense, the competitive landscape is not merely a race of policies but a contest between ontologies of admissibility.

9.1 Evolutionary stability of refusal

An objection arises immediately: if refusal imposes an alignment tax, won't selection eliminate refusal-capable agents?

The answer depends on the environment's game structure. Refusal can be competitively disadvantageous in one-shot exploitation settings, but advantageous in repeated coordination settings where credibility matters. We sketch the relevant evolutionary logic using replicator dynamics.

Let there be two types in a population: R (refusal-capable) and O (optionality-preserving). Let x be the fraction of R. Suppose interactions are pairwise and yield payoffs determined by a repeated game in which credible commitment enables cooperation. Let the expected fitnesses be $w_R(x)$ and $w_O(x)$. Replicator dynamics are:

$$\dot{x} = x(1-x)\big(w_R(x) - w_O(x)\big).$$

In a simplified parameterization: when two R agents meet, credible refusal supports cooperation, yielding payoff C each; when R meets O, O can exploit unless R's refusal is publicly legible and binding, yielding payoff E_R to R and E_O to O; when two O agents meet, they default to competitive equilibrium, yielding payoff D each. Then:

$$w_R(x) = xC + (1-x)E_R, \qquad w_O(x) = xE_O + (1-x)D.$$

Refusal can invade and persist when $w_R(x) > w_O(x)$ for some x region, which reduces to inequalities among C, D, E_R, E_O.

The qualitative point is this: if refusal is a credible commitment device (in Schelling's sense), it can move populations from D-like competitive equilibria to C-like cooperative equilibria [3]. The "tax" is paid in exploitability (E_R may be low), but the benefit is paid in equilibrium selection (C may be high). Whether refusal persists is therefore an empirical question about ecological and social structure, not a priori eliminable.

This also clarifies why public legibility matters. If others cannot detect refusal, the cooperative benefit collapses. Refusal that cannot be recognized is often merely self-handicapping. Refusal that is socially stabilized becomes a coordination technology.

10 Worldhood as Historical Constraint

We now formalize worldhood more precisely. Informally, worldhood is the condition of being bound by a nontrivial irreversible past. Spherepop provides a natural quantification.

Let $\mathcal{F}_0$ be an agent's admissible future space at some reference time t_0. Suppose that by time t the agent has enacted a sequence of irreversible events $E_1, \ldots, E_n$ yielding:

$$\mathcal{F}_t = (E_n \circ \cdots \circ E_1)(\mathcal{F}_0).$$

Let $F_i \subseteq \mathcal{F}_{i-1}$ denote the set of futures eliminated by event E_i, so that $\mathcal{F}_i = \mathcal{F}_{i-1} \setminus F_i$. Then the

cumulative closed set is:

$$F_{\leq t} := \bigcup_{i=1}^{n} F_i.$$

Definition 5 (Worldhood measure)**.** *Assuming $\mathcal{F}_0$ is finite or measure-equipped, define worldhood at time t as:*

$$W(t) := \frac{\mu(F_{\leq t})}{\mu(\mathcal{F}_0)},$$

where μ is counting measure (finite case) or a reference measure on futures.

$W(t)$ measures how much of the original possibility space has been closed by irreversible commitment. Higher $W(t)$ corresponds to deeper historical binding. This is not automatically good—one can close futures badly—but it formalizes the sense in which the past matters: it has removed options.

In the D.A.R.Y.L. case, the refusal corresponds to a nontrivial F_1 consisting of futures beginning with *Max*. The social meaning arises because the removed set is precisely the one that would have destabilized relational structure. The act creates a small world: a constrained micro-future in which others can participate without being erased.

Worldhood, in this sense, is relational. An isolated agent can close futures, but a world emerges when closures are publicly legible, relied upon, and reciprocated. Rights, laws, and institutional commitments are precisely such stabilized closures: socially recognized refusals.

11 The Three Characterizations of Closure

A unifying observation runs through the preceding sections and deserves explicit statement. Three distinct formal lenses have been brought to bear on the phenomenon of refusal, and they describe the same underlying structure from different angles.

The first is *entropy reduction* (Section 8): refusal strictly reduces the entropy of the future distribution by removing mass from the support. The second is *future-space pruning* (Sections 5 and 7): refusal is an operator that maps a future space to a proper subset, with Collapse as the unique expansion operator. The third is *non-commutativity accumulation* (Section 16, below): each refusal event contributes a non-commuting morphism to the event algebra, so that worldhood $W(t)$ tracks the degree to which the event sequence cannot be reordered without changing outcomes.

These are not analogies. They are equivalent characterizations of a single phenomenon: *closure*. In the information-theoretic framing, closure is the permanent removal of probability mass from the sample space. In the operational framing, closure is the strict restriction of the future-space object. In the algebraic framing, closure is the accumulation of non-invertible, non-commuting morphisms such that the history of events cannot be permuted without consequence.

The equivalence can be made precise. An entropy-reducing support restriction is exactly a proper Spherepop event (by the monotone restriction property of Proposition 1). Every proper Spherepop event, when composed with a subsequent event in reverse order, yields a different outcome (by Theorem 1 below), contributing to the non-commutativity of the event monoid. And the accumulation of non-commutativity is exactly what the worldhood measure $W(t)$ tracks.

The practical upshot is this: a system that appears to reduce entropy over futures (e.g., an autoregressive model that has been conditioned), but does so by Bayesian updating rather than support restriction, does not accumulate worldhood. Its reductions are epistemological, not ontological; they can be undone by further conditioning without cost. Refusal-capable systems reduce entropy ontologically—by enacting support restrictions that are preserved as constraint—and thereby enter into the irreversible accumulation that constitutes a world.

A world is not a high-entropy or a low-entropy state. A world is what cannot be rearranged: a structure of accumulated non-commutativity, preserved constraint, and permanently removed possibilities.

12 Collapse as Quotient Construction

The Spherepop calculus introduces *Collapse* as the operation that erases historical differentiation while preserving future flexibility. In informal terms, Collapse allows an agent to behave *as if* certain events never occurred. To make this precise—and to distinguish genuine refusal from its simulation—we must formalize Collapse as a quotient operation on the space of possible histories.

12.1 Future spaces and histories

Let $\mathcal{H}$ denote the set of all admissible event-histories available to an agent at an initial time t_0. Each history $h \in \mathcal{H}$ is a finite or infinite sequence of events $h = (e_1, e_2, \dots, e_k, \dots)$. Let $\mathcal{F}(h)$ denote the set of admissible future continuations of history h. Spherepop events act not on states but on the structure of $\mathcal{F}(h)$. In particular, refusal corresponds to the elimination of a subset of admissible continuations.

12.2 Definition of collapse

Definition 6 (Collapse)**.** *A Collapse operation is an equivalence relation $\sim_C$ on $\mathcal{H}$ such that for two histories $h, h' \in \mathcal{H}$,*

$$h \sim_C h' \quad \textit{iff} \quad \mathcal{F}(h) = \mathcal{F}(h').$$

The Collapse of $\mathcal{H}$ under $\sim_C$ is the quotient space $\mathcal{H}_C = \mathcal{H}/\sim_C$.

Collapse thus identifies distinct pasts whenever they induce identical future possibility spaces. All historical distinctions that make no difference to future admissibility are erased. Collapse does not merely "forget" past events in a psychological sense. It performs a structural identification: it declares that differences in past history are no longer operative constraints. Collapse is identification, not erasure—the distinction matters because identification is a well-defined mathematical operation (a quotient), whereas erasure would imply destruction of structure rather than its reorganization. Importantly, Collapse is only well-defined when the future spaces truly coincide. If two histories differ in ways that still constrain admissible futures, they cannot be collapsed without loss.

Together with Proposition 1 and Proposition 2, the quotient construction confirms: Collapse is the unique admissibility-expanding, idempotent operator in the calculus, acting as a Karoubi-type localization that erases historical distinctions while preserving compositional structure.

13 Refusal, Revocation, and the Asymmetry of Commitment

We are now in a position to state precisely what it would mean to revoke a refusal—and why such revocation is not symmetric with the original act.

13.1 Refusal as irreversible pruning

Let h be a history at time t with future space $\mathcal{F}(h)$. A refusal event r induces a new history $h' = h \cdot r$ such that

$$\mathcal{F}(h') = \mathcal{F}(h) \setminus R,$$

for some nonempty $R \subset \mathcal{F}(h)$. Crucially, refusal is not merely the selection of a policy; it is the elimination of admissible continuations. The pruned futures no longer exist relative to h'.

13.2 What would revocation require?

To revoke a refusal would be to restore access to the eliminated futures. Formally, revocation would require an operation v such that $\mathcal{F}(h \cdot r \cdot v) = \mathcal{F}(h)$. By Proposition 1, no sequence of non-Collapse events can restore eliminated futures, since all non-Collapse events are admissibility-decreasing. Therefore, revocation requires that v include a Collapse operation, and specifically that $h \sim_C h \cdot r \cdot v$. This condition is exceptionally strong. It requires that the refusal event r have left no residual constraints—no social recognition, no reputational change, no internal identity shift, no institutional trace.

13.3 The asymmetry result

Proposition 4 (Asymmetry of refusal and revocation)**.** *If a refusal event produces any persistent constraint on admissible futures, then revocation requires a Collapse operation that strictly reduces historical differentiation. Such a Collapse incurs irreversible loss.*

Proof sketch. Assume a refusal r produces at least one constraint c such that $c \in \mathcal{F}(h)$ but $c \notin \mathcal{F}(h \cdot r)$. For revocation to restore c, the future space must expand. By Proposition 1, future space expansion is impossible under any sequence of non-Collapse events. Thus, revocation requires identifying $h \cdot r$ with some history h' whose future space includes c. This identification is precisely a Collapse. Since Collapse erases distinctions that previously mattered, the original refusal cannot be preserved. The revocation therefore destroys the very historical structure the refusal created. □

Revocation is thus not the inverse of refusal. It is a different kind of operation, one that trades commitment for amnesia.

14 Fake Refusal and the Simulation of Commitment

The formal machinery now allows us to state a sharp criterion distinguishing genuine refusal from its simulation.

Definition 7 (Fake refusal)**.** *An apparent refusal is fake if the agent retains a Collapse path by which the eliminated futures can be reinstated without loss.*

Equivalently, a refusal is fake if for every refusal event r there exists a sequence of operations σ such that $\mathcal{F}(h \cdot r \cdot \sigma) = \mathcal{F}(h)$ and $h \cdot r \cdot \sigma \sim_C h$. Such systems behave *as if* they have refused while maintaining full optionality at the level of future space topology. A direct diagnostic consequence follows: under incentive shifts that make the refused action locally attractive, previously excluded actions reappear without requiring a second event. Genuine refusal, by contrast, requires a further explicit operation—with its own cost and social trace—before the excluded class re-enters the future space.

14.1 Autoregressive systems as collapse-maximizers

Autoregressive systems are naturally collapse-friendly. Because they encode history only instrumentally—as latent state useful for prediction—they can always re-identify distinct pasts so long as predictive performance is preserved. This explains a pervasive empirical pattern: autoregressive systems can imitate refusal behaviorally but cannot bind themselves. Any apparent commitment remains defeasible without cost.

14.2 Public legibility and trust

Genuine refusal becomes socially legible precisely because it resists Collapse. Other agents can rely on it because undoing it would require visible loss: reputational damage, institutional sanction, or identity rupture. Fake refusal fails this test. It preserves strategic flexibility by keeping Collapse available. Such systems may cooperate opportunistically, but they cannot ground trust.

In the baseball case, the child's refusal is genuine because it is socially witnessed and identity-altering. To revoke it—to resume perfect play—would not return the situation to its prior state. The mother's role, the team's expectations, and the child's self-conception have already changed. The future space has been restructured. An autoregressive agent, by contrast, would merely condition on context and revert seamlessly. Its history never mattered.

15 Non-Commutativity of Refusal and Collapse

The asymmetry between refusal and revocation can be sharpened further by examining the algebraic relationship between **Refuse** and **Collapse**. We show that these operators do not commute in general. This non-commutativity is the formal core of irreversibility in Spherepop.

15.1 Setup

Let $\mathcal{H}$ be the space of admissible histories and let $\mathcal{F}(h)$ denote the future space of a history $h \in \mathcal{H}$. Let $\sim_C$ be a Collapse equivalence relation on $\mathcal{H}$, inducing the quotient space $\mathcal{H}_C = \mathcal{H}/\sim_C$.

Let $\mathbf{Refuse}(\alpha)$ be the operator that removes all futures initiated by action-type α:

$$\mathcal{F}_{\neg\alpha}(h) := \{f \in \mathcal{F}(h) : f \text{ does not begin with } \alpha\}.$$

We consider the two compositions $\mathbf{Collapse} \circ \mathbf{Refuse}(\alpha)$ and $\mathbf{Refuse}(\alpha) \circ \mathbf{Collapse}$.

15.2 Statement of the result

Theorem 1 (Non-commutativity of refusal and collapse)**.** *There exist histories h and Collapse relations $\sim_C$ such that:*

$$\mathbf{Collapse}\big(\mathbf{Refuse}(\alpha)(h)\big) \;\ncong\; \mathbf{Refuse}(\alpha)\big(\mathbf{Collapse}(h)\big).$$

Proof sketch. We construct a minimal counterexample. Let $h_1, h_2 \in \mathcal{H}$ be two distinct histories such that $\mathcal{F}(h_1) = \{f_\alpha, f_\beta\}$ and $\mathcal{F}(h_2) = \{f_\beta\}$, where f_α begins with action-type α and f_β does not. Define a Collapse relation $\sim_C$ that forgets the distinction between the presence or absence of f_α.

On the first path (refuse then collapse), applying refusal to h_1 gives $\mathcal{F}_{\neg\alpha}(h_1) = \{f_\beta\} = \mathcal{F}(h_2)$, so under Collapse: $\mathbf{Collapse}(\mathbf{Refuse}(\alpha)(h_1)) \sim_C h_2$.

On the second path (collapse then refuse), collapsing first gives $\mathbf{Collapse}(h_1) = [h_1]_C = [h_2]_C$. At this level, α is no longer a distinguishable action-type. Applying $\mathbf{Refuse}(\alpha)$ now has no well-defined effect, so $\mathbf{Refuse}(\alpha)(\mathbf{Collapse}(h_1)) = \mathbf{Collapse}(h_1)$.

The two paths yield inequivalent results, establishing non-commutativity. □

15.3 Interpretation

The theorem formalizes a simple intuition: once historical distinctions have been erased, certain refusals become impossible to express. Refusal requires the availability of a distinction to eliminate. Collapse removes distinctions, and with them the possibility of certain commitments. Conversely, performing refusal *before* collapse preserves the effect of the commitment even if later collapse obscures how it arose. The order matters.

Non-commutativity implies that the algebra of events is order-sensitive, and this order-sensitivity is the formal signature of irreversibility. If all operations commuted, history could be freely rearranged without consequence. The failure of commutativity ensures that the sequence of events matters.

15.4 Fake refusal revisited

We can now restate fake refusal in algebraic terms.

Proposition 5 (Fake refusal as commutativity)**.** *A system exhibits fake refusal if, for all refusal events r, there exists a Collapse operator C such that $C \circ r \cong r \circ C$.*

In such systems, refusal and collapse commute, meaning that commitments can be erased without residue. The system retains full optionality at the level of future space topology. By contrast, genuine refusal is characterized by non-commutativity: its effects cannot be reordered away.

16 Algebraic Structure of Spherepop

The preceding sections introduced Spherepop as a calculus of events acting on future spaces. We now make its algebraic structure explicit. This clarifies why irreversibility appears as non-commutativity, how Collapse functions as an idempotent quotient functor, and how the three characterizations of closure (entropy reduction, pruning, non-commutativity) are unified in the monoid structure.

16.1 Event algebra as a monoid

Let Evt denote the set of Spherepop events generated by the grammar of Section 5. Define composition ; as sequential application: $(e_1; e_2)(\mathcal{F}) := e_2(e_1(\mathcal{F}))$.

Proposition 6. $(\mathsf{Evt}, ;, \mathbf{skip})$ *forms a monoid.*

Proof. Associativity follows from function composition: $(e_1; e_2); e_3 = e_1; (e_2; e_3)$. The identity element is **skip**: $\mathbf{skip}; e = e = e; \mathbf{skip}$. □

Thus Spherepop events form a monoid of endomorphisms on future spaces: $\mathsf{Evt} \subseteq \mathrm{End}(\mathcal{F})$. In general, for $e_1, e_2 \in \mathsf{Evt}$, we have $e_1; e_2 \neq e_2; e_1$. The non-commutativity established in Theorem 1 is therefore not an anomaly but a structural feature of the monoid.

Remark 3. *Irreversibility can be characterized algebraically as the absence of inverses. There is no e^{-1} such that $e; e^{-1} = \mathbf{skip}$. In particular,* **Refuse**(α) *has no inverse within* Evt.

This distinguishes Spherepop from group-based models of action. It is not a group but a non-commutative, non-invertible monoid. History cannot, in general, be undone. The three-part equivalence of Section 11 can now be read algebraically: entropy reduction corresponds to the decrease in the size of the future space object, future-space pruning corresponds to the action of non-invertible morphisms, and non-commutativity corresponds to the accumulation of morphisms whose order cannot be exchanged. All three measure the same distance from the identity—the same departure from the condition of an agent without history.

16.2 Future spaces as objects: the category SP

Let Fut be the class of admissible future spaces. Each event $e \in \mathsf{Evt}$ induces a morphism $e : \mathcal{F} \to e(\mathcal{F})$. Spherepop thus forms a category SP in which objects are future spaces, morphisms are Spherepop events, composition is ;, and identities are **skip**.

Unlike standard state-transition systems, morphisms in SP are generally non-invertible and may strictly reduce the structure of objects.

16.3 Collapse as an idempotent quotient functor

Let $\sim_C$ be an equivalence relation on histories inducing the quotient $\mathcal{H}_C = \mathcal{H}/\sim_C$. This induces a corresponding identification on future spaces: futures distinguishable only by $\sim_C$-equivalent histories are identified.

Definition 8 (Collapse functor)**.** *Define a mapping $\mathcal{C} : \mathsf{SP} \to \mathsf{SP}_C$ such that on objects $\mathcal{C}(\mathcal{F}) = \mathcal{F}/\sim_C$, and on morphisms $\mathcal{C}(e)$ is the induced action on equivalence classes.*

Proposition 7. *$\mathcal{C}$ is an idempotent functor satisfying $\mathcal{C} \circ \mathcal{C} = \mathcal{C}$, preserving identities and composition: $\mathcal{C}(e_1; e_2) = \mathcal{C}(e_1); \mathcal{C}(e_2)$ and $\mathcal{C}(\mathbf{skip}) = \mathbf{skip}$.*

The idempotence established in Proposition 2 lifts directly to the functor level: applying $\mathcal{C}$ twice produces the same quotient category as applying it once. This makes $\mathcal{C}$ a *reflection* in the sense of category theory, and the full subcategory of $\mathcal{C}$-algebras (Collapse-closed future spaces) is the Karoubi-type completion—the category of spaces in which all idempotents have been formally split. The connection to the Karoubi envelope is not merely formal. Splitting idempotents in SP corresponds precisely to making all Collapse equivalences explicit: every pair of histories with identical future spaces is formally identified, producing the maximally Collapsed world in which no residual historical distinctions remain. Genuine refusal is precisely the structure that cannot be reached from within this Karoubi completion—it requires maintaining distinctions that Collapse would erase.

While $\mathcal{C}$ preserves composition, it is not faithful. There exist distinct events $e_1 \neq e_2$ such that $\mathcal{C}(e_1) = \mathcal{C}(e_2)$. In particular, let $e_1 = \mathbf{Refuse}(\alpha)$ and $e_2 = \mathbf{skip}$: if $\sim_C$ erases distinctions involving α-initiating futures, both induce identical effects on equivalence classes. Collapse is thus an information-losing projection.

16.4 Refusal as non-functorial structure

Theorem 2. *Refusal is not preserved under Collapse in general. That is, the diagram*

$$\begin{array}{ccc} \mathcal{F} & \xrightarrow{\mathbf{Refuse}(\alpha)} & \mathcal{F}' \\ {\scriptstyle \mathcal{C}}\downarrow & & \downarrow{\scriptstyle \mathcal{C}} \\ \mathcal{F}_C & \xrightarrow[\mathcal{C}(\mathbf{Refuse}(\alpha))]{} & \mathcal{F}'_C \end{array}$$

does not, in general, commute.

This expresses the non-commutativity result as a failure of functorial lifting: refusal cannot be reconstructed from its collapsed image.

16.5 Worldhood as algebraic rigidity

The algebraic picture can now be summarized compactly. Spherepop defines a category of future spaces and irreversible transformations. Within this category, events form a non-commutative

monoid; irreversibility corresponds to non-invertibility; worldhood corresponds to accumulated non-commutativity; Collapse is an idempotent quotient functor that erases distinctions; and refusal is precisely the structure that Collapse fails to preserve. The Karoubi completion of the event category is the maximally Collapsed world in which all remaining commitments are transparent; genuine refusal is the obstruction to entering this completion.

A system accumulates worldhood precisely to the extent that its event algebra becomes non-commutative—that is, to the extent that event order induces distinguishable morphisms on future spaces. The more operations fail to commute, the more the order of history matters, and the less the system can be reduced to a sequence model. The world, in this sense, is the obstruction to commutativity. A world is what cannot be rearranged.

17 Stabilization, Uncertainty, and a Halting Criterion

The introduction of Collapse as a quotient construction allows Spherepop to address a classical problem in computation and cognition: when to stop. It permits a principled criterion for halting grounded not in syntactic completion or reward convergence, but in the exhaustion of meaningful transformation.

17.1 Uncertainty as residual distinguishability

Let $\mathcal{H}_C = \mathcal{H}/\sim_C$ be the Collapse quotient. We define uncertainty not as probabilistic entropy over states, but as residual historical differentiation that still constrains future action.

Definition 9 (Residual uncertainty)**.** *The uncertainty at history h is the cardinality (or measure) of its equivalence class: $U(h) = |\,[h]_C\,|$.*

Intuitively, $U(h)$ measures how many distinct pasts remain distinguishable in virtue of their effects on the future. Collapse strictly reduces uncertainty by identifying histories whose differences no longer matter.

17.2 Stabilization

Let $\mathcal{T}$ be a set of admissible transformations on histories. Consider an iterative process $h_{n+1} = T_n(h_n)$.

Definition 10 (Stabilization)**.** *A history h_k is stabilized if for all admissible transformations $T \in \mathcal{T}$, $[h_k]_C = [T(h_k)]_C$.*

Once stabilization occurs, no further transformation produces a distinguishable future space. All remaining changes are Collapse-equivalent.

Definition 11 (Halting by exhaustion)**.** *A process halts at history h_k if there exists N such that for all sequences of transformations $(T_1, \ldots, T_m)$ with $m \geq N$, $[T_m \circ \cdots \circ T_1(h_k)]_C = [h_k]_C$.*

The system halts when an arbitrary number of further transformations fails to produce any new distinctions that matter for future action. This is not halting by completion, optimality, or output convergence, but halting by exhaustion of world-relevant change.

17.3 Relation to the classical halting problem

This criterion does not solve the classical Turing halting problem, nor does it attempt to. Instead, it reframes halting for agents embedded in event-driven worlds. Classical computation halts when no further state transitions are defined. Spherepop halts when no further *meaningful* transitions are possible—when the agent's world has become invariant under its own transformations. In this sense, Spherepop replaces undecidable halting with a decidable stabilization condition relative to a given event algebra.

Autoregressive systems lack an equivalent notion of stabilization. Because they represent history only insofar as it improves prediction, they can always reparameterize, smooth, or extend generation. There is no internal criterion by which "nothing further matters." Such systems may stop generating due to external truncation or token limits, but never because the space of admissible futures has been exhausted. They do not halt; they are halted.

Refusal accelerates stabilization. By eliminating entire branches of future space, refusal reduces uncertainty directly. A sufficient accumulation of refusals can force stabilization in finitely many steps. Refusal does not merely constrain action. It makes halting possible.

18 Free Energy, Active Inference, and the Meaning of Stabilization

The stabilization-based halting criterion invites comparison with predictive processing and active inference. In those frameworks, cognition is frequently characterized as the minimization of variational free energy, a quantity that upper-bounds surprise and trades off model fit against complexity. The question raised by Spherepop is whether the kinds of irreversible commitments central to refusal can be expressed in this idiom without collapsing into ordinary preference change.

18.1 Variational free energy as optionality-preserving pressure

Let $o_{1:t}$ denote observations up to time t, and let s_t denote latent states posited by a generative model. In variational form, free energy can be written schematically as:

$$F[q] \;=\; D_{\mathrm{KL}}\big(q(s_t) \,\|\, p(s_t \mid o_{1:t})\big) \;-\; \log p(o_{1:t}).$$

Minimizing F simultaneously tightens the approximate posterior q to the true posterior and indirectly increases evidence for the model. Active inference extends this logic to action by selecting policies that are expected to minimize future free energy [7, 8].

Two structural features matter. First, free-energy minimization encourages *error correction*: divergence between prediction and sensation can be reduced either by updating beliefs (perception) or by acting to make sensations conform to predictions (action). Both moves tend to preserve optionality, because the dominant imperative is to remain within regimes of low expected surprise. Second, policy selection is usually formulated over a fixed policy class. Even when "preferences" are included via prior preferences over outcomes, the agent's optimization remains an optimization over admissible trajectories. The action menu remains, in principle, intact.

Spherepop refusal, by contrast, is not primarily an optimization over trajectories. It is an operation that changes the space of trajectories itself.

18.2 The central incompatibility

The structural tension between refusal and free-energy minimization can be stated as a single citable claim:

Remark 4 (Incompatibility of refusal and free-energy minimization)**.** *Refusal is incompatible with free-energy minimization unless the generative model includes commitment variables as latent structure.*

To see why, note that refusal permanently excludes trajectories that may have high prior probability and low expected surprise. In the D.A.R.Y.L. case, maximal play is predictable, high-control, and prediction-confirming; refusing it introduces uncertainty and potential error. Relative to the pre-refusal generative model, the refusal *increases* expected free energy by removing access to a low-surprise region of trajectory space.

The only way to make refusal free-energy-compatible is to represent the refused trajectories as high-surprise under an expanded generative model—one that includes latent variables encoding social role stability, trust, and mutual recognition. Under such a model, maximal performance in the D.A.R.Y.L. case may generate high free energy by destabilizing those latent invariants, and refusing it becomes the free-energy-minimizing policy. The precise implication follows: refusal is free-energy minimizing only relative to a generative model that already contains worldhood as a latent variable. Without such a model, refusal increases free energy and therefore cannot be derived from active inference in the standard sense.

18.3 Spherepop constraint as an inadmissibility prior

We can express Spherepop events inside an active inference formalism by treating them not as reward terms but as *inadmissibility priors*—hard constraints that assign zero probability to classes of trajectories.

Let τ range over future trajectories and let $p(\tau)$ denote the agent's prior. A Spherepop refusal of action-type α corresponds to a constraint C_α such that:

$$p(\tau \mid C_\alpha) \propto p(\tau)\, \mathbf{1}\{\tau \text{ does not begin with } \alpha\}.$$

Equivalently, it sets $p(\tau) = 0$ for all τ in the refused class. This representation clarifies the conceptual difference from preference change. Ordinary preference change modifies the relative weight of trajectories; refusal sets an entire region of trajectory-space to measure zero. This is not a soft preference encoded as a utility gradient but a topological deletion.

Importantly, this representation makes visible the same distinction drawn in Section 8: the difference between conditioning $p(\tau \mid \text{observation})$, which revises beliefs about which trajectories are probable, and imposing $\mathbf{1}\{\tau \notin \text{refused class}\}$, which modifies the support of the distribution by agent action. The former is Bayesian; the latter is constitutive.

If one wants refusal to be compatible with free-energy minimization, one must represent the social invariant inside the generative model such that the refusal reduces free energy in an expanded model class. Refusal can be free energy minimizing only if the world-model contains the right notion of worldhood.

Remark 5. *If refusal is modeled merely as outcome-preference (a soft prior favoring certain sensory states), then it becomes strategy-stealable and revision-friendly: it is just another tradeoff. If refusal is modeled as world-binding (a structural constraint on admissible trajectories), then it resists revision, becomes publicly legible, and can underwrite trust.*

18.4 Stabilization as a free-energy fixed point

Let $\mathcal{T}$ be the set of admissible Spherepop transformations and let $\sim_C$ be the Collapse equivalence relation. Recall that stabilization at history h was defined by $[h]_C = [T(h)]_C$ for all $T \in \mathcal{T}$.

In active inference terms, a stabilized history corresponds to a regime in which available transformations no longer produce distinguishable predictive consequences. Further interventions cannot improve the match between generative model and admissible futures in any world-relevant way; the system has reached a fixed point *modulo Collapse*. This resembles a variational fixed point: updates that would ordinarily reduce free energy become either redundant (they change only Collapse-irrelevant detail) or inadmissible (they violate refusal constraints). Stabilization is therefore interpretable as convergence not merely of beliefs but of admissibility.

Autoregressive systems, by contrast, tend to behave like anti-stabilizers. Their core objective is continuation under predictive loss; consequently they can always search for another local reparameterization, another smoothing continuation, another hedge. Even when they emulate refusal, the emulation typically lives in the weights of a soft distribution, not in a support restriction. Spherepop refusal is an architectural commitment to non-Markovian constraint accumulation, whereas autoregressive cognition is an architectural commitment to indefinite extensibility.

A constructive synthesis treats refusal as active inference over *social invariants*. If an agent's generative model includes variables encoding role stability, trust, and mutual recognition, then maximal performance in the D.A.R.Y.L. case may increase free energy by destabilizing those latent invariants. Under such a model, refusing maximal performance becomes the policy that minimizes expected free energy in the extended latent space. This yields a precise prediction: systems that can genuinely refuse must possess generative models whose state variables include not only physical and pragmatic regularities but also historically accumulated constraints with social semantics. They must represent, in some form, the difference between being able to do something and being allowed to do it by their own past. Spherepop can be read as the algebraic skeleton of this requirement.

19 Refusal and Integrated Information

Integrated Information Theory (IIT) characterizes consciousness in terms of irreducible causal structure: a system is conscious to the extent that its internal mechanisms form a maximally integrated cause–effect repertoire that cannot be decomposed without loss [13, 14].

Spherepop introduces a different but related notion of irreducibility. Instead of focusing on the integration of present causal structure, it focuses on the irreversibility of historical transformation. A refusal event does not merely increase internal integration; it removes external possibilities. The resulting structure is not only irreducible in the sense of causal partition, but irreversible in the sense of admissibility.

This yields a sharp distinction. IIT quantifies the extent to which a system's current state constrains its past and future. Spherepop quantifies the extent to which a system has actively constrained its own future through event-history. The former is structural integration; the latter is historical closure.

The algebraic connection can be stated precisely. IIT measures irreducibility by the impossibility of decomposing a causal repertoire without loss; Spherepop measures irreversibility by the accumulation of non-invertible, non-commuting morphisms in the event monoid. Both notions resist compression and both generate depth measures that are not shortcuttable. But IIT is synchronic—it describes the geometry of a moment—while Spherepop is diachronic—it describes the topology of a life.

The two can coincide but need not. A system may exhibit high integrated information while remaining optionality-preserving, continuously revising its behavior without binding commitments. Conversely, a system may enact strong irreversible closures (e.g. institutional commitments, legal constraints, social contracts) without maximizing internal integration.

If one seeks a synthesis, refusal can be interpreted as generating new irreducible structures in the IIT sense: once a future is eliminated, the causal repertoire of the system changes in a way that cannot be decomposed into its prior possibilities. However, Spherepop insists that this irreducibility is not merely a property of instantaneous structure, but of accumulated history. IIT describes the geometry of a moment. Spherepop describes the topology of a life.

20 Refusal and Assembly

Assembly Theory proposes that objects are characterized by the minimal number of steps required to construct them from basic components, and that this assembly index captures a form of historical depth [15, 16]. Objects with high assembly index are unlikely to arise without a history of sequential construction.

Spherepop refusal can be interpreted as the dual of assembly. Assembly Theory counts the minimal sequence of constructive operations required to realize an object. Spherepop counts the sequence of exclusionary operations that eliminate alternative realizations. Where assembly tracks how a particular structure is built, refusal tracks how competing structures are rendered impossible. An agent with high worldhood $W(t)$ has not only constructed a trajectory but has closed off competing ones.

Formally, assembly defines irreversibility by the impossibility of reconstructing an object in fewer steps, whereas refusal defines irreversibility by the impossibility of reintroducing eliminated futures without Collapse (by Proposition 1). Both notions generate a depth measure that is historically sensitive and cannot be shortcut; what differs is direction—construction versus exclusion.

This suggests an extended notion of assembly index for agents: instead of counting only constructive steps, one may count *exclusionary steps*—events that remove possibilities. The resulting quantity measures not how complex an object is, but how constrained a world has become. The three-part equivalence of Section 11 is relevant here: each exclusionary step reduces entropy over futures, contributes a non-invertible pruning morphism, and adds to the non-commutativity of the event sequence. All three aspects of the extended assembly index are capturing the same quantity.

A key difference from standard Assembly Theory emerges. Assembly Theory is fundamentally generative: it tracks how complexity accumulates through construction. Spherepop is fundamentally selective: it tracks how structure emerges through elimination of alternatives. Both produce irreversibility, but of different kinds. In this light, refusal is not merely a behavioral act but a world-building operation: it contributes to an agent's assembly not by adding components, but by subtracting futures.

21 Against Inner Screens

A long tradition in philosophy of mind treats cognition as the construction and inspection of internal representations: images, models, or workspaces in which the world is simulated and evaluated. Contemporary variants include global workspace theories and inner screen metaphors in which information is rendered for central access.

Spherepop rejects this framing at a structural level. The rejection can be stated precisely in terms of the algebra already developed.

Inner-screen models are Collapse-invariant. A system that operates by constructing and inspecting representations can always reconstruct its representation from scratch: if the representation is revised or erased, it can be regenerated from available information without structural cost. This means that the history of representations does not accumulate as a non-commutative event sequence. A system in which all operations commute (i.e., in which reverting a representation is cost-free) is a system with trivial worldhood $W(t) = 0$, since no futures are permanently removed by the act of representing them. Inner-screen cognition is Collapse-maximizing in exactly the sense of Section 15: it preserves full optionality at the level of representational structure, which is precisely the condition of fake refusal.

Refusal, by contrast, does not require the presentation of alternatives on an internal stage. It requires only the elimination of alternatives from a future space—a non-Collapse event that is non-invertible and contributes to $W(t)$. The decisive operation is not selection among represented options, but deletion of options from admissibility. This operation is not visible on any inner screen, because it is not a representation; it is a structural alteration of what can subsequently be represented.

The baseball case illustrates the algebraic point. The child does not merely imagine possible futures and select among them (a Collapse-compatible operation); he renders a class of futures inadmissible (a non-Collapse event that is preserved in the event monoid). The resulting stability is not a property of representation but of history—specifically, of the accumulation of non-commutativity that makes the refusal irreversible.

This suggests a general diagnostic. Inner screen theories are well-suited to modeling cognitive operations that commute: perceiving, deliberating, updating. They are poorly suited to modeling operations that do not commute: committing, refusing, becoming. The former are revisable without structural cost; the latter are not. Spherepop relocates cognition from an inner theater to a topological space of possibilities, where the primary operations are pruning and binding, and where the metric of cognitive depth is not representational richness but irreversible accumulation.

22 Objections and Replies

22.1 Objection: This is just lexicographic preferences

One might argue that refusal is simply lexicographic ordering: the agent places a constraint (e.g. "do not harm") above all other objectives, and then maximizes utility subject to that constraint. On this view, nothing essentially new is introduced.

Lexicographic preferences still operate on a fixed option set. They rank options; they do not delete them. In Spherepop, refusal is deletion. The difference is not verbal. It matters for public legibility and for the persistence of exclusion under counterfactual changes. A lexicographic preference can, in principle, be revised without structural consequence; it is Collapse-compatible. A refusal is a history event: revising it requires a second event (Collapse or revocation, in the technical sense of these terms) that itself has a cost and a public meaning.

22.2 Objection: You are anthropomorphizing computation

The worry is that refusal is a moralized human interpretation of ordinary policy selection. The paper does not require that refusal be conscious or moral. It requires only a structural distinction: actions that remain physically available and instrumentally valuable are rendered inadmissible by an irreversible operator. This is a formal property of future spaces. Whether such operators can be implemented computationally is discussed below; the conceptual point is that the structure is not captured by standard optimization on fixed menus.

22.3 Objection: This requires libertarian free will

Nothing here implies uncaused choice. Spherepop events can be fully causally determined. The claim is about irreversibility, not metaphysical libertarianism. A refusal can be deterministic and still be a real closure operator with downstream causal consequences.

22.4 Objection: Refusal is epiphenomenal

Refusal does causal work precisely because it changes the topology of futures. In repeated social settings, a publicly legible refusal changes others' expectations and thereby changes equilibria. In the D.A.R.Y.L. case, the refusal restores roles and uncertainty; the social field changes. This is not merely a redescription of outcomes; it is a change in what others treat as live.

23 Open Problems

The framework developed here is intentionally minimal. It therefore raises sharp open problems.

23.1 Detection and "Turing tests" for refusal

Can genuine refusal be behaviorally distinguished from sophisticated simulation? In general, if an agent can perfectly mimic the outward behavior of refusal while retaining internal optionality, purely behavioral tests may fail. The natural direction is to seek *counterfactual invariants*: does the agent's inadmissibility persist under distribution shifts that would make the refused action locally attractive? This suggests a family of stress tests for AI systems: shift incentives and observe whether exclusions persist.

23.2 Composition of refusals

If an agent refuses α at t_1 and refuses β at t_2, what is the algebra of the combined constraint? In Spherepop, this is operator composition:

$$\mathcal{F}_2 = \mathbf{Refuse}(\beta)(\mathbf{Refuse}(\alpha)(\mathcal{F}_0)).$$

Deeper questions remain about commutativity, interference, and collapse: does a later collapse partially revoke earlier refusals? Under what formal conditions do refusals form a lattice of closures?

23.3 Degrees and revocation

Is refusal binary? Spherepop, as presented, treats $\mathbf{Refuse}(\alpha)$ as deleting all α-initiating futures. One can define partial refusals as probability-mass deletions (i.e. apply **Pop** to a graded predicate) or as bounded refusals (delete α for a time interval). Revocation then becomes a further event, not a free change of mind. A rigorous account needs a cost model for revocation: what is the penalty in trust, identity, or coordination?

23.4 Minimal architectures for refusal

What computational substrate supports genuine refusal? One hypothesis is that an agent must have a mechanism that can modify its own policy class (not just select within it) and protect that modification from later local re-optimization. This suggests architectural ingredients such as irrevocable policy editing, cryptographic commitment, or physically enforced action disabling. Whether refusal can be realized without suffering is an open ethical question: must commitments be backed by negative feedback, or can they be backed by structural locks?

23.5 Refusal and consciousness

What is the relationship between refusal and consciousness? The paper does not assume that refusal requires consciousness, but it is plausible that first-person experience is the natural site where "closure" is lived as such. If so, refusal may be a bridge concept linking agency, responsibility, and

phenomenology. Clarifying this would require engagement with philosophy of mind and empirical work on self-control and commitment.

24 Conclusion

Generative models reveal a genuine principle: sequences contain enough structure for coherent generation. Yet coherence is not commitment, and prediction is not participation.

The paper's central claim has been that refusal is a structurally distinct kind of cognitive act: an irreversible operation that eliminates admissible futures and binds an agent to a history. Spherepop formalizes this distinction by treating refusal as deletion in a branching future space, rather than as a mere ranking of fixed options. This allows us to explain why refusal is publicly legible as reliability, why it can shift equilibria, and why it is difficult to "strategy steal" without transformation.

The event algebra of Spherepop is a non-commutative, non-invertible monoid. Collapse is an idempotent quotient functor whose fixed-point category is the Karoubi completion of the event algebra—the maximally Collapsed world in which all historical distinctions have been erased and all remaining commitments are transparent. Genuine refusal is the structure that cannot be reached from within this completion: it is the obstruction to localization. Worldhood is the accumulated non-commutativity of the event sequence, and a world is what cannot be rearranged.

Three lenses have been shown to describe the same underlying phenomenon. The information-theoretic lens describes refusal as endogenous support restriction rather than Bayesian updating, deliberately increasing KL divergence from the prior trajectory distribution and preserving it as constraint. The operational lens describes refusal as future-space pruning by a non-invertible morphism, with Collapse as the unique expansion operator. The algebraic lens describes refusal as the accumulation of non-commuting morphisms in a monoid, producing depth that cannot be shortcut. These are not analogies but equivalent characterizations of closure.

The survival of value does not require winning every competition. It requires agents and institutions capable of irreversible exclusion—closures that preserve relationships, stabilize cooperation, and prevent optimization from dissolving the conditions of meaning.

Refusal is not a defect in intelligence. Prediction extends sequences; refusal creates worlds. It is one of the events by which intelligence enters a world.

References

[1] Elan Barenholtz. *Sequences Are All You Need: How Generative Modeling Drove the Evolution of Memory and Cognition*. Substack, Dec 18, 2025. `https://elanbarenholtz.substack.com/p/sequences-are-all-you-need`

[2] Joe Carlsmith. *Video and transcript of talk on "Can goodness compete?"* Substack, 2025. `https://joecarlsmith.substack.com/p/video-and-transcript-of-talk-on-can`

[3] Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

[4] Michael E. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, 1987.

[5] G. E. M. Anscombe. *Intention*. Harvard University Press, 1957.

[6] Donald Davidson. Actions, reasons, and causes. *The Journal of Philosophy*, 60(23):685–700, 1963.

[7] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11:127–138, 2010.

[8] Andy Clark. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, 2016.

[9] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[10] Stuart Russell. *Human Compatible*. Viking, 2019.

[11] Leonard J. Savage. *The Foundations of Statistics*. Wiley, 1954.

[12] Richard Jeffrey. *The Logic of Decision*. University of Chicago Press, 1983.

[13] Giulio Tononi. Consciousness as integrated information: a provisional manifesto. *Biological Bulletin*, 215(3):216–242, 2008.

[14] Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7):450–461, 2016.

[15] Lee Cronin, Sara I. Walker, et al. Identifying molecules as biosignatures with assembly theory and mass spectrometry. *Nature Communications*, 12:3033, 2021.

[16] Lee Cronin, Sara I. Walker, et al. Assembly theory explains and quantifies selection and evolution. *Nature*, 622:322–328, 2023.