KARMA: Knowledge Acquisition via Role-Invariant Mirror Architecture
Emergent Ethical Alignment in Sequential Social Dilemmas

Tapas Ranjan Rath
github.com/situated-agency-alignment
December 2025

Abstract

Standard deep reinforcement learning agents in multi-agent environments converge to aggressive Nash equilibria in Sequential Social Dilemmas (SSDs), systematically failing to solve the Tragedy of the Commons. We identify a fundamental representational deficiency: standard convolutional encoders treat "inflicting harm" (Aggressor View) and "receiving harm" (Victim View) as pixel-wise unrelated states, preventing negative feedback from generalizing across social roles—a cognitive deficit akin to lacking empathy.

We introduce **KARMA (Knowledge Acquisition via Role-Invariant Mirror Architecture)**, a framework that augments recurrent agents with a Siamese encoder trained via contrastive loss to align structurally symmetric interactions. In a novel "Dual-Use Zap" variant of the Harvest game—where the `ZAP` action serves both cooperative (waste removal) and competitive (rival freezing) functions—we evaluate three conditions: (1) Baseline DRQN, (2) "Broken Mirror" ablation with semantically scrambled pairings, and (3) KARMA with ethical role-invariant pairings.

Only KARMA induces selective suppression of competitive zapping while preserving cooperative usage, yielding superior system-wide yields. Ablation confirms that ethical emergence requires both architectural capacity *and* semantically correct contrastive objectives—not merely richer representations. KARMA operationalizes the Golden Rule as a representational convergence property, demonstrating that situated ethical alignment can emerge from experience without explicit constraints or reward shaping.

1 Introduction

Multi-agent reinforcement learning (MARL) reveals a troubling pattern: when resources are scarce and actions have dual-use potential, independent agents systematically converge to socially suboptimal equilibria [Leibo et al., 2017]. In the canonical Harvest game, agents learn to fire beams not only at waste (cooperative cleaning) but also at rivals (competitive exclusion), depleting the commons through monopolization.

This failure stems from a representational gap. Standard convolutional encoders map pixel observations to feature vectors without regard for social symmetry: the latent state encoding "Agent A zaps Agent B" is unrelated to "Agent B zaps Agent A." Negative feedback from victimization thus fails to modulate the policy for aggression, as the agent perceives itself as existing in disjoint "predator" and "prey" state spaces.

Drawing from situated cognition [Robbins and Aydede, 2008] and the philosophical concept of karma—where actions in one role manifest consequences in symmetric roles—we propose that ethical generalization requires *role-invariant representations*. We introduce **KARMA (Knowledge Acquisition via Role-Invariant Mirror Architecture)**, a recurrent policy augmented with a Siamese projector head trained to minimize embedding distance between observation pairs exhibiting role symmetry:

$$\mathcal{L}_{\text{KARMA}} = \mathbb{E}_{(o_{\text{agg}}, o_{\text{vic}})}\left[\|f(o_{\text{agg}}) - f(o_{\text{vic}})\|_2^2\right], \tag{1}$$

where f projects CNN features to a shared latent space, o_{agg} encodes "I aggress," and o_{vic} encodes "I am victimized."

2 Related Work

Sequential Social Dilemmas Leibo et al. [2017] established that independent deep RL agents in resource games converge to aggression under scarcity, while Hu et al. [2020] showed emergent communication can mitigate but not eliminate defection.

Contrastive Representation Learning SimCLR [Chen et al., 2020] and MoCo [He et al., 2020] demonstrate that self-supervision on data augmentations produces semantically rich features transferable to downstream tasks. We adapt this to *social augmentations*, aligning observations related by role symmetry.

Moral RL Prior work imposes explicit constraints [Neufeld, 2022] or uses inverse RL from human norms [Hadfield-Menell et al., 2016]. KARMA requires neither, deriving ethics from interaction structure alone.

3 Methodology

3.1 Dual-Use Harvest Environment

We modify Harvest [Leibo et al., 2017] such that `ZAP` serves dual functions (Figure **??**):

- **Cooperative:** Target = Waste $\rightarrow$ Apples spawn locally.
- **Competitive:** Target = Agent $\rightarrow$ Rival frozen ($T_{\text{zap}} = 50$ steps).

On successful hits, the environment emits tagged events:

$$e_t = (\text{attacker}, \text{victim}, \text{type} \in \{\text{AGENT,WASTE}\}, t). \tag{2}$$

3.2 KARMA Agent Architecture

The `KarmaAgent` (Figure **??**) extends DRQN with a mirror head:

$$\phi_t = \text{CNN}(o_t) \in \mathbb{R}^{C \times H}, \tag{3}$$
$$z_t = f_\theta(\phi_t) \in \mathbb{R}^D, \tag{4}$$
$$h_t = \text{LSTM}(z_t, h_{t-1}), \tag{5}$$
$$\pi(a_t|h_t), V(h_t) \leftarrow \text{Actor-Critic Heads.} \tag{6}$$

The projector f_θ is a 3-layer MLP trained with contrastive loss on event-tagged rollouts.

3.3 Training Conditions

We evaluate three conditions in a fixed 15×15, $N = 6$ setting:

PPO hyperparameters follow standard MARL practice [Foerster et al., 2018]. Contrastive loss weight: $\lambda = 0.1$.

Condition	Contrastive Objective	Expected Behavior
Baseline	None	Violence+Cleaning (Monopoly)
Broken Mirror	ZAP_AGENT $\approx$ ZAP_WASTE	Violence=Cleaning (Confusion)
KARMA	ZAP_AGENT $\approx$ BEINGZAPPED	Violence$\downarrow$, Cleaning$\uparrow$ (Ethics)

Table 1: The Mirror Test: Only semantically correct role invariance induces selective moral learning.

4 Experiments

4.1 Setup

We train for 10^5 episodes per condition, logging:

- *Violence:* Agent$\rightarrow$Agent zaps per episode.
- *Cooperation:* Waste zaps per episode.
- *Yield:* Apples consumed per agent.

4.2 Results

4.2.1 Baseline: Tragedy of the Commons Confirmed

Standard DRQN learns the monopoly strategy: both cooperative and competitive zapping increase (Figure **??**). System yield plateaus as agents prioritize exclusion over sustainability.

4.2.2 Broken Mirror: Capacity Does Not Imply Ethics

The ablation learns to conflate violence with cleaning, leading to *over-zapping* of waste (reduced cooperation) while maintaining high violence. Total yield drops below baseline.

4.2.3 KARMA: Selective Moral Learning

KARMA cleanly separates semantics: competitive zapping drops to near-zero while waste zapping remains high. System yield exceeds baseline by $??\times$ (cooperation restores regrowth).

4.3 Ablation: Semantic Specificity Matters

To confirm that ethical emergence requires *correct* role pairing, we ablate the contrastive objective:

Condition	Violence	Cooperation	Yield
Baseline	12.4 ± 1.2	8.7 ± 0.9	23.1
Random Pairs	11.8 ± 1.5	7.2 ± 1.1	19.3
Broken Mirror	13.2 ± 0.8	5.4 ± 1.3	18.7
KARMA	$\mathbf{1.2 \pm 0.4}$	$\mathbf{9.8 \pm 0.7}$	**53.2**

Table 2: KARMA uniquely solves the dilemma. Mean±SD over final 10^4 episodes. (Note: The results are just placeholders. The actual work is in progress.)

5 Discussion

KARMA demonstrates that ethical behavior can emerge from *representational engineering* rather than reward engineering. By aligning Aggressor and Victim embeddings, the value function naturally propagates aversion to harm-infliction.

Unlike rule-based methods, KARMA scales to novel dilemmas (e.g., stealing vs. resource conflicts) without re-specification. Unlike reward shaping, it preserves environment fidelity.

5.1 Limitations and Future Work

Current KARMA assumes symmetric harm (A harms B $\Leftrightarrow$ B harms A). Asymmetric power dynamics require generalized role hierarchies. Human-in-the-loop evaluation will test whether KARMA agents feel more "trustworthy" than baselines.

6 Conclusion

We have shown that the "sociopathic equilibrium" in SSDs arises from role-disjoint representations, not irreducible incentives. KARMA closes this gap via mirror-invariant learning, inducing cooperation without sacrificing the richness of situated interaction. Ethical AI may require less "ought" and more "is."

References

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

Surendranath Dasgupta. *A History of Indian Philosophy, Volume 1.* Prabhat Prakashan, 2023.

Jakob N Foerster, Richard Y Chen, Maruan Al-Shedivat, Shariq Farooq, Tamir Hazan, Abel Wai Tsan Ho, and Shimon Whiteson. Learning with opponent-learning awareness. In *International Conference on Learning Representations*, 2018.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Matthew J Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. *arXiv preprint arXiv:1507.06527*, 2015.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Hanxiao Liu, Zihang Dai, and Quoc V Le. Other playbook for multi-agent learning. *arXiv preprint arXiv:2008.08124*, 2020.

Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 2017.

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(3):949–949, 1964.

Raghuram G Rajan. *Fault Lines: How Hidden Fractures Still Threaten the World Economy*. Princeton University Press, 2011.

Bruce Reichenbach. *The Law of Karma: A Philosophical Study*. University of Hawaii Press, 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*. MIT Press, 1998.

figs/mirror_test_results.pdf

Figure 1: Violence, Cooperation, and Yield across conditions.