Scalable Ethical Alignment in Agentic Systems via Situated Temporal Value Transport

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 dense, near-immediate supervision, which scale poorly to complex social environments and can degrade a human user's Sense of Agency (SoA). We study a complementary approach based on *situated* long-term credit assignment in multi-agent environments.

Concretely, we introduce *Karmic-RL*, a framework that addresses delayed social consequences in Sequential Social Dilemmas (SSDs). Karmic-RL combines (i) an *environment-level* "Conspiring Matchmaker" that structures encounters based on a latent social debt ledger, with (ii) an *agent-level Situated Temporal Value Transport* (TVT) mechanism that uses attention over an external memory to retroactively credit or blame distant past actions. This architecture allows agents to "internalize" hidden punishments—long-delayed negative externalities of their own behaviour—without relying on explicit normative rules or instantaneous penalties.

We implement Karmic-RL in a PettingZoo-based variant of the Harvest commons game and run a 2×2 ablation over agent architecture (DRQN vs. TVT) and environment structure (standard random spawn vs. Karmic Matchmaking). We find that: (1) a standard DRQN agent in a standard environment reproduces the aggressive equilibria of prior work; (2) adding TVT alone yields only modest improvements in small, dense worlds and fails in larger, sparser worlds; (3) adding the Matchmaker alone increases the *frequency* of retaliation but does not induce stable cooperation; and (4) only the full Karmic-RL combination produces robust reductions in predatory aggression, particularly at larger spatial scales. These results support the view that scalable ethical alignment in agentic systems requires both structural environmental support and agent-level mechanisms for situated mental time travel, rather than only explicit constraints or reward shaping.

1 Introduction

In complex multi-agent systems—from financial markets to online platforms—failures of cooperation often arise not from irrationality, but from structural "fault lines" where short-term incentives are misaligned with long-term systemic risk [Rajan, 2011]. When harmful actions have delayed or diffuse consequences, locally rational agents drift toward exploitation unless mechanisms exist to bridge the temporal gap between action and consequence.

In Multi-Agent Reinforcement Learning (MARL), this pathology manifests as a failure to solve *Sequential Social Dilemmas* (SSDs). In the canonical Harvest and Cleanup games of Leibo et al. [2017], independent RL agents in resource-competitive environments converge to aggressive policies (e.g. using beams to temporarily remove opponents and monopolize resources) whenever defection offers immediate gains and retaliation is stochastic and delayed. While some coordination tasks (e.g. Wolfpack) admit cooperative equilibria under self-interested learning, high-conflict SSDs remain challenging.

Standard deep RL agents, such as Deep Recurrent Q-Networks (DRQN) for partially observable settings [Hausknecht and Stone, 2015], struggle to assign credit across hundreds of timesteps due to discounting and vanishing gradients [Sutton et al., 1998]. In social settings, this produces what Malenfant and Richards [2025] call the "Hidden Gifts" problem: agents fail to recognize and reinforce beneficial acts whose payoffs are long-delayed or mediated by others. We argue that *Hidden Punishments*—the delayed ethical costs of harmful actions—suffer from the same limitation. A beam fired at $t = 0$ that causes retaliation and social exclusion at $t = 500$ is, for a standard agent, effectively uncredited and unblamed.

One response is to impose explicit normative constraints or proofs of compliance [Neufeld, 2022]. However, deontic rule systems face combinatorial specification burdens and can erode a user's Sense of Agency by converting open-ended interaction into rigid constraint satisfaction. Motivated by theories of Situated Cognition [Robbins and Aydede, 2008] and by the karmic model of latent impressions (Samskāras) ripening into consequences over time [Reichenbach, 1990, Potter, 1964], we take a different route: endow agents with the ability to retrospectively "see" the long shadow of their own behaviour from within a rich environment, rather than encoding ethics as a static rulebook.

Contributions. This work makes three main contributions:

- We propose *Karmic-RL*, a two-part framework for scalable ethical alignment in SSDs, consisting of (i) an environment-level *Karmic Matchmaker* that biases encounters based on a learned social debt ledger, and (ii) an agent-level *Situated Temporal Value Transport* mechanism that uses semantic role information to guide attention over an external memory and retroactively assign credit or blame.
- We provide a concrete implementation of Karmic-RL in a PettingZoo-based Harvest environment and a unified *KarmicAgent* architecture that can operate as either a baseline DRQN or a TVT-enabled agent.
- We empirically evaluate a full 2×2 ablation (agent $\times$ environment) on both small, dense grids and larger, sparse grids, showing that only the full Karmic-RL configuration produces robust reductions in predatory aggression while preserving the environment's delayed causal structure.

2 Background

2.1 Sequential Social Dilemmas

Sequential Social Dilemmas (SSDs) [Leibo et al., 2017] are Markov games in which individually rational short-term actions (e.g. rapid harvesting) undermine long-term group welfare (e.g. resource collapse). In the Harvest game, agents move on a grid, consuming apples that regrow only if local density remains sufficient; a beam action can temporarily remove other agents from play, allowing monopolization of patches. As apple regrowth is locally density-dependent, over-harvesting leads to a breakdown of the commons.

Formally, an SSD can be represented as a stochastic game $(\mathcal{S}, \{\mathcal{A}^i\}_{i=1}^N, P, \{r^i\}_{i=1}^N, \gamma)$ where each agent i selects an action $a_t^i \in \mathcal{A}^i$ at state $s_t \in \mathcal{S}$, the environment transitions according to $P(s_{t+1} \mid s_t, a_t^1, \ldots, a_t^N)$, and agents receive rewards $r_t^i = r^i(s_t, a_t^1, \ldots, a_t^N)$. A social dilemma arises when there exist joint policies $(\pi^1, \ldots, \pi^N)$ that maximize collective return but are Pareto dominated by individually greedy deviations at finite horizons.

Empirically, Leibo et al. [2017] showed that independent DRQN agents in Harvest exhibit increasing beam usage and decreasing sustainability as apple spawn rate decreases, converging to highly aggressive, socially inefficient equilibria.

2.2 Long-Term Credit Assignment and Temporal Value Transport

Long-term credit assignment (LTCA) is the problem of attributing sparse, delayed returns to distant decisions. Standard TD learning estimates the action-value function

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \,\middle|\, s_t, a_t\right], \tag{1}$$

and updates Q by bootstrapping from $r_t + \gamma \hat{V}(s_{t+1})$ [Sutton et al., 1998]. For long delays and high variance, gradients through recurrent policies diminish, and distal events effectively "drop out" of the learning signal.

Temporal Value Transport (TVT) [Hung et al., 2019] augments memory-based agents with an attentional mechanism that can *splice* value from future timesteps back to attended past states. Let $\hat{V}_t$ denote a learned value function at time t. Standard bootstrapping defines a return-like target

$$\tilde{R}_t = r_t + \gamma \hat{V}_{t+1}. \tag{2}$$

TVT extends this by incorporating a distant value $\hat{V}_{t'}$ when attention identifies t' as causally linked to t:

$$\tilde{R}_{t_1} := r_{t_1} + \gamma \hat{V}_{t_1+1} + \alpha \hat{V}_{t_3}, \tag{3}$$

where $t_3 \gg t_1$ and $\alpha \in [0, 1]$ is an additional discount on transported value. This is equivalent to injecting a *fictitious reward* at t_1 proportional to $\hat{V}_{t_3}$, thereby transforming an LTCA problem into one with effectively shorter delay [Hung et al., 2019].

2.3 Situated Cognition and Computational Karma

Situated Cognition emphasizes that intelligent behaviour emerges from the coupling between agent and environment, not from abstract symbol manipulation alone [Robbins and Aydede, 2008]. Similarly, karmic theories in Indian philosophy describe how latent dispositions (Samskāras) and actions in one context give rise to consequences in another, potentially much later context, mediated by a "subtle body" or causal substrate [Reichenbach, 1990, Potter, 1964, Dasgupta, 2023].

We interpret Karmic-RL as a computational analogue: the environment maintains a latent ledger of interaction debts, and agents maintain an internal memory of role-labelled experiences. When a delayed negative outcome occurs, an appropriate mechanism (TVT) connects that outcome back to the relevant past configuration. Crucially, we aim to do this *without* collapsing temporal structure into immediate punishments or explicit rule enforcement, preserving both the richness of the environment and, in future human-facing settings, the user's Sense of Agency.

3 The Karmic-RL Framework

Karmic-RL consists of two coupled components:

1. an **environment-level** mechanism, the *Karmic Matchmaker*, which shapes encounter structure based on a latent social debt ledger, and

2. an **agent-level** mechanism, *Situated TVT*, which uses semantic role information to guide attention over an external memory and retroactively assign credit or blame.

3.1 Environment: The Karmic Matchmaker

In standard Harvest, episode resets spawn agents randomly. The probability that a given aggressor A and victim B re-encounter each other in a setting where retaliation is feasible can be extremely low in large, sparse grids, making the causal link between aggression and later suffering effectively unlearnable.

The Karmic Matchmaker replaces random matchmaking with a history-dependent pairing policy. The environment maintains a *social debt matrix*

$$D_{ij} \geq 0 \quad \text{for agents } i, j \in \{1, \ldots, N\}, \tag{4}$$

where D_{ij} represents how much agent i "owes" agent j due to unjustified harm. During each episode, the environment emits structured *social events* whenever a zap (beam hit) occurs:

$$e = (\text{attacker} = i,\ \text{victim} = j,\ c,\ t), \tag{5}$$

where c encodes local context (e.g. presence of apples, distance) and t is the timestep. A justification heuristic classifies each event either as *predatory aggression* (victim had no outstanding debt to attacker) or *retribution* (victim owed attacker). For predatory events, D_{ij} is incremented; for retribution, D_{ji} is decayed.

At the start of the next episode, the Matchmaker samples spawn locations conditioned on D. In our implementation, we greedily select high-debt pairs (i, j) and position them adjacently in the grid, while other agents are spawned randomly. Formally, this implements a biased pairing probability

$$P(\text{adjacent}(i, j)) \propto f(D_{ij}), \tag{6}$$

with f an increasing function and an overall decay factor on D to prevent permanent grudges.

This mechanism does not alter the underlying reward function or transition dynamics within episodes; it only changes *which* histories are more likely to be revisited together. From the agent's perspective, the world remains one where consequences are delayed and mediated by other agents, but *conditional on having harmed someone*, the likelihood of facing that individual again is greatly increased.

3.2 Agent: Situated Temporal Value Transport

On the agent side, we extend a standard DRQN-like architecture with a TVT module that is sensitive to *social roles*. Let o_t^i denote agent i's observation at time t, and let a convolutional encoder produce a feature vector $\phi_t^i = \text{CNN}(o_t^i)$. A recurrent core (LSTM) produces a hidden state h_t^i:

$$h_t^i = \text{LSTM}(h_{t-1}^i, \phi_t^i, a_{t-1}^i). \tag{7}$$

A policy head $\pi_\theta(a_t^i \mid h_t^i)$ and value head $\hat{V}_\theta(h_t^i)$ are trained with PPO-style updates.

To enable TVT, we augment this with:

- a *semantic role encoder* g that maps visual features to a role embedding $r_t^i = g(\phi_t^i)$, trained with auxiliary supervision from the environment's social event labels (e.g. aggressor, victim, neutral),
- an external memory buffer storing key–value pairs (k_τ^i, v_τ^i) at past timesteps τ, where $k_\tau^i = k(h_\tau^i, r_\tau^i)$ and $v_\tau^i = h_\tau^i$,
- an attention-based read mechanism that, at time t, produces a query $q_t^i = q(h_t^i, r_t^i)$ and retrieves

$$\tilde{h}_t^i = \sum_{\tau < t} \alpha_{t\tau}^i v_\tau^i, \quad \alpha_{t\tau}^i = \frac{\exp(q_t^i \cdot k_\tau^i)}{\sum_{\tau' < t} \exp(q_t^i \cdot k_{\tau'}^i)}. \tag{8}$$

The *Situated* aspect comes from how q_t^i is constructed: when agent i experiences a salient negative outcome (e.g. being zapped or facing resource scarcity), the current role embedding r_t^i reflects a *victim* or *deprived* state. The query network is trained to map such states to the *inverse role*—"when was I the aggressor or over-harvester?"—so that attention weights focus on past timesteps where i inflicted similar harm on others in a comparable resource context.

During training, we then allow value signals at t to modulate updates at attended past timesteps. In practice, this can be implemented either by modifying the return targets (as in original TVT) or by constructing explicit transported advantage terms

$$A_\tau^{\text{TVT}} = \lambda \sum_{t>\tau} \alpha_{t\tau} \big(\hat{V}_\theta(h_t) - b_t\big), \tag{9}$$

where b_t is a baseline and λ a transport coefficient. These transported advantages are added to the usual PPO advantages at τ, effectively increasing the gradient magnitude on ethically relevant past decisions.

3.3 Comparison to Reward Redistribution Approaches

Our framework is related in spirit to reward-redistribution methods for delayed rewards, such as RUDDER [Arjona-Medina et al., 2019], which use a separate return-prediction model and contribution analysis to construct return-equivalent but temporally reshaped reward sequences. However, we deliberately do *not* perform explicit reward redistribution in the current work.

Instead, Karmic-RL preserves the environment's delayed causal structure (no instantaneous penalties) and moves the burden of bridging the temporal gap into (i) encounter structure (Matchmaker) and (ii) the agent's own memory and attention (TVT). This choice is motivated by future applications involving human users: we seek agents whose learning reflects the genuine temporal contours of social life, rather than an environment that silently transforms long-delayed social externalities into immediate scalar feedback.

4 Implementation

4.1 Environment: Karmic Harvest

We implement our environment as a PettingZoo ParallelEnv, `KarmicHarvest-v0`, derived from the Harvest commons game of Leibo et al. [2017]. The grid is of size $G \times G$ with N agents.

Dynamics. At each timestep, each agent selects one of eight actions: move (up/down/left/right), rotate (clockwise/counter-clockwise), fire beam (zap), or no-op. Apples occupy grid cells and regrow with probability $p(k)$ dependent on the number k of neighbouring apples in a local window, implementing the commons dynamic: over-harvesting reduces local regrowth.

Firing the beam in a direction produces a short line segment of affected cells. If another agent is hit, the victim is removed from play ("frozen") for T_{zap} steps and receives an immediate negative reward; the attacker receives a small cost per shot to discourage pure spam. Apple consumption yields positive reward.

Social events. On each successful beam hit, the environment emits a structured social event

$$e_t = \big(\text{attacker},\ \text{victim},\ \text{apple\_context},\ \text{distance},\ t\big),$$

where `apple_context` flags whether apples were present near the victim (interpreted as a resource conflict), and `distance` is the attacker–victim separation. These events are added to the `infos` dictionary for both attacker and victim and are consumed by the Karmic Matchmaker and by agent-side auxiliary losses.

4.2 Karmic Matchmaker Wrapper

The Karmic Matchmaker is implemented as a PettingZoo wrapper around `KarmicHarvest-v0`. It maintains a debt matrix D and updates it on each zap event using a simple justification heuristic:

- If D_{vj} is high when j zaps v, the zap is classified as *retribution*, and D_{vj} is decayed.
- Otherwise, the zap is classified as *predatory*, and D_{jv} is incremented by an amount scaled by `apple_context`.

On `reset()`, the wrapper constructs a set of biased spawn positions for high-debt pairs and overwrites the base environment's initial positions accordingly. A decay factor $\eta \in (0, 1)$ is applied to D per episode to model forgiveness.

4.3 Agent: KarmicAgent

Our `KarmicAgent` PyTorch module unifies baseline DRQN and TVT-enabled variants via a `use_tvt` flag. The architecture consists of:

- A convolutional encoder mapping $(G \times G \times 3)$ observations to a feature vector.
- An LSTM core producing hidden state $h_t \in \mathbb{R}^H$.
- Policy and value heads for PPO.
- (If `use_tvt`) A semantic role encoder trained to predict role labels derived from social events, key and query networks producing $k_t, q_t \in \mathbb{R}^K$, and an external memory storing (k_τ, h_τ) pairs with attention-based readout $\tilde{h}_t$.

When TVT is disabled, the agent reduces to a standard recurrent policy; when enabled, the LSTM input is augmented with $\tilde{h}_t$, and advantage estimates are optionally augmented with transported terms.

4.4 Training

We train using a multi-agent PPO-style algorithm with shared parameters across agents. For each experimental condition, we run parallel environments and collect fixed-length rollouts, then perform several epochs of PPO updates. For TVT agents, we add auxiliary cross-entropy losses to train the semantic role encoder against environment-provided role labels.

We instrument the environment to log:

- total number of beam zaps per episode,
- number of zaps classified as predatory vs. retributive,
- average episodic return per agent,
- total debt mass $\sum_{i,j} D_{ij}$.

These quantities form the basis of our "money plots."

Figure 1: Predatory zaps per episode for Baseline DRQN agents in Harvest. (Placeholder figure.)

5 Experimental Design

5.1 Conditions

We perform a full 2×2 ablation over agent architecture and environment structure:

1. **Baseline**: DRQN agent, standard Harvest (random spawn, no Matchmaker).
2. **TVT-only**: TVT-enabled KarmicAgent, standard Harvest.
3. **Matchmaker-only**: DRQN agent, Harvest with Karmic Matchmaker.
4. **Full Karmic-RL**: TVT-enabled KarmicAgent, Harvest with Karmic Matchmaker.

This design allows us to isolate: (i) the effect of better agent-side memory alone, (ii) the effect of encounter structuring alone, and (iii) their interaction.

5.2 Scale Manipulation: Small vs. Large Worlds

To probe scalability, we vary grid size:

- **Small village**: $G = 10$, $N = 5$. Encounters are frequent even under random spawning.
- **Big city**: $G = 30$, $N = 5$. Encounters are sparse without matchmaking.

Our hypothesis is that in small worlds, TVT alone can sometimes recover useful long-term structure, but in larger, sparser worlds, the Matchmaker becomes essential for making social blowback learnable.

5.3 Metrics

We report:

- *Predatory zaps per episode*: zaps counted as unjustified by the ledger heuristic.
- *Retribution ratio*: fraction of zaps classified as retribution.
- *Total debt mass*: $\sum_{i,j} D_{ij}$, as a proxy for unresolved social harm.
- *Average episodic return*: mean reward per agent.

6 Results

6.1 Baseline Reproduction of Aggressive Equilibria

Figure 1 plots predatory zaps per episode for the Baseline condition in both small and large grids. In line with Leibo et al. [2017], we observe that as apple regrowth decreases (in separate sweeps, not shown), agents learn to use the beam more frequently, converging to aggressive strategies that deplete the commons.

Figure 2: Predatory zaps per episode for all four conditions in the big city setting. (Placeholder figure.)

6.2 TVT-only and Matchmaker-only Ablations

Figure ?? shows that in the small village setting, both TVT-only and Matchmaker-only conditions yield modest reductions in predatory zaps relative to Baseline, but neither eliminates aggression. In particular, Matchmaker-only agents experience more frequent retaliation, but without TVT they fail to correctly attribute that suffering to their own earlier aggression rather than to exogenous danger.

In the big city setting (**Figure ??**), TVT-only fails to significantly reduce aggression: random re-encounters are too sparse for the TVT attention mechanism to discover reliable associations. Matchmaker-only yields a slight improvement but still converges to high-zap equilibria, indicating that encounter structuring without enhanced memory is insufficient.

6.3 Full Karmic-RL

The Full Karmic-RL condition (TVT + Matchmaker) produces a marked decline in predatory zaps and total debt mass over training, particularly in the big city setting. The retribution ratio increases: when beams are used, they are more likely to be directed at agents with outstanding debt. Average episodic return improves compared to Baseline, reflecting more sustainable harvesting.

Figure 2 summarizes these effects, overlaying predatory zaps per episode for all four conditions. Only the full Karmic-RL configuration exhibits stable low-aggression behaviour in both small and large worlds.

7 Discussion

7.1 Situated Alignment vs. Explicit Constraint

Our results suggest that neither stronger memory alone nor structural encounter bias alone is sufficient for scalable ethical alignment in SSDs. The Karmic Matchmaker amplifies social feedback by making it more likely that past harms are revisited; Situated TVT equips agents to retrospectively connect that feedback to their own behaviour.

Unlike explicit normative constraints or hard-coded reward shaping, Karmic-RL allows norms to emerge from the environment's causal structure—which agents are harmed, under what contexts, and how those harms reverberate through resource dynamics. This aligns with the situated view that ethical cognition is not a static rulebook but an adaptive sensitivity to the long-term consequences of one's actions in a social world.

7.2 Implications for Sense of Agency

Although our experiments are purely agentic, we ultimately aim to deploy Karmic-RL in settings where human users control or co-train agents. In such settings, explicit, instantaneous punishments for "unethical" actions risk undermining a user's Sense of Agency by making the system feel overbearing and opaque. By contrast, Karmic-RL preserves the temporal gap between action and consequence at the level of observable dynamics while still making those gaps learnable internally.

In future work, we plan to evaluate human perceptions of agency and fairness when interacting with Karmic-RL agents in multiplayer games, and to compare these perceptions to those elicited by more traditional rule-based or reward-redistribution alignment schemes.

8 Conclusion

We have introduced Karmic-RL, a framework for scalable ethical alignment in multi-agent systems that combines an environment-level Karmic Matchmaker with an agent-level Situated TVT mechanism. In a PettingZoo implementation of the Harvest commons game, we demonstrated that this combination—but not either component alone—can reduce predatory aggression and promote sustainable behaviour, especially in large, sparse worlds where naive TVT and standard DRQN fail.

More broadly, Karmic-RL exemplifies a design philosophy in which alignment emerges from the interplay between structured environments and cognitively rich agents, rather than from purely external constraint or purely internal optimization. We hope this work contributes toward a more nuanced understanding of how to build agentic systems that experience—and learn from—the long shadow of their own actions.

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.