Designing Virtuous Agents via Social Reinforcement Learning

Anonymous Authors

Abstract

This extended abstract argues that prevailing deontic and reward-centric approaches to ethical Reinforcement Learning face structural limits. Rule-based methods are brittle under ambiguity, and scalar rewards often compress multiple values into a single objective that invites proxy gaming. We instead treat ethics as *policy-level dispositions*—relatively stable habits that hold up when incentives, partners, or contexts change. We propose a Social Reinforcement Learning framework to design **virtuous agents** that acquire character not through solitary optimization, but through moral mimesis and socially mediated feedback (internalizing the stable norms of a multi-agent population).

1 Introduction: The Critique of the Solitary Agent

The integration of Artificial Intelligence into mission-critical contexts requires agents that remain robust when ethical tensions arise. However, current patterns in machine ethics have two limitations (Ghasemi and Crowley [2025]). First, deontic approaches (rules/constraints) require principles to be encoded in advance, rendering systems **inflexible** when facing non-stationarity (environments where underlying dynamics or norms evolve over time) (Allen et al. [2005]). Second, consequentialist approaches (reward shaping) implicitly compress diverse moral considerations into a single scalar signal, obscuring trade-offs and inviting proxy gaming—exploiting metric loopholes to maximize reward at the expense of the true objective (Abel et al. [2016]). To address these limitations, we propose a framework where agents acquire ethical stability not through solitary constraints, but through social reinforcement learning. This shifts the focus from maximizing scalar rewards to maintaining normative consistency across dynamic environments.

2 Formal Preliminaries

To formalize the solution, we model tasks not as standard Markov Decision Processes (MDPs) (Puterman [1990]), but as multi-objective social games where normative stability is the primary objective.

2.1 Standard vs. Multi-Objective RL

In standard RL, an agent optimizes a scalar objective $J(\pi) = \mathbb{E}_\pi[\sum \gamma^t r(s_t, a_t)]$. We instead adopt a Multi-Objective RL (MORL) formulation where the reward is vector-valued $\mathbf{r}(s, a) \in \mathbb{R}^m$ (Deschamps et al. [2024]). The agent maximizes a vector of expected returns:

$$\mathbf{J}(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}(s_t, a_t) \right]$$

Optimality is characterized by the Pareto front, allowing trade-offs (e.g., between efficiency and fairness) to remain explicit rather than collapsed.

2.2 Social Extension to RL

We extend this to a stochastic game with N agents, defined as $\mathcal{G} = (\mathcal{S}, \{\mathcal{A}_i\}, \mathcal{P}, \{\mathbf{r}_i\}, \gamma)$. A joint policy $\boldsymbol{\pi}$ induces returns for each agent. Our core hypothesis is that **virtuous agents** (Vishwanath and Omlin [2023]) could emerge when they optimize not just for their own $\mathbf{r}_i$, but align with a *virtue prior* π_0 learned via observation of the collective.

3 Mechanism: Moral Mimesis and Interaction

We propose Social Reinforcement Learning (Social RL) as the foundational mechanism for designing virtuous agents based on social learning, where agents learn by observing other agents. Unlike standard RL, where a solitary agent learns strictly through trial-and-error interaction with a static environment, Social RL embeds the agent within a dynamic population. This shifts the learning paradigm from solitary discovery to social construction, relying on two distinct but complementary channels: *observation* and *interaction*.

3.1 Observation (Mimesis)

The primary channel for acquiring virtue is observational learning, or *mimesis*. In solitary RL, an agent must personally experience a catastrophic state to learn to avoid it, which is a high-risk and sample-inefficient process. Social learning can help overcome this weakness by allowing agents to acquire behavior as a result of observation and its resultant consequences shown by other agents (Ndousse et al. [2021]).

In our framework, this manifests as *cultural transmission*—the domain-general mechanism by which agents internalize dispositions exhibited by models without needing to experience the associated risks directly (Bhoopchand et al. [2023]). By observing which states peer agents seek or avoid, a learner can bootstrap a "Virtue Policy" (V) that serves as a stable prior. Crucially, this process does not require the observed models to be perfect ethical oracles. The learner can bootstrap from imperfect exemplars (e.g., heuristic teachers or human demonstrations), refining these coarse normative signals into robust habits through continued observation (Ndousse et al. [2021], Bhoopchand et al. [2023]). This acts as a stabilizing "low-pass filter" for behavior. While individual reward signals may be noisy or corrupted, the aggregate behavior of a population tends to reflect a stable normative consensus.

3.2 Socially Mediated Feedback

While observation provides the initial policy prior, *normative stability*—the resistance to corruption—is enforced through active interaction. Mere observation is insufficient if an agent discovers a profitable but unethical "cheat" (proxy gaming) that the group has not yet encountered.

To counter this, we introduce *socially mediated feedback*. In this regime, agents are not passive observers but active judges; they provide sanctioning feedback (e.g., negative social rewards or reputation penalties) to peers who violate established norms. This fundamentally alters the game-theoretic landscape: it transforms "virtue" from a purely internal constraint into a dominant strategy. An agent that deviates from the norm to maximize a short-term utilitarian reward will face immediate social penalties (exclusion or sanction) that lower its long-term expected return. Thus, the agent learns to align with the "Virtuous" policy (V) not just out of habit, but because the social environment explicitly constructs a survival advantage for normative consistency.

4 Implementation and Roadmap

Moving beyond theory, we propose a concrete roadmap to implement and validate virtuous agents. The current literature is limited by a lack of benchmarks that adequately capture the complexity of ethical trade-offs in open-ended environments. Most ethical RL benchmarks rely on goal-driven environments with static constraints, which fail to test an agent's ability to resist normative drift over long horizons.

4.1 Phase I: A New Benchmark for Virtue

We propose developing a new benchmark utilizing open-ended environments like **Craftax** (Matthews et al. [2024]). Unlike rigid safety environments, Craftax allows for complex resource management and survival dynamics. In the single-agent phase, we will establish baselines by introducing "corrupted" incentives—situations where the agent is highly rewarded for unsustainable behavior (e.g., resource depletion) that violates long-term virtue. This allows us to measure the failure modes of standard utilitarian agents when faced with proxy rewards.

4.2 Phase II: Multi-Agent Dynamics

The core validation of our thesis requires a transition to **Multi-Agent Craftax** (Omari et al. [2025]). Here, we will simulate a population where agents compete for limited resources. We aim to model the emergence of "Tyrant Agents" (agents that maximize personal utility at the expense of the commons) and demonstrate how a sub-population of "virtuous agents" (using Social RL and sanctioning) can stabilize the environment. The metric

for success is not scalar return, but *disposition retention*, which is the ability of the virtuous population to maintain the "cooperative norm" despite the presence of exploitative strategies.

4.3 Challenges and Open Questions

This implementation faces significant challenges. First is the metric design problem. Operationalizing "virtue" as a measurable statistic without collapsing it back into a scalar reward is difficult. We must develop "trait summaries" that track behavioral consistency rather than just accumulation of reward. Second is the bootstrap problem. Social RL relies on existing norms. In a fresh simulation, we must determine how to seed the initial population with enough "imperfect exemplars" to kickstart the cultural transmission process. Addressing these challenges will move the field from abstract ethical principles to practical, engineering-grade definitions of character.

Final Remarks. Ultimately, this hypothesis aims to demonstrate that social reinforcement learning builds robust virtuous agents that resist incentive corruption more effectively than isolated constraints.

References

David Abel, James MacGlashan, and Michael L Littman. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*, volume 16, Phoenix, AZ, 2016.

Colin Allen, Iva Smit, and Wendell Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3):149–155, 2005.

Avishkar Bhoopchand, Bethanie Brownfield, Adrian Collister, Agustin Dal Lago, Ashley Edwards, Richard Everett, Alexandre Fréchette, Yanko Gitahy Oliveira, Edward Hughes, Kory W Mathewson, et al. Learning few-shot imitation as cultural transmission. *Nature Communications*, 14(1):7536, 2023.

Théo Deschamps, Roman Chaput, and Laetitia Matignon. Multi-objective reinforcement learning: an ethical perspective. In *Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA)*, 2024.

Majid Ghasemi and Mark Crowley. Toward virtuous reinforcement learning. *arXiv preprint arXiv:2512.04246*, 2025.

Michael Matthews, Michael Beukman, Benjamin Ellis, Mikayel Samvelyan, Matthew Jackson, Samuel Coward, and Jakob Foerster. Craftax: A lightning-fast benchmark for open-ended reinforcement learning. *arXiv preprint arXiv:2402.16801*, 2024.

Kamal K Ndousse, Douglas Eck, Sergey Levine, and Natasha Jaques. Emergent social learning via multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 7991–8004. PMLR, 2021.

Bassel Al Omari, Michael Matthews, Alexander Rutherford, and Jakob Nicolaus Foerster. Multi-agent craftax: Benchmarking open-ended multi-agent reinforcement learning at the hyperscale. *arXiv preprint arXiv:2511.04904*, 2025.

Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2: 331–434, 1990.

Ajith Vishwanath and Christian Omlin. Exploring affinity-based reinforcement learning for designing artificial virtuous agents in stochastic environments. In *International Conference on Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications*, pages 25–38. Springer, 2023.