

Virtue by Observation: Social Reinforcement Learning for Intrinsically Ethical Agents

Majid Ghasemi¹

¹University of Waterloo, Canada

Introduction and Motivation

Reinforcement learning is increasingly deployed in contexts with ethical stakes, yet current alignment frameworks share two structural limitations. Rule-based (deontological) approaches encode duties as hard constraints or runtime shields; they offer compliance guarantees but are brittle under ambiguity and distributional shift, and do not cultivate lasting dispositions (Allen, Smit, and Wallach 2005). Scalar-reward (utilitarian) approaches compress diverse moral considerations into a single signal, obscuring legitimate trade-offs and inviting proxy gaming: agents satisfy the proxy rather than the underlying objective (Ouyang et al. 2022; Ghasemi and Crowley 2025).

I argue that both failures share a common cause, which I term *feedback collapse*: when plural human values are compressed into a static scalar reward by a fixed evaluation model, the training gradient can diverge from the latent objectives of factual accuracy and normative safety (Casper et al. 2023; Ghasemi and Crowley 2026b,a). Hallucination, the production of fluent but ungrounded content, is the epistemic manifestation of this divergence; moral drift, convergence toward norm-violating behavior under sustained deployment, is its ethical manifestation (Everitt et al. 2021; Ghasemi and Crowley 2025). These pathologies are typically studied in isolation; I propose that they share an etiology and call for a unified response.

My research pursues a *virtue-centric* alternative. Drawing on virtue ethics’ emphasis on stable, context-sensitive dispositions rather than rule compliance or return maximization (Crisp 2014), I treat ethical behavior as a policy-level property acquired through habituation, social interaction, and multi-agent feedback. The central claim is that robust alignment requires distributing evaluative authority across a dynamic social environment rather than concentrating it in a frozen reward model.

Research Questions

1. **Structural diagnosis.** Can feedback collapse explain hallucination and moral drift as joint consequences of a single training pathology, and can this be formalized in a way that is both precise and empirically testable?

2. **Social correction.** Can peer critique, reputation, and social sanction correct a training signal that has diverged from latent normative objectives, and under what conditions does decentralized peer enforcement remain stable rather than eroding through free-riding or collusion?
3. **Evaluation validity.** Do standard multi-agent RL metrics capture the quality of emergent coordination, or do they conflate legitimate consensus with coerced compliance in ways that matter for deployment?

Completed Work

Virtuous RL: Critique and Roadmap. In Ghasemi and Crowley (2025) I articulate the structural limits of deontic and scalar-reward approaches and propose a virtue-centered research agenda with four components: social learning in multi-agent RL to acquire virtue-like dispositions from imperfect but normatively informed exemplars; multi-objective formulations that preserve value conflicts; affinity-based regularization toward updatable virtue priors; and operationalizing ethical traditions beyond the standard Western triad as control signals with explicit value assumptions.

Social RL as Deliberative Infrastructure. Building on the roadmap, I develop Social Reinforcement Learning (Social RL) as a framework for restoring feedback integrity, formalized as a gradient alignment condition: a pipeline preserves epistemic (respectively, ethical) feedback integrity if the inner product between the proxy gradient and the latent epistemic (normative) gradient is non-negative (Ghasemi and Crowley 2026c). I show that social feedback introduces a positive gradient component that can rotate the effective training gradient back toward truth and ethical behavior even when the proxy gradient alone is misaligned, identify three political risks (majority capture, opponent shaping, and social washing), and derive governance recommendations for researchers, developers, and policymakers.

Epistemic Source Alignment (ESA). A specific vulnerability arises when the social layer itself is sycophantic: evaluators reward whatever the agent already does rather than ground truth, so the observed objective permanently separates from the latent one (Ghasemi and Crowley 2026a). I prove that under naive aggregation, any no-regret algorithm on the observed reward incurs linear regret on the latent reward when a sycophantic majority exists. I propose ESA

(Ghasemi and Crowley 2026b), which uses sparse reference signals (e.g., hard safety constraints or factual anchors) to audit evaluator reliability in real time via a multiplicative weights update. Under an informational dominance condition, ESA concentrates trust on truthful sources exponentially fast, recovering the optimal policy even when 80% of the social layer is misaligned.

The Forager’s Dilemma. I introduce the Forager’s Dilemma, a multi-agent environment that couples epistemic and ethical failure modes in a shared resource economy: agents decide whether to communicate honestly or deceptively and whether to harvest sustainably or destructively, with both axes economically coupled. Across four governance regimes with tabular Q-learning and IPPO agents in PettingZoo, DERL, which combines intrinsic alignment priors with emergent social enforcement, achieves approximately 1% deception, 76–90% sustainable gathering, and the highest cumulative reward. Crucially, peer-only enforcement is unstable: free-riding erodes it against individual defectors, and it grows markedly more vulnerable when higher-capacity agents form adversarial coalitions.

Planned Work

My ongoing research asks how Social RL can serve as a principled mechanism for producing better-aligned and more genuinely virtuous agents, along two axes.

The first concerns *social feedback as an alignment signal*. Peer critique, reputation, and sanction can partially correct a misaligned training gradient; I plan to characterize when this correction is reliable. Under what population compositions, interaction structures, and aggregation strategies does Social RL steer agents toward latent normative objectives rather than socially dominant but misspecified norms? This requires formal analysis and controlled experiments in which the ethical quality of the social environment is varied independently of task reward, to establish where Social RL improves on static scalar alignment and where it remains fragile.

The second, more foundational, axis concerns *intrinsic ethical development through observation*. Rather than treating ethics as an externally imposed objective, I ask whether agents can acquire stable ethical dispositions simply by being situated among agents that already behave virtuously. Social learning research shows that observing others transmits behavioral patterns without explicit supervision (Ndousse et al. 2021; Bhoopchand et al. 2023); I plan to extend this to the ethical domain. Can an agent, through sustained observation of honest and cooperative peers, internalize those dispositions as a prior that persists when the social context changes, incentives shift, or exploitative alternatives appear? If so, this would point toward agents that are ethical not because deviation is penalized, but because virtuous behavior has become their default mode of engagement (Ghasemi and Crowley 2025). Testing this requires environments where ethical exemplars and naive learners co-exist, measuring disposition retention under partner-swap and incentive-flip interventions, and distinguishing genuine internalization from mimicry that collapses under pressure.

Fit to Conference Themes

This program speaks to AIES themes in AI ethics, multi-agent systems, and AI safety. It contributes (i) a framework unifying two prominent failure modes under one structural diagnosis, (ii) algorithms with theoretical guarantees and empirical validation, (iii) a methodological critique of standard MARL evaluation, and (iv) governance recommendations grounded in technical analysis and political philosophy. I aim to engage the AIES community on understanding alignment not merely as an engineering problem, but as an ongoing social and institutional practice.

References

- Allen, C.; Smit, I.; and Wallach, W. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3): 149–155.
- Bhoopchand, A.; Brownfield, B.; Collister, A.; Dal Lago, A.; Edwards, A.; Everett, R.; Fr  chette, A.; Gitahy Oliveira, Y.; Hughes, E.; Mathewson, K. W.; et al. 2023. Learning few-shot imitation as cultural transmission. *Nature Communications*, 14(1): 7536.
- Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.
- Crisp, R. 2014. *Aristotle: Nicomachean Ethics*. Cambridge University Press.
- Everitt, T.; Hutter, M.; Kumar, R.; and Krakovna, V. 2021. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27): 6435–6467.
- Ghasemi, M.; and Crowley, M. 2025. Toward virtuous reinforcement learning: A critique and roadmap. ArXiv preprint arXiv:2512.04246. Accepted to the Machine Ethics Workshop, AAAI 2026.
- Ghasemi, M.; and Crowley, M. 2026a. Learning When to Trust in Contextual Bandits. *arXiv preprint arXiv:2603.13356*.
- Ghasemi, M.; and Crowley, M. 2026b. Objective Decoupling in Social Reinforcement Learning: Recovering Ground Truth from Sycophantic Majorities. *arXiv preprint arXiv:2602.08092*.
- Ghasemi, M.; and Crowley, M. 2026c. Rethinking AI Alignment: From Static Rewards to Social Reinforcement Learning. In *Pluralistic Alignment Workshop at ICML 2026*.
- Ndousse, K. K.; Eck, D.; Levine, S.; and Jaques, N. 2021. Emergent social learning via multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 7991–8004. PMLR.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.