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

Flyxion

December 20, 2025

1 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 that refusal cannot be represented as ordinary utility maximization on a fixed action set without smuggling event-history into the state description. This result underpins the later claims about "strategy stealing" and the competitive disadvantage of value-sustaining agents [2].

The argument proceeds in five stages. First, we introduce an illustrative case in which refusal is unusually clean: deliberate underperformance in the service of relationship (§2). Second, we characterize the autoregressive baseline and identify its structural bias toward optionality preservation (§3). Third, we develop Spherepop as a calculus with syntax, operational semantics, and worked examples (§4). Fourth, we distinguish refusal from preference change and related phenomena (§5), then prove a non-representability result (§6). Finally, we use these tools to revisit alignment taxes, competition, and worldhood as historically accumulated constraint (§8–§9), and we close by posing open problems about verification, composition, and implementation (§16).

2 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.

2.1 Counterfactual analysis

The counterfactual "continue playing perfectly" is not simply "win more." It is a different world. It produces a drift toward (i) competitive collapse (the game becomes an exhibition), (ii) relational flattening (others become spectators), and (iii) 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 (i) the admissible futures, (ii) the cost of eliminating them, and (iii) the public legibility of the resulting commitment.

3 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 (i) language models trained by next-token prediction, (ii) sequence models for video and audio trained by next-step prediction or denoising objectives, and (iii) 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 contradiction 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.

3.1 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 DARYL 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.

3.2 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.

4 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.

4.1 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:*

$$E : \mathcal{F}(h_t) \to \mathcal{F}'(h_t) \subset \mathcal{F}(h_t),$$

with $\mathcal{F}'(h_t) \neq \mathcal{F}(h_t)$.

Intuitively, an event is characterized by exclusion: it removes at least one admissible continuation. This exclusion is not mere preference; it is structural.

4.2 Formal syntax

We present a minimal syntax sufficient for composition.

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

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

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.

4.3 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.*

- $\textbf{skip}(\mathcal{F}) = \mathcal{F}$.
- $e_1; e_2(\mathcal{F}) = e_2(e_1(\mathcal{F}))$.
- $\textbf{Pop}(\varphi)(\mathcal{F}) = \{f \in \mathcal{F} : \neg\varphi(f)\}$*, i.e. remove futures matching predicate* φ.
- $\textbf{Refuse}(\alpha)(\mathcal{F}) = \{f \in \mathcal{F} : f \text{ does not begin with action-type } \alpha\}$.
- $\textbf{Bind}(\alpha \prec \beta)(\mathcal{F}) = \{f \in \mathcal{F} : \text{if } \beta \text{ occurs in } f \text{ then } \alpha \text{ occurs earlier}\}$.
- $\textbf{Collapse}(\psi)(\mathcal{F})$ *identifies or forgets distinctions among futures while preserving a coarser set of options; formally, it quotients* $\mathcal{F}$ *by an equivalence relation* $\sim_\psi$ *induced by predicate* ψ*, yielding a reduced representation in which some historical differentiation is erased.*

Two notes matter for later sections. First, **Refuse**(α) 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.

4.4 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:

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

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*:

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

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.

4.5 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 and supports accumulation into worldhood.

5 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 DARYL 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.

6 Formal Consequences and Proposition 1

We now return to the non-representability claim stated in earlier drafts and prove it under explicit assumptions.

6.1 Formal setup

Let $\mathcal{A}$ be a fixed, time-indexed action set. Let h_t be the agent's history up to time t. Let π be a policy mapping histories to distributions over actions. Standard decision theory represents rational choice by maximizing an expected utility functional:

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

where outcomes o depend on h_t and the chosen action.

To isolate the point, consider a simplified setting in which U is defined over immediate consequences and $\mathcal{A}$ is fixed across time.

6.2 What it means to "represent refusal"

We say that a utility representation *represents refusal* if it can distinguish (i) deliberate exclusion of an action-type that remains instrumentally high-value from (ii) ordinary preference change in which the action-type becomes low-value.

More precisely, suppose there is an action-type $\alpha \in \mathcal{A}$ that maximizes instrumental value for some salient evaluation functional V (e.g. winning). A refusal of α at time t is a policy constraint of the form:

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

That is: the policy assigns probability zero to α even though α is evaluatively optimal with respect to V.

6.3 Proposition and proof

Proposition 1 (Non-representability without history)**.** *Fix a time-indexed action set $\mathcal{A}$ that does not depend on history. Suppose an agent exhibits refusal of an action-type α at time t in the sense above. Then no utility function U defined solely over immediate outcomes and a fixed action set $\mathcal{A}$ can represent refusal without encoding event-history as a primitive state variable.*

Proof sketch. Assume for contradiction that refusal can be represented by maximizing expected utility $\mathbb{E}[U \mid h_t, a]$ on the fixed set $\mathcal{A}$, where U is defined without reference to refusal events. Since the agent refuses α, we must have that α is not utility-maximizing under U given h_t:

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

for some $a^* \neq \alpha$.

But refusal, as characterized, requires that α remain instrumentally high-value with respect to a salient evaluation functional V the agent continues to recognize (e.g. winning), and indeed that the agent's competence and knowledge about this remain intact. Therefore, if U makes α strictly suboptimal, U must be assigning disvalue to α's outcomes in a way that treats those outcomes as worse *in themselves* relative to U.

That move collapses refusal into preference change: it says the agent simply prefers a^* to α. Yet the refusal phenomenon is precisely the distinction between (i) treating α as bad and (ii) treating α as inadmissible despite its recognized instrumental value. In order to preserve that distinction inside a utility framework, the utility function must condition on whether a refusal event has occurred, i.e. on a historical variable $r_t \in \{0, 1\}$ indicating prior commitment. Put differently, U must be replaced by $U(\cdot; r_t)$ or the state must be augmented from h_t to (h_t, r_t). But r_t is exactly an event-history primitive. Contradiction. □

6.4 Concrete counterexample

Return to the baseball case. Let $\alpha = \mathit{Max}$, 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 winning.

Any attempt to represent this as ordinary utility maximization must either: (i) assign disvalue to winning when achieved via *Max* (thus redefining the agent's preferences so that refusal is no longer refusal), or (ii) condition the utility on the refusal event (thus importing event-history).

This counterexample is not about baseball; it is about the structural role of inadmissibility. When inadmissibility is real, it cannot be represented as a mere ranking over a fixed menu without treating commitments as state.

7 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.

7.1 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 $\mathcal{F}'(h_t) \subset \mathcal{F}(h_t)$. 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 $H'_t = H(P'_t)$.

7.2 Quantifying the cost of refusal

The entropy eliminated by an event can be quantified as:

$$\Delta H := H_t - H'_t.$$

For refusal, $\mathcal{F}'(h_t)$ deletes all futures beginning with action-type α. 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 not merely Shannon entropy; it 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.

7.3 Constraint conservation

The central interpretive claim of the paper 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. If one wishes to connect the two frameworks, refusal appears as a kind of "counter-homeostatic" act: a deliberate increase in model tension preserved as commitment rather than resolved by adaptation.

8 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.

8.1 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, assume:

- If two R agents meet, credible refusal supports cooperation, yielding payoff C each.
- If R meets O, O can exploit unless R's refusal is publicly legible and binding; call the payoff to R in such encounters E_R and to O E_O.
- If 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.

9 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 $E_1, \dots, E_n$ 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 $\mathcal{F}_i = \mathcal{F}_{i-1} \setminus F_i$ (ignoring collapse quotients for the moment). 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 DARYL 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.

10 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.

10.1 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.

10.2 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.

10.3 Interpretation

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. 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.

This will prove decisive in distinguishing genuine refusal from its revocation.

11 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.

11.1 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'.

11.2 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

$$\mathcal{F}(h \cdot r \cdot v) = \mathcal{F}(h).$$

But this equality cannot hold unless the histories h and $h \cdot r \cdot v$ are equivalent under Collapse. That is, revocation requires:

$$h \sim_C h \cdot r \cdot v.$$

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.

11.3 The Asymmetry Result

Proposition 2 (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.

However, future space expansion is impossible under event composition alone; Spherepop primitives other than Collapse only restrict or order futures. 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.

12 Fake Refusal and the Simulation of Commitment

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

12.1 Definition of Fake Refusal

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.

12.2 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.

12.3 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.

12.4 Return to the Illustrative Case

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.

13 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. In particular, it permits a principled criterion for halting grounded not in syntactic completion or reward convergence, but in the exhaustion of meaningful transformation.

13.1 Uncertainty as Residual Distinguishability

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

Definition 8 (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.

13.2 Transformations and Stabilization

Let $\mathcal{T}$ be a set of admissible transformations on histories (e.g., Pop, Bind, Refuse, Collapse). Consider an iterative process applying transformations to a history:

$$h_{n+1} = T_n(h_n), \quad T_n \in \mathcal{T}.$$

We are interested in whether continued application of transformations yields genuinely new structure.

Definition 9 (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.

13.3 A Spherepop Halting Criterion

This yields a halting criterion internal to Spherepop:

Definition 10 (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 $m \geq N$,*

$$[T_m \circ \cdots \circ T_1(h_k)]_C = [h_k]_C.$$

That is, 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 convergence of output, but halting by exhaustion of world-relevant change.

13.4 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.

13.5 Why Autoregressive Systems Cannot Halt This Way

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, token limits, or optimization thresholds, but never because the space of admissible futures has been exhausted. They do not halt; they are halted.

Spherepop systems halt because the world has closed.

13.6 Connection to Refusal

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.

This provides a formal sense in which refusal is computationally efficient: it trades optionality for decisiveness, prediction for commitment, and infinite continuation for closure.

Refusal does not merely constrain action. It makes halting possible.

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

The stabilization-based halting criterion introduced above 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. On a standard reading, an agent *ought* to reduce the mismatch between its generative model and the sensory stream, thereby keeping itself on a high-probability manifold of expected observations [7]. 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.

14.1 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:

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

so 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 for our purposes.

First, free-energy minimization encourages *error correction*: when the agent encounters divergence between prediction and sensation, it can reduce F 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 (e.g. 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.

14.2 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 (action-observation sequences) and let $p(\tau)$ denote the agent’s prior over trajectories induced by its generative model and policy. 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 is a prior that sets:

$$p(\tau) = 0 \quad \text{for all } \tau \text{ in the refused class.}$$

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.

In active inference language, refusal is therefore not well-modeled as introducing a new term in the expected free energy objective. It is better modeled as changing the support of the distribution over trajectories.

14.3 Free Energy vs. World-Binding: A Structural Tension

Once refusal is expressed as a support restriction, an immediate tension with free-energy minimization emerges. If refusal deletes trajectories that are instrumentally powerful or prediction-confirming, then, relative to the pre-refusal model, it may *increase* expected free energy. In the DARYL case, maximal play yields predictable outcomes and high control; refusing it introduces uncertainty and potential error.

Thus, refusal appears as a deliberate willingness to incur predictive cost for the sake of social world-stability. This matches the paper's earlier information-theoretic characterization: refusal is an intentional KL "spike" relative to the prior trajectory distribution, preserved as constraint rather than resolved as noise.

The important point is that this does not refute active inference. It clarifies where the burden shifts. If one wants refusal to be compatible with free-energy minimization, one must represent the social invariant (e.g. preserving relational roles) *inside the generative model* such that the refusal reduces free energy in an expanded model class. Put bluntly: refusal can be free energy minimizing only if the world-model contains the right notion of worldhood.

14.4 Refusal as Model-Expansion Rather Than Outcome-Preference

This suggests a useful diagnostic distinction.

Remark 2. *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.*

Active inference can therefore accommodate refusal only by treating it as structural: a modification of the model's admissibility structure rather than a preference reweighting within a fixed admissibility structure.

14.5 Stabilization as a Free-Energy Fixed Point

We can now connect the Spherepop halting criterion to free-energy dynamics.

Let $\mathcal{T}$ be the set of admissible Spherepop transformations (Pop, Bind, Refuse, Collapse), and let $\sim_C$ be the Collapse equivalence relation. Recall that stabilization at history h was defined by:

$$[h]_C = [T(h)]_C \quad \text{for all } T \in \mathcal{T}.$$

In active inference terms, a stabilized history corresponds to a regime in which available transformations no longer produce distinguishable predictive consequences. Put differently, 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 (i) redundant (they change only Collapse-irrelevant detail) or (ii) inadmissible (they violate refusal constraints). Stabilization is therefore interpretable as convergence not merely of beliefs but of admissibility: the agent's world has become invariant under its own event algebra.

14.6 Autoregressive Agents as Anti-Stabilizers

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. Their learning dynamics are naturally aligned with minimizing local surprise, not with enforcing hard deletions of trajectory support. Even when they emulate refusal, the emulation typically lives in the weights of a soft distribution, not in a support restriction.

In this sense, Spherepop refusal is an architectural commitment to non-Markovian constraint accumulation, whereas autoregressive cognition is an architectural commitment to indefinite extensibility.

14.7 A Synthesis: Refusal as Active Inference Over Social Invariants

A constructive way to integrate these perspectives is to treat 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 DARYL 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.

15 Objections and Replies

15.1 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.

Reply: 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. A refusal is a history event: revising it requires a second event (e.g. collapse or revocation) that itself has a cost and a public meaning.

15.2 Objection: You are anthropomorphizing computation

The worry is that refusal is a moralized human interpretation of ordinary policy selection.

Reply: 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.

15.3 Objection: This requires libertarian free will

Reply: 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.

15.4 Objection: Refusal is epiphenomenal

Reply: 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 DARYL 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.

16 Open Problems

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

16.1 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, then 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.

16.2 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:

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

But 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?

16.3 Degrees and revocation

Is refusal binary? Spherepop, as presented, treats $\mathbf{Refuse}(\alpha)$ as deleting all α-initiating futures. But 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?

16.4 Minimal architectures for refusal

What computational substrate supports genuine refusal? One hypothesis is that an agent must have a mechanism that can (i) modify its own policy class, not just select within it, and (ii) 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?

16.5 Refusal and consciousness

Finally, 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.

17 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 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. 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.