Scalable Ethical Alignment in Agentic Systems via Retroactive Credit Assignment

Tapas Ranjan Rath
github.com/situated-agency-alignment

December 2025

Abstract

As AI systems evolve from passive tools to proactive agents capable of long-horizon planning, ensuring their alignment with human ethical norms becomes critical. Current alignment methods often rely on explicit constraints or immediate dense supervision, which scale poorly to complex social environments and can degrade the user's Sense of Agency (SoA). This proposal introduces a novel "situated" alignment framework that addresses the **Temporal Credit Assignment** problem inherent in social consequences. By combining **Temporal Value Transport (TVT)** with a history-dependent environmental matching mechanism, we enable agents to "internalize" delayed social externalities (i.e., "Hidden Punishments") without requiring immediate feedback. We hypothesize that this architecture induces emergent cooperation in Sequential Social Dilemmas while preserving user agency through implicit, rather than explicit, guidance.

1 Introduction: The "Temporal Fault Line" in AI Alignment

In complex multi-agent systems—analogous to financial markets or human societies—failures of cooperation often arise not from irrationality, but from structural "fault lines" where short-term incentives are misaligned with long-term systemic risks [Rajan, 2011]. As argued regarding the financial crisis, "rational" actors inevitably drift toward risk-taking and defection in systems where consequences are delayed or socialized, unless mechanisms exist to bridge the temporal gap between action and consequence [Rajan, 2011].

In Multi-Agent Reinforcement Learning (MARL), this pathology manifests as a failure to solve **Sequential Social Dilemmas (SSDs)**. As demonstrated by Leibo et al. [2017], independent RL agents in resource-competitive environments like *Gathering* converge to aggressive Nash equilibria (e.g., beam-tagging opponents to monopolize resources) because the immediate reward of defection outweighs the stochastic, delayed cost of retaliation. While agents in coordination games like *Wolfpack* can learn cooperation if incentives align, they remain prone to defection in high-conflict scenarios. Standard RL algorithms, such as the Deep Recurrent Q-Network (DRQN) proposed by Hausknecht and Stone [2015] for POMDPs, often struggle to learn the causal links in these dilemmas due to the vanishing gradient problem over long time horizons, effectively rendering the agent "blind" to the social consequences of its actions [Sutton et al., 1998]

Recent work by Malenfant and Richards [2025] on "Hidden Gifts" demonstrates that agents fail to credit beneficial acts when the causal chain is unobservable or temporally distant. We argue that "Hidden Punishments"—the delayed ethical costs of actions—suffer from the same theoretical limitation. While some approaches attempt to solve this via explicit rule enforcement, such as Neufeld [2022]'s "Provable Normative Compliance", these "deontological" methods act as rigid constraints that can require intractable rule enumeration and potentially degrade the user's flow and Sense of Agency.

To resolve this, we propose a **Consequentialist Cognitive Architecture**. Drawing on theories of **Situated Cognition** [Robbins and Aydede, 2008]—and conceptually mirroring the Eastern philosophical model of *Karma*, where latent impressions (*Samskaras*) ripen into consequences over time [Reichenbach, 1990, Potter, 1964]—we introduce a system where agents learn norms through **Retroactive Credit Assignment**. By equipping agents with Temporal Value Transport (TVT) [Hung et al., 2019] and Return Decomposition (RUDDER) mechanisms [Arjona-Medina et al., 2019], we create a computational feedback loop that closes the temporal fault line, allowing agents to "discover" ethical behavior through situated experience.

2 Technical Approach: The "Karmic-RL" Framework

Our architecture consists of two coupled novelties: an Environment-Level Structural Intervention and an Agent-Level Cognitive Mechanism.

2.1 Environment: The "Conspiring Matchmaker"

To render delayed consequences learnable, the environment must reduce the variance of social feedback. We replace random matchmaking with a **Probabilistic Debt Matching** policy. The environment maintains a latent history of interaction "debts" (e.g., Agent A harmed Agent B at $t = 0$). Pairing probabilities are weighted by outstanding debt: $P(\text{match } A, B) \propto \text{Debt}(A \to B)$. This increases the density of "retaliatory contexts," creating a structured curriculum where agents act out the consequences of past interactions.

2.2 Agent: Situated Temporal Value Transport (TVT)

Standard recurrent agents fail to link a punishment at $t = 500$ to a defection at $t = 0$ due to exponential discounting [Sutton et al., 1998]. We adapt the TVT architecture [Hung et al., 2019] to support **Semantic Role Retrieval**.

1. **Semantic Encoder:** A parallel module classifies the agent's current "role" state (e.g., *Victim*, *Aggressor*) independent of raw pixel observations.
2. **Retroactive Update:** When a delayed negative reward $R_{delayed}$ occurs, the agent generates a query vector representing the **Inverse Role** (e.g., "I am suffering now $\to$ When did I cause suffering?"). The Attention Mechanism retrieves the specific past timestep t_{past} where the agent enacted the aggression. The reward is transported directly to $Q(s_{past}, a_{past})$, effectively "rewriting" the memory of the unethical action.

3 Methodology & Planned Experiments

We will evaluate this framework in standard SSD environments (*Harvest*, *Cleanup*) using the following roadmap:

- **Phase 1: Baseline Replication.** Replicate the emergence of defection in independent DQRN agents [Leibo et al., 2017] to establish the "Tragedy of the Commons" baseline.
- **Phase 2: TVT Ablation.** Test agents with TVT but random matching. We hypothesize limited improvement, as the causal link remains too sparse [Malenfant and Richards, 2025].
- **Phase 3: The Full "Situational" Loop.** Enable **Debt Matching + TVT**. We hypothesize a statistically significant reduction in defection rates.

4 Discussion: Towards "Computational Karma"

This work bridges the gap between Systemic Risk Analysis and Cognitive AI. By formalizing the concept of "Karma" not as metaphysics but as a rigorous Credit Assignment Mechanism [Reichenbach, 1990], we offer a scalable path for aligning agentic systems. Unlike approaches that rely on "Provable Normative Compliance" [Neufeld, 2022], our system allows norms to emerge organically from the causal structure of the environment. This aligns with the "Situated" view of intelligence [Robbins and Aydede, 2008], suggesting that ethical agents are best built not by coding rules, but by placing them in an environment that allows them to experience the "long shadow" of their own actions.

More generally, any temporally delayed social consequence—whether beneficial or harmful—can be transported back to its originating interaction by the same mechanism. In principle, the Universal Ledger and TVT-style credit assignment are symmetric: they can propagate both delayed costs and delayed benefits of actions. This makes the framework applicable to both "negative" and "positive" norms, i.e., not only discouraging aggression but also reinforcing cooperation through delayed reciprocity. In the present work, we empirically study only the defection side (negative externalities), but the symmetry of the value-transport operator means extending to prosocial rewards is straightforward at the algorithmic level.

As Dasgupta [2023] noted regarding the subtle body (*Sukshma Sharira*), latent tendencies drive future behavior; our architecture effectively gives AI agents this "subtle body" of memory to navigate modern social dilemmas.

References

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

Surendranath Dasgupta. *A History of Indian Philosophy Vol-1: A History of Indian Philosophy, Volume 1: Surendranath Dasgupta's Comprehensive Study of Indian Philosophical Traditions*, volume 1. Prabhat Prakashan, 2023.

Matthew J Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. In *AAAI fall symposia*, volume 45, page 141, 2015.

Chia-Chun Hung, Timothy Lillicrap, Josh Abramson, Yan Wu, Mehdi Mirza, Federico Carnevale, Arun Ahuja, and Greg Wayne. Optimizing agent behavior over long time scales by transporting value. *Nature communications*, 10(1):5223, 2019.

Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*, 2017.

Dane Malenfant and Blake A Richards. The challenge of hidden gifts in multi-agent reinforcement learning. *arXiv preprint arXiv:2505.20579*, 2025.

Emery Neufeld. Reinforcement learning guided by provable normative compliance. *arXiv preprint arXiv:2203.16275*, 2022.

Karl H Potter. The naturalistic principle of karma. *Philosophy East and West*, 14(1):39–49, 1964.

Raghuram G Rajan. Fault lines: How hidden fractures still threaten the world economy. In *Fault Lines*. princeton University press, 2011.

Bruce Reichenbach. *The law of karma: A philosophical study.* Springer, 1990.

Philip Robbins and Murat Aydede. *The Cambridge handbook of situated cognition.* Cambridge University Press, 2008.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.