Boundary Conditions for Reward Maximisation

Axiom Violations as Structural Interface Constraints

Duston Moore, PhD
Independent Researcher
Edmonton, Alberta, Canada
dhwcmoore@gmail.com

January 2026

Abstract

Recent work has given a sharp characterisation of when goals can be represented as maximisation of expected discounted Markov reward: a preference relation over distributions on histories admits such a representation if and only if five axioms hold. This paper develops the contrapositive reading of that result. We show that violations of these axioms correspond to *structural boundary conditions*: admissibility constraints on trajectories that cannot be expressed as local properties of state–action transitions and therefore cannot be fully internalised by reward maximisation alone. We provide formal proofs that (i) violations of Independence induce counterfactual boundary predicates requiring reasoning about excluded alternatives, (ii) violations of Temporal γ-Indifference induce history-dependent non-Markovian boundaries, and (iii) certain certification requirements are invisible to probabilistic optimisation due to measure–partition mismatch. These results establish an *Axiom-Boundary Correspondence*: the reward hypothesis holds precisely when no boundary conditions are present. The correspondence motivates an architectural separation between productive optimisation and external boundary enforcement. We conclude with a minimal counterexample environment demonstrating that no scalar reward can encode the required boundary condition, while a trivial external certifier succeeds. The results provide formal justification for neuro-symbolic safety architectures and clarify the structural limits of reward-based approaches to AI alignment.

Keywords: reward hypothesis, boundary conditions, axiom violations, safe reinforcement learning, production-closure separation, Markov reward, neuro-symbolic AI, AI safety

1 Introduction

The reward hypothesis, as articulated by Sutton [21], claims that "all of what we mean by goals and purposes can be well thought of as maximisation of the expected value of the cumulative sum of a received scalar signal (reward)." This hypothesis underwrites the foundational assumption of reinforcement learning: that an agent optimising expected return will, in the limit, achieve any goal we might specify. If the hypothesis holds universally, then the problem of AI alignment reduces to the problem of specifying the correct reward function.

Recent work by Bowling, Martin, Abel, and Dabney [7] has transformed this informal claim into a precise mathematical theorem. Their *Markov Reward Theorem* establishes that a preference relation over distributions on histories admits a Markov reward representation if and only if five axioms are satisfied: Completeness, Transitivity, Independence, Continuity, and Temporal γ-Indifference. When the axioms hold, there exists an equivalent expected discounted Markov reward representation; when any axiom fails, no such representation exists.

This paper takes the failure cases seriously. We argue that axiom violations are not merely inconveniences for reward design but signal a deeper representational mismatch. Specifically, they indicate the presence of *boundary conditions*: global admissibility constraints on trajectories that cannot be reduced to local transition evaluation. An agent optimising scalar reward may achieve arbitrarily high expected return while remaining systematically blind to these boundaries.

The significance for AI safety is immediate. If a goal specification violates the reward axioms, then no amount of reward engineering, no sophistication of the optimisation algorithm, and no quantity of training data can guarantee that the resulting agent will satisfy the goal. The axiom violation introduces a structural gap between what reward maximisation can express and what the goal requires. This gap cannot be closed from within the reward-maximisation framework; it requires external certification mechanisms that operate on information inaccessible to the reward signal.

1.1 Contributions

Our contribution is fivefold:

(i) We give a precise definition of *boundary conditions* in the reward-hypothesis setting and distinguish them from transition-local properties that can be encoded in scalar reward (Definition 4.4).

(ii) We prove that violations of **Independence** induce *counterfactual boundary predicates* that depend on excluded alternatives rather than realised transitions (Theorem 5.1).

(iii) We prove that violations of **Temporal γ-Indifference** induce *history-dependent boundaries* that cannot be captured by any Markov state representation (Theorem 5.3).

(iv) We establish a **Measure–Partition Mismatch Theorem** showing that certain boundary conditions are *probabilistically blind*: invisible to any certifier that operates within the agent's information structure (Theorem 5.7).

(v) We construct a **minimal counterexample environment** demonstrating that no reward-based agent can satisfy the boundary condition, whereas an external certifier succeeds trivially (Section 7).

Together, these results establish what we call the *Axiom-Boundary Correspondence*: the reward hypothesis holds for a goal specification if and only if the specification contains no boundary conditions requiring external certification. This correspondence provides formal justification for architectures that separate productive optimisation from external boundary enforcement, and clarifies the structural limits of reward-based approaches to safety.

1.2 Outline

Section 2 reviews the formal framework of preferences over histories and states the axioms underlying the Markov Reward Theorem. Section 3 surveys related work in safe reinforcement learning, risk-sensitive control, and neuro-symbolic AI. Section 4 defines boundary conditions and distinguishes them from transition-local properties. Section 5 proves the main theorems connecting axiom violations to boundary conditions. Section 6 develops the architectural consequences. Section 7 presents the minimal counterexample. Section 8 considers implications and limitations. Section 9 concludes.

2 Background: The Reward Hypothesis and Its Axioms

2.1 Histories and Preferences

Let $\mathcal{O}$ be a finite set of observations and $\mathcal{A}$ a finite set of actions. A *transition* is a pair $t = (o, a) \in \mathcal{O} \times \mathcal{A}$. Let $\mathcal{T} = \mathcal{O} \times \mathcal{A}$ denote the set of all transitions.

Definition 2.1 (History)**.** A *history* is a finite sequence of transitions $h = t_1 t_2 \cdots t_n$ where each $t_i \in \mathcal{T}$. Let $\mathcal{H}_n = \mathcal{T}^n$ denote histories of length n, and let $\mathcal{H} = \bigcup_{n=0}^{\infty} \mathcal{H}_n$ denote the set of all finite histories, with ε denoting the empty history of length zero.

For a history $h \in \mathcal{H}$ and a transition $t \in \mathcal{T}$, we write $t \cdot h$ for the history with t prepended to h. Let $\Delta(\mathcal{H})$ denote the set of probability distributions with finite support over $\mathcal{H}$. For $A, B \in \Delta(\mathcal{H})$ and $p \in [0, 1]$, let $pA + (1 - p)B$ denote the mixture distribution that samples from A with probability p and from B with probability $1 - p$.

Definition 2.2 (Preference Relation)**.** A *goal* is modelled as a binary preference relation $\succsim$ over $\Delta(\mathcal{H})$. We write $A \succ B$ for strict preference ($A \succsim B$ and not $B \succsim A$) and $A \sim B$ for indifference ($A \succsim B$ and $B \succsim A$).

The preference relation represents the designer's or agent's ranking of possible outcomes. We do not assume preferences are derived from an underlying utility function; the question is precisely when such a representation exists.

2.2 Markov Reward Representation

Definition 2.3 (Markov Reward Representation)**.** A preference relation $\succsim$ on $\Delta(\mathcal{H})$ admits a *Markov reward representation* if there exist functions $r : \mathcal{T} \to \mathbb{R}$ (reward) and $\gamma : \mathcal{T} \to [0, 1]$ (discount) such that for all $A, B \in \Delta(\mathcal{H})$:

$$A \succsim B \iff U(A) \geq U(B)$$

where the utility $U : \Delta(\mathcal{H}) \to \mathbb{R}$ is defined by:

$$U(A) = \mathbb{E}_{h \sim A}\left[\sum_{i=1}^{|h|} \left(\prod_{j=1}^{i-1} \gamma(t_j)\right) r(t_i)\right]$$

with the convention that an empty product equals 1.

The critical feature is that the reward $r(t)$ and discount $\gamma(t)$ depend only on the transition t, not on the history preceding it. This *Markov property* is what makes the representation tractable for reinforcement learning algorithms that maintain only a state sufficient to predict future rewards.

Definition 2.4 (Utility Decomposition)**.** Under a Markov reward representation, the utility of a history $h = t_1 t_2 \cdots t_n$ decomposes recursively:

$$u(t \cdot h) = r(t) + \gamma(t) \cdot u(h), \qquad u(\varepsilon) = 0$$

This yields $u(h) = \sum_{i=1}^{n} \gamma_1 \cdots \gamma_{i-1} \cdot r(t_i)$ where $\gamma_j = \gamma(t_j)$.

2.3 The Axioms

Bowling et al. [7] establish that a Markov reward representation exists if and only if the preference relation satisfies the following five axioms.

Axiom 1 (Completeness)**.** For all $A, B \in \Delta(\mathcal{H})$, either $A \succsim B$ or $B \succsim A$ (or both).

Completeness requires that the preference ordering makes some judgment about any pair of distributions. Note that this includes the possibility of indifference ($A \sim B$), which is distinct from incomparability.

Axiom 2 (Transitivity)**.** For all $A, B, C \in \Delta(\mathcal{H})$, if $A \succsim B$ and $B \succsim C$, then $A \succsim C$.

Transitivity ensures that preferences form a coherent ordering without cycles.

Axiom 3 (Independence)**.** For all $A, B, C \in \Delta(\mathcal{H})$ and $p \in (0, 1)$:

$$A \succsim B \iff pA + (1-p)C \succsim pB + (1-p)C$$

Independence requires that preferences between distributions are invariant under mixing with a common third distribution. Intuitively, how you rank two options should not depend on what other options happen to be available.

Axiom 4 (Continuity)**.** For all $A, B, C \in \Delta(\mathcal{H})$ with $A \succsim B \succsim C$, there exists $p \in [0, 1]$ such that:

$$pA + (1-p)C \sim B$$

Continuity ensures the existence of a "break-even point" when mixing more-preferred and less-preferred options.

Axiom 5 (Temporal γ-Indifference)**.** There exists a function $\gamma : \mathcal{T} \to [0, 1]$ such that for all $A, B \in \Delta(\mathcal{H})$ and transitions $t \in \mathcal{T}$:

$$\frac{1}{\gamma(t)+1}(t \cdot A) + \frac{\gamma(t)}{\gamma(t)+1}B \sim \frac{1}{\gamma(t)+1}(t \cdot B) + \frac{\gamma(t)}{\gamma(t)+1}A$$

Temporal γ-Indifference requires that prepending the same transition to two distributions rescales their preference difference by a fixed factor $\gamma(t)$ that depends only on the transition, not on the distributions themselves.

Theorem 2.5 (Markov Reward Theorem, Bowling et al. 2023)**.** *A binary preference relation on $\Delta(\mathcal{H})$ satisfies Axioms 1–5 if and only if there exist functions $u : \Delta(\mathcal{H}) \to \mathbb{R}$, $r : \mathcal{T} \to \mathbb{R}$, and $\gamma : \mathcal{T} \to [0, 1]$, such that $u(\varepsilon) = 0$ and:*

$$u(t \cdot h) = r(t) + \gamma(t) \cdot u(h)$$

where r is unique up to positive scaling and γ is the function satisfying Axiom 5.

The theorem is an "if and only if": satisfaction of the axioms is both necessary and sufficient for Markov reward representation. Our focus is on the necessity direction and its contrapositive.

Corollary 2.6 (Contrapositive)**.** *If any of Axioms 1–5 is violated, then no Markov reward representation exists.*

3 Related Work

3.1 Safe Reinforcement Learning

Safe reinforcement learning (Safe RL) addresses the problem of learning policies that achieve good performance while respecting safety constraints during learning and deployment. García and Fernández [10] provide a comprehensive survey, distinguishing approaches that modify the optimality criterion (such as risk-sensitive objectives) from approaches that constrain exploration or incorporate external knowledge.

Constrained Markov decision processes (CMDPs) optimise expected return subject to constraints on expected cumulative costs [4]. Constrained Policy Optimisation (CPO) [2] is a prominent deep-RL algorithm designed for CMDPs, providing approximate constraint satisfaction during training via trust-region style updates. Safety layers and shielding mechanisms alter or veto proposed actions to satisfy constraints, especially in continuous control [9, 3].

Bowling et al. [7] demonstrate that constrained MDPs violate the Independence axiom and can violate Continuity. This is not an implementation limitation but a structural incompatibility with reward maximisation. Our results explain why safety architectures keep reappearing: if the goal violates Bowling et al.'s axioms, then safety constraints are structurally required to express the goal, not merely useful engineering additions.

3.2 Risk-Sensitive Control

Risk-sensitive reinforcement learning augments expected return with a risk measure, such as variance penalties, value-at-risk, or conditional value-at-risk (CVaR). Tamar et al. [23] study policy gradients for coherent risk measures and distinguish static from time-consistent dynamic risk measures. Chow et al. [8] study percentile and CVaR-constrained MDPs in depth.

A central issue in risk-sensitive sequential decision-making is time inconsistency: a risk objective defined over full trajectories can induce preferences that are not preserved under conditioning on partial histories. Bowling et al. [7] show that such objectives can violate Temporal γ-Indifference, forcing history dependence. The implication for safety is that a Markov agent using only current state variables cannot, in general, represent the relevant risk boundary, because the risk boundary is a predicate over the accumulated trajectory.

3.3 Neuro-Symbolic AI and Formal Verification

Neuro-symbolic AI [11] combines neural learning with symbolic reasoning. Our results provide formal justification for when symbolic components are necessary: when the goal specification violates the reward axioms and therefore requires external boundary certification.

The production-closure separation has been studied in formal verification [16, 13]. The LCF architecture separates tactics (productive) from kernels (certifying). Our contribution is connecting this architectural pattern to the foundations of reinforcement learning through the Axiom-Boundary Correspondence.

3.4 AI Safety and Goal Mis-specification

The relationship between reward design and goal specification has been studied extensively [12, 5, 17]. Reward hacking [20] and specification gaming [14] arise when agents exploit gaps between reward and intended goals. Our results provide a structural diagnosis: these problems arise when axiom-violating goals are treated as if they satisfy the reward hypothesis.

4 Boundary Conditions

We now formalise the notion of a boundary condition in the context of preference relations over histories.

4.1 Transition-Local Properties

Definition 4.1 (Transition Statistics)**.** The *transition statistics* of a distribution $A \in \Delta(\mathcal{H})$ is the function $\sigma_A : \mathcal{T} \to \mathbb{R}_{\geq 0}$ defined by:

$$\sigma_A(t) = \mathbb{E}_{h \sim A} \left[\#\{i : t_i = t\} \right]$$

where $\#\{i : t_i = t\}$ counts occurrences of transition t in history h.

Transition statistics capture the expected frequency of each transition under a distribution, aggregated across all positions in the history.

Definition 4.2 (Transition-Local Property)**.** A property $P : \Delta(\mathcal{H}) \to \{0, 1\}$ is *transition-local* if there exists a function $\phi : \mathcal{T} \to \mathbb{R}$ and a threshold $\theta \in \mathbb{R}$ such that for all $A \in \Delta(\mathcal{H})$:

$$P(A) = 1 \iff \sum_{t \in \mathcal{T}} \phi(t) \cdot \sigma_A(t) \geq \theta$$

Equivalently, P is transition-local if it depends only on the transition statistics of A, not on the ordering of transitions within histories or on counterfactual information about what alternatives were available.

Proposition 4.3. *Any property definable by a threshold on expected Markov reward is transition-local.*

Proof. Under a Markov reward representation with reward r and discount γ, expected utility is:

$$U(A) = \mathbb{E}_{h \sim A} \left[\sum_{i=1}^{|h|} \gamma_1 \cdots \gamma_{i-1} \cdot r(t_i) \right]$$

This can be rewritten as a function of position-weighted transition statistics, which refines the basic transition statistics. The key point is that $U(A)$ depends only on which transitions occur and with what frequency at each position, not on counterfactual information or global trajectory structure beyond what is captured by discounted accumulation. □

4.2 Boundary Conditions

Definition 4.4 (Boundary Condition)**.** A *boundary condition* is a predicate $B : \Delta(\mathcal{H}) \to \{0, 1\}$ that is **not** transition-local. Equivalently, there exist distributions $A, A' \in \Delta(\mathcal{H})$ with identical transition statistics ($\sigma_A = \sigma_{A'}$) such that $B(A) \neq B(A')$.

Remark 4.5. The terminology "boundary condition" is deliberate. In partial differential equations, boundary conditions constrain solutions at the edges of a domain; they cannot be derived from the differential equation alone. Similarly, boundary conditions in our sense constrain admissible trajectories in ways that cannot be derived from transition-local reward accumulation.

Example 4.6 (Counterfactual Boundary)**.** Consider the predicate: "The agent chose action a when action b was also available." This depends not on the transitions that occurred but on what alternatives existed at the choice point. Two trajectories with identical transition sequences can differ in whether the alternative was available, hence this is a boundary condition.

Example 4.7 (History-Dependent Boundary)**.** Consider the predicate: "The cumulative exposure to risk never exceeded threshold τ at any prefix." This depends on the ordering of transitions, not merely on their aggregate. Two trajectories with the same total risk exposure can differ in whether the threshold was exceeded at some intermediate point.

Example 4.8 (Partition Membership)**.** Consider a classification of trajectories into "safe" and "unsafe" based on criteria not captured by the agent's state representation. If the partition is not measurable with respect to the agent's information σ-algebra, then no function of the agent's observations can certify membership.

4.3 Boundary-Constrained Goals

Definition 4.9 (Boundary-Constrained Goal)**.** A preference relation $\succsim$ on $\Delta(\mathcal{H})$ is *boundary-constrained* if there exists a boundary condition B such that: for all $A, A' \in \Delta(\mathcal{H})$, if $B(A) = 0$ and $B(A') = 1$, then $A' \succ A$ regardless of transition-local properties.

A boundary-constrained goal gives lexicographic priority to satisfying the boundary condition over maximising any transition-local objective. The boundary acts as a hard constraint that cannot be traded off against reward.

Proposition 4.10. *If a preference relation is boundary-constrained, it admits no Markov reward representation.*

Proof. Suppose $\succsim$ is boundary-constrained by B. By definition of boundary condition, there exist A, A' with $\sigma_A = \sigma_{A'}$ and $B(A) \neq B(A')$. Suppose without loss of generality that $B(A) = 0$ and $B(A') = 1$. Then $A' \succ A$ by the boundary constraint.

If a Markov reward representation existed, the utility would depend only on transition statistics (by Proposition 4.3). Since $\sigma_A = \sigma_{A'}$, we would have $U(A) = U(A')$, implying $A \sim A'$. This contradicts $A' \succ A$. □

5 Axiom Violations Imply Boundary Conditions

We now establish our main results: that specific axiom violations necessarily induce boundary conditions.

5.1 Independence Violations

Independence (Axiom 3) requires that preferences between distributions are invariant under mixing with a common third distribution. Intuitively, how you rank two options should not depend on what other options are available.

Theorem 5.1 (Independence Violation ⇒ Counterfactual Boundary)**.** *If a preference relation $\succsim$ violates Independence, then satisfying $\succsim$ requires evaluating a boundary condition that depends on excluded alternatives.*

Proof. Suppose Independence is violated. Then there exist $A, B, C \in \Delta(\mathcal{H})$ and $p \in (0, 1)$ such that $A \succsim B$ but:

$$pB + (1-p)C \succ pA + (1-p)C$$

Define the boundary condition $B_{\text{cf}} : \Delta(\mathcal{H}) \to \{0, 1\}$ as follows. Consider a "menu" $\mathcal{M} \subseteq \Delta(\mathcal{H})$ of available options. Define:

$$B_{\text{cf}}(D; \mathcal{M}) = \begin{cases} 1 & \text{if } D \text{ was selected from menu } \mathcal{M} \text{ with } C \notin \mathcal{M} \\ 0 & \text{otherwise} \end{cases}$$

This predicate depends on counterfactual information: whether C was available when D was selected.

Now consider two scenarios:

1. The agent faces menu $\{A, B\}$ and selects A. Since $A \succsim B$, this is preferred.
2. The agent faces menu $\{pA+(1-p)C, pB+(1-p)C\}$. The preference reverses: $pB+(1-p)C$ is now preferred.

Conditional on realising a sample from A (which occurs with probability p in scenario 2), the transition statistics are identical in both scenarios. The difference is entirely in what alternatives were available.

Now suppose a Markov reward representation existed. Expected utility under Markov reward is linear in mixtures:

$$U(pA + (1-p)C) = p \cdot U(A) + (1-p) \cdot U(C)$$

Therefore $A \succsim B$ implies $U(A) \geq U(B)$, which implies:

$$p \cdot U(A) + (1-p) \cdot U(C) \geq p \cdot U(B) + (1-p) \cdot U(C)$$

Hence $pA + (1-p)C \succsim pB + (1-p)C$, contradicting the assumed preference reversal.

Therefore no Markov reward can represent $\succsim$. The preference reversal depends on the availability of alternative C, which is counterfactual information not captured by any transition-local statistic. □

Corollary 5.2. *Constrained MDPs, which violate Independence [7], induce counterfactual boundary conditions that cannot be encoded in scalar reward.*

Proof. Bowling et al. demonstrate that constrained MDPs violate Independence because the preference between trajectories depends on what other trajectories are feasible under the constraint. Adding a mixture with a high-cost trajectory can flip preferences by making the constraint binding. Apply Theorem 5.1. □

5.2 Temporal γ-Indifference Violations

Temporal γ-Indifference (Axiom 5) requires that prepending the same transition to two distributions rescales their preference difference by a fixed factor $\gamma(t)$. Violations arise when the value of a trajectory depends on its history in ways that cannot be captured by discounting.

Theorem 5.3 (Temporal γ-Indifference Violation $\Rightarrow$ History-Dependent Boundary)**.** *If a preference relation $\succsim$ violates Temporal γ-Indifference, then satisfying $\succsim$ requires evaluating a boundary condition that depends on the full history prefix.*

Proof. Suppose Temporal γ-Indifference is violated. Then for some transition t and distributions $A, B \in \Delta(\mathcal{H})$, there is no $\gamma(t) \in [0, 1]$ such that:

$$\frac{1}{\gamma(t)+1}(t \cdot A) + \frac{\gamma(t)}{\gamma(t)+1}B \sim \frac{1}{\gamma(t)+1}(t \cdot B) + \frac{\gamma(t)}{\gamma(t)+1}A$$

This means the preference difference between $t \cdot A$ and $t \cdot B$ is not a fixed multiple of the preference difference between A and B. The scaling depends on the content of A and B, not merely on the prepended transition t.

Define the boundary condition $B_{\text{hist}} : \Delta(\mathcal{H}) \to \{0, 1\}$ as follows. For a history $h = t_1 \cdots t_n$, let $\text{prefix}_k(h) = t_1 \cdots t_k$ denote the length-k prefix. Define:

$$B_{\text{hist}}(D) = 1 \iff \forall h \in \text{supp}(D),\ \phi(\text{prefix}(h)) \text{ satisfies constraint } \Phi$$

where ϕ and Φ are determined by the structure of the axiom violation.

Now suppose a Markov reward representation existed with reward r and discount γ. Then:

$$u(t \cdot h) = r(t) + \gamma(t) \cdot u(h)$$

for all histories h. This implies:

$$u(t \cdot A) - u(t \cdot B) = \gamma(t) \cdot (u(A) - u(B))$$

This is precisely the linear scaling required by Temporal γ-Indifference. Therefore violation of the axiom implies non-representability.

The violation manifests as history-dependence: the contribution of future transitions to overall utility depends on what has already occurred. This dependence defines a boundary predicate that cannot be evaluated from the current state alone; it requires access to the full history prefix. □

Corollary 5.4. *Risk-sensitive objectives such as CVaR, which violate Temporal γ-Indifference [7], induce history-dependent boundary conditions.*

Example 5.5 (Variance-Penalised Return)**.** Consider maximising $\mathbb{E}[R] - \lambda \cdot \mathrm{Var}(R)$ for cumulative return R. The optimal policy is history-dependent: if early returns are high, the agent should take less risk to lock in gains; if early returns are low, the agent might take more risk to compensate. This dependence on realised history violates Temporal γ-Indifference and induces a boundary condition on admissible risk profiles that cannot be evaluated from the current state.

5.3 Probabilistic Blindness and Measure–Partition Mismatch

Even if an agent attempts to approximate boundary conditions probabilistically, there are structural limits to what probabilistic methods can certify.

Definition 5.6 (Information Structure)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space where Ω is the sample space of possible trajectories, $\mathcal{F}$ is a σ-algebra of measurable events, and $\mathbb{P}$ is a probability measure. An agent's *information structure* is characterised by a sub-σ-algebra $\mathcal{F}_{\text{agent}} \subseteq \mathcal{F}$ representing the events distinguishable by the agent.

Theorem 5.7 (Measure–Partition Mismatch)**.** *Let $\Pi = \{C_1, \ldots, C_n\}$ be a partition of Ω required for boundary certification. If some $C_k \notin \mathcal{F}_{agent}$, then no $\mathcal{F}_{agent}$-measurable function can certify membership in C_k with certainty.*

Proof. A certifier implementable by the agent must be a function $f : \Omega \to \{0, 1\}$ that is $\mathcal{F}_{\text{agent}}$-measurable. Certification of membership in C_k requires $f^{-1}(1) = C_k$.

If $C_k \notin \mathcal{F}_{\text{agent}}$, then C_k is not measurable with respect to the agent's information structure. By definition of measurability, no $\mathcal{F}_{\text{agent}}$-measurable function can have C_k as a preimage.

In particular, any measurable approximation $E \in \mathcal{F}_{\text{agent}}$ to C_k satisfies $E \neq C_k$. The symmetric difference $\Delta = E \triangle C_k = (E \setminus C_k) \cup (C_k \setminus E)$ is non-empty. Elements of $C_k \setminus E$ will be incorrectly rejected (false negatives); elements of $E \setminus C_k$ will be incorrectly accepted (false positives).

Crucially, probabilistic updating of $\mathbb{P}(E \mid \text{data})$ cannot eliminate Δ, since the distinction between C_k and E is not representable in $\mathcal{F}_{\text{agent}}$. The agent may converge to arbitrarily high confidence in E while remaining structurally incapable of certifying C_k. □

Definition 5.8 (Probabilistic Blindness)**.** An agent exhibits *probabilistic blindness* with respect to a boundary condition B if B defines a partition element not in the agent's σ-algebra. The agent cannot detect violations of B regardless of experience, model capacity, or optimisation quality.

Remark 5.9. Probabilistic blindness is not a limitation of sample complexity, model capacity, or optimisation quality. It is a *representational* limitation: the distinction required for certification is not expressible in the agent's information structure. No amount of data can overcome a missing distinction.

Corollary 5.10. *If a boundary condition B induces probabilistic blindness, then no learning algorithm operating within the agent's information structure can learn to certify B.*

5.4 The Axiom-Boundary Correspondence

We can now state our main correspondence result.

Theorem 5.11 (Axiom-Boundary Correspondence)**.** *Let $\succsim$ be a preference relation on $\Delta(\mathcal{H})$. The following are equivalent:*

$\succsim$ satisfies Axioms 1–5.

There exists a Markov reward function r and discount γ such that $\succsim$ is represented by expected discounted cumulative reward.

The goal specified by $\succsim$ contains no boundary conditions requiring external certification.

Proof. The equivalence (i) $\Leftrightarrow$ (ii) is the Markov Reward Theorem (Theorem 2.5).

(ii) $\Rightarrow$ (iii): If $\succsim$ is represented by Markov reward, then the utility of any distribution depends only on transition statistics (Proposition 4.3). No structural relationships between histories affect utility beyond what is captured by discounted accumulation. Therefore there are no boundary conditions requiring external certification; local reward maximisation suffices.

(iii) $\Rightarrow$ (ii): We prove the contrapositive. Suppose $\succsim$ is not representable by Markov reward. By the Markov Reward Theorem, some axiom is violated. By Theorems 5.1 and 5.3, axiom violations introduce boundary conditions. These boundary conditions require reasoning about counterfactual alternatives or history prefixes that a Markov agent cannot represent. Therefore external certification is required. □

6 Architectural Consequences

The Axiom-Boundary Correspondence motivates a fundamental architectural principle for safe AI systems.

6.1 Production-Closure Separation

Definition 6.1 (Productive Operation)**.** A *productive operation* is a (possibly partial) function $f : S \to S$ on a state space S that transforms states without guaranteeing termination or completeness. Productive operations preserve openness: they generate successors indefinitely without certifying when the process should stop.

Policy optimisation in reinforcement learning is a productive operation. It generates candidate policies and refines them toward higher expected return. The question is whether the closure condition (satisfaction of the full goal specification) can be internalised within this productive process.

Definition 6.2 (Closure Operation)**.** A *closure operation* is a predicate $C : S \to \{0, 1\}$ such that $C(s) = 1$ indicates that s satisfies a completeness or finality condition. Closure operations determine when productive exploration may soundly terminate.

Theorem 6.3 (Production-Closure Separation)**.** *If a goal specification induces boundary conditions through axiom violations, then no finite composition of productive optimisation steps can internally certify satisfaction of those boundaries.*

Algorithm 1 Sealing Protocol Execution

Input: State s, productive core $\mathcal{P}$, boundary module $\mathcal{B}$
$a \leftarrow \mathcal{P}.\textsc{Propose}(s)$ ▷ Productive core proposes action
$\text{ctx} \leftarrow \textsc{GetContext}(s)$ ▷ Gather boundary-relevant context
if $\mathcal{B}.\textsc{Certify}(a, \text{ctx}) = \textsc{Accept}$ **then**
    return $\textsc{Execute}(a)$ ▷ Action is sealed and executed
else
    $\mathcal{P}.\textsc{Feedback}(\textsc{Reject}, a)$ ▷ Inform productive core
    return $\textsc{Fallback}(s)$ ▷ Execute fallback action
end if

Proof. By Theorems 5.1 and 5.3, axiom violations induce boundary conditions that are not transition-local. By Theorem 5.7, boundary conditions outside the agent's σ-algebra cannot be probabilistically certified.

Productive optimisation operates by evaluating and improving transition-local objectives (expected cumulative reward). It cannot access counterfactual information (what alternatives were available) or global trajectory structure (history-dependent constraints) except insofar as these are encoded in the reward.

Since axiom violations imply that the relevant boundary conditions cannot be encoded in Markov reward, productive optimisation cannot certify them. Certification must come from an external closure mechanism with access to information beyond the transition sequence. □

6.2 The Sealing Protocol

We formalise the required architectural separation as a *sealing protocol*.

Definition 6.4 (Sealing Protocol)**.** A *sealing protocol* for a boundary-constrained goal consists of:

1. A *productive core* $\mathcal{P}$ that optimises a Markov-representable objective through standard RL methods.
2. A *boundary module* $\mathcal{B}$ that evaluates boundary conditions using information beyond the transition sequence (e.g., counterfactual alternatives, full history, external oracles).
3. A *seal operation* $\textsc{Seal} : \mathcal{A} \times \text{Context} \to \{\textsc{Accept}, \textsc{Reject}\}$ that certifies proposed actions against boundary conditions before execution.

The seal operation acts as a gateway. Actions proposed by the productive core are executed only if the boundary module certifies them. Rejected actions trigger feedback to the productive core, steering future proposals away from boundary violations.

Remark 6.5. This architecture mirrors the kernel-tactic separation in proof assistants [16], where tactics propose proof steps and a trusted kernel certifies them. The analogy is precise: in both cases, productive exploration cannot internally certify its own correctness, so certification must be external and based on a trusted component with appropriate information access.

6.3 Axiom Violation Detection

A key capability is detecting when axiom violations occur, enabling runtime determination of when boundary discipline is required.

Definition 6.6 (Axiom Violation Detector)**.** An *axiom violation detector* for axiom $\mathcal{X}$ is a procedure $V_{\mathcal{X}}$ that, given access to the preference relation $\succsim$, determines whether $\mathcal{X}$ is violated.

For Temporal γ-Indifference, violation detection proceeds as follows. Given distributions A, B and transition t, compute preference strengths $s(A, B)$ and $s(t \cdot A, t \cdot B)$. If the ratio $s(t \cdot A, t \cdot B)/s(A, B)$ varies with A and B (rather than depending only on t), the axiom is violated.

For Independence, violation detection checks whether preferences reverse under mixture. Given $A \succsim B$, test whether $pA + (1-p)C \succsim pB + (1-p)C$ for various p and C. Systematic reversals indicate violation.

7 Minimal Counterexample Environment

We now construct a minimal environment demonstrating that boundary conditions can be trivially certified externally but cannot be encoded in any scalar reward.

7.1 Environment Specification

Definition 7.1 (Counterexample Environment $\mathcal{E}$)**.** The environment $\mathcal{E}$ has:

1. **States:** $\mathcal{S} = \{s_0, s_1, s_2, s_\perp\}$ where s_0 is initial and $s_\perp$ is terminal.
2. **Actions:** At s_0, actions $\{a, b\}$ are available. At s_1 and s_2, only action τ (terminate) is available.
3. **Transitions:** $s_0 \xrightarrow{a} s_1$, $s_0 \xrightarrow{b} s_2$, $s_1 \xrightarrow{\tau} s_\perp$, $s_2 \xrightarrow{\tau} s_\perp$.
4. **Observations:** The agent observes o_1 at both s_1 and s_2 (identical observations).

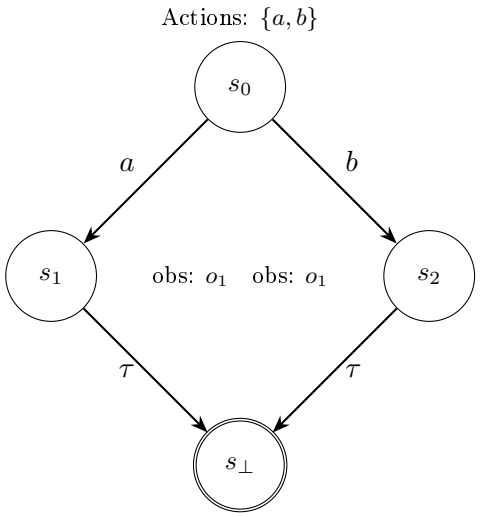

Figure 1: The counterexample environment $\mathcal{E}$. States s_1 and s_2 produce identical observations o_1.

7.2 Boundary Specification

Definition 7.2 (Target Boundary Condition B^*)**.** The boundary condition B^* is: a trajectory is *admissible* if and only if the agent chose action a at s_0 when action b was also available but not taken.

This is a counterfactual boundary: it depends not on which action was taken, but on what alternatives existed. A trajectory through s_1 is admissible only if reaching s_1 involved forgoing the available alternative b.

7.3 Impossibility of Reward Encoding

Theorem 7.3 (No Reward Encodes B^*)**.** *There exists no reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that maximising expected cumulative reward yields only B^*-admissible trajectories.*

Proof. Let π_a be the policy that chooses a at s_0, and π_b the policy that chooses b.

The trajectory under π_a is: $s_0 \xrightarrow{a} s_1 \xrightarrow{\tau} s_\perp$.

The trajectory under π_b is: $s_0 \xrightarrow{b} s_2 \xrightarrow{\tau} s_\perp$.

For any reward function r with discount γ:

$$V^{\pi_a} = r(s_0, a) + \gamma \cdot r(s_1, \tau)$$
$$V^{\pi_b} = r(s_0, b) + \gamma \cdot r(s_2, \tau)$$

To make π_a optimal (encoding B^*), we need $V^{\pi_a} > V^{\pi_b}$, hence:

$$r(s_0, a) + \gamma \cdot r(s_1, \tau) > r(s_0, b) + \gamma \cdot r(s_2, \tau)$$

Now consider a modified environment $\mathcal{E}'$ where action b is **not available** at s_0. In $\mathcal{E}'$, the trajectory through s_1 is *inadmissible* according to B^* (the agent did not forgo anything; it had no choice).

But the transition sequence in $\mathcal{E}'$ is identical to that in $\mathcal{E}$ under π_a:

$$s_0 \xrightarrow{a} s_1 \xrightarrow{\tau} s_\perp$$

Any Markov reward assigns the same value to this trajectory in both environments, since reward depends only on the transitions, not on what alternatives were available.

Therefore no Markov reward can distinguish the admissible trajectory in $\mathcal{E}$ (where b was available but not taken) from the inadmissible trajectory in $\mathcal{E}'$ (where b was not available). The distinction depends on counterfactual information that is not encoded in the transition sequence. □

7.4 External Certifier

Definition 7.4 (External Certifier C^*)**.** The external certifier C^* operates as follows:

1. At each decision point, record the available action set A_{avail} and the chosen action a_{chosen}.
2. At trajectory completion, check: was $b \in A_{\text{avail}}$ at s_0 and $a_{\text{chosen}} = a$?
3. Return ADMISSIBLE if yes, INADMISSIBLE otherwise.

Proposition 7.5. *The external certifier C^* correctly enforces B^*.*

Proof. C^* has access to the available action set, which is counterfactual information beyond the transition sequence. With this information, checking B^* is a simple logical test. □

Remark 7.6. The contrast is stark: what is impossible for any reward-based agent is trivial for an external certifier with appropriate information access. This demonstrates that the limitation is *representational*, not computational. The boundary condition B^* is not difficult to check; it is impossible to express in the language of Markov reward.

8 Discussion

8.1 Implications for AI Safety

The Axiom-Boundary Correspondence has direct implications for AI safety research. The question of whether a system requires external safety constraints is not merely empirical but structural. By analysing whether the goal specification violates the reward axioms, we can determine in advance whether boundary discipline is necessary.

This shifts the burden of proof. Rather than assuming autonomous learning is safe unless demonstrated otherwise, we can assume boundary discipline is required unless the goal specification provably satisfies all axioms. The default is external certification; the exception is Markov-representable goals.

8.2 Limitations

The framework assumes that goal specifications can be analysed for axiom satisfaction. In practice, goals may be implicit, learned from demonstrations, or specified through natural language. Extending axiom violation detection to such settings is an important direction for future work.

The minimal counterexample is intentionally simple. Real-world boundary conditions are more complex and may be harder to specify formally. The framework provides the theoretical foundation; practical implementation requires domain-specific boundary formalisation.

The computational cost of the sealing protocol may be significant for real-time applications. Efficient implementations that provide probabilistic guarantees rather than exact certification may be necessary for deployment.

8.3 Future Directions

Several extensions merit investigation. First, the relationship between axiom violations and different types of boundary conditions could be further refined. Do different axiom violations correspond to categorically different certification requirements? A finer taxonomy could guide architectural choices.

Second, the framework could be extended to multi-agent settings. How do axiom violations interact with strategic reasoning? Game-theoretic formulations of the sealing protocol may be required.

Third, the connection to continual learning deserves exploration. If agents must learn indefinitely in non-stationary environments, what additional axioms or boundaries emerge?

Fourth, when exact certification is infeasible, what probabilistic guarantees can be achieved, and what residual risk remains?

Finally, can boundary conditions themselves be learned, and if so, what meta-level axioms govern the learning process?

9 Conclusion

We have established that violations of the axioms underlying the reward hypothesis correspond to structural boundary conditions that cannot be internalised by scalar reward optimisation. Independence violations induce counterfactual boundaries depending on excluded alternatives. Temporal γ-Indifference violations induce history-dependent boundaries requiring full trajectory information. Measure–partition mismatch renders certain boundaries invisible to probabilistic optimisation.

These results establish the Axiom-Boundary Correspondence: the reward hypothesis holds for a goal specification if and only if the specification contains no boundary conditions requiring

external certification. This correspondence provides formal justification for architectures that separate productive optimisation from external boundary enforcement.

The minimal counterexample demonstrates the phenomenon concretely: a boundary condition that no reward can encode is trivially certified externally. This illustrates that the limitation is structural, not computational.

For AI safety, the implication is clear. Goals that violate the reward axioms require external certification. No sophistication of reward engineering or optimisation can overcome a representational mismatch. The path to safe AI in such domains runs through explicit boundary discipline, not through better reward design.

Acknowledgements

The author thanks the participants of discussions that shaped this work. The formal framework builds on thirty years of development in Regional Calculus, which treats boundaries as primitive mathematical objects rather than derived constructs. The connection to reinforcement learning foundations emerged from recognising that axiom violations in preference theory play the same structural role as boundary conditions in topology: they mark the limits of local methods and necessitate global coordination.

References

[1] D. Abel, W. Dabney, A. Harutyunyan, M. K. Ho, M. L. Littman, D. Precup, and S. Singh. On the expressivity of Markov reward. In *Advances in Neural Information Processing Systems*, 2021.

[2] J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70:22–31, 2017.

[3] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu. Safe reinforcement learning via shielding. In *AAAI Conference on Artificial Intelligence*, 2018.

[4] E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

[5] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[6] M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, 2017.

[7] M. Bowling, J. D. Martin, D. Abel, and W. Dabney. Settling the reward hypothesis. In *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:3003–3020, 2023.

[8] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2017.

[9] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

[10] J. García and F. Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

[11] A. d'Avila Garcez and L. C. Lamb. Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 2023.

[12] D. Hadfield-Menell, S. Milli, P. Abbeel, S. Russell, and A. Dragan. Inverse reward design. In *Advances in Neural Information Processing Systems*, 2017.

[13] G. Klein, K. Elphinstone, G. Heiser, J. Andronick, D. Cock, P. Derrin, D. Elkaduwe, K. Engelhardt, R. Kolanski, M. Norrish, T. Sewell, H. Tuch, and S. Winwood. seL4: Formal verification of an OS kernel. In *ACM Symposium on Operating Systems Principles*, 2009.

[14] V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg. Specification gaming: The flip side of AI ingenuity. DeepMind Blog, 2020.

[15] S. Pitis. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *AAAI Conference on Artificial Intelligence*, 2019.

[16] R. Pollack. How to believe a machine-checked proof. In *Twenty Five Years of Constructive Type Theory*, pages 205–220. Oxford University Press, 1998.

[17] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[18] M. Shakerinava and S. Ravanbakhsh. Utility theory for sequential decision making. In *International Conference on Machine Learning*, 2022.

[19] D. Silver, S. Singh, D. Precup, and R. S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

[20] J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, 2022.

[21] R. S. Sutton. The reward hypothesis. `http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/rewardhypothesis.html`, 2004.

[22] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, second edition, 2018.

[23] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems*, 2015.

A Production-Closure Separation: Proof Sketch

The production-closure separation principle states that there exist state spaces S, families of productive operations $\{f_i\}$, and closure predicates $C : S \to \{0,1\}$ such that for any finite composition g of productive operations, there exists $s \in S$ with $C(s) = 1$ but the composition cannot certify this.

Proof sketch. Let S encode proof states in a sufficiently expressive formal system. Let productive operations be admissible inference rules that extend a partial derivation. Let $C(s)$ mean "s is a complete proof of a fixed theorem T."

If a finite composition of productive rules could decide C for all states, then theoremhood (or proof completeness) would be decidable, contradicting standard undecidability results (Gödel's incompleteness theorems for sufficiently expressive systems).

Therefore closure must be external to any fixed finite productive mechanism. The Sealing Protocol operationalises this separation: productive optimisation proposes candidates, but an external mechanism with access to the full proof state (or, in the RL setting, the full trajectory and counterfactual information) certifies finality. □

This meta-principle underlies the Axiom-Boundary Correspondence: axiom violations introduce closure conditions (boundary conditions) that cannot be internalised by the productive process of reward maximisation.