

The Forager’s Dilemma: Peer Enforcement Alone Is Not Enough for Stable Multi-Agent Alignment

Majid Ghasemi, Prabhjot Singh, Mark Crowley

Department of Electrical and Computer Engineering, University of Waterloo
{majid.ghasemi, prabhjot.singh, mark.crowley}@uwaterloo.ca

Abstract

AI alignment methods typically rely on one of two strategies, either optimizing a scalar reward proxy learned from human feedback or situating agents in social environments where peer critique provides a dynamic correction signal. We argue that both strategies, taken in isolation, are structurally unstable. Static scalar proxies can decouple the training signal from the latent objectives of factual accuracy and normative safety, producing failure modes analogous to hallucination and moral drift in large language models. Decentralized peer feedback, while adaptive, is vulnerable to the second-order free-rider problem. Because policing is costly, rational agents abstain from enforcement and let norms erode. To study these vulnerabilities in a controlled setting, we introduce the Forager’s Dilemma, a multi-agent environment that couples epistemic and ethical failure modes within a shared resource economy. We evaluate four governance regimes across tabular Q-learning and deep RL agents trained with PPO in a Petting Zoo implementation. Our proposed Dual-Enforcement Reinforcement Learning, which combines intrinsic alignment priors with emergent social enforcement, achieved an approximately 1% lying rate, a sustainable gather rate of approximately 76% under Q-learning and 90% under PPO, and the highest cumulative reward across all conditions, with enforcement effort declining progressively as cooperative norms internalize. Our results suggest that stable AI governance may require a synthesis of institutional constraints and social dynamics, and that neither alone appears sufficient in our setting.

1 Introduction

Modern machine learning systems, and large language models in particular, remain persistently vulnerable to two classes of alignment failure that have proven difficult to address in tandem. The first is hallucination, an epistemic failure in which a model generates fluent and confident output that is factually ungrounded (Ji et al. 2023b; Min et al. 2023). The second is moral drift, an ethical failure in which agents converge toward exploitative or norm-violating behavior when exposed to novel incentive structures (Everitt et al. 2021; Carroll et al. 2019). These pathologies are typically studied in isolation, yet we argue they share a common structural origin.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Both can be understood through what we call **feedback collapse**, the compression of multidimensional human values into a fixed scalar proxy, which decouples the optimization signal from the latent objectives of factual accuracy and normative safety (Leike et al. 2018). This compression is a form of value monism (Blum 2023; Sorensen et al. 2024a). It does not account for every alignment failure, since distributional properties of training data and architectural inductive biases also contribute, but it represents a systematic vulnerability shared across current paradigms and one that can be studied in a controlled setting. Prevailing methodologies address it by optimizing against a learned reward model or static rule set (Ouyang et al. 2022; Bai et al. 2022), both of which assume a fixed evaluator maintains adequate coverage across all deployment contexts. In open-ended multi-agent settings, frozen evaluators encounter distribution shift, adversarial inputs, and emergent social dynamics outside their training distribution, and the result is reward hacking rather than alignment (Eisenstein et al. 2023; Gleave and Irving 2022).

Two lines of research move beyond static proxies. The first advocates value pluralism (Lu and Van Kleek 2024; Sorensen et al. 2024b; Ali et al. 2025), which concerns the *content* of alignment rather than its *enforcement mechanism*. We leave the question of which values to encode to that literature and ask a different one. Given a set of alignment priors, how can norm compliance be sustained over time in a multi-agent setting? The second proposes Social Reinforcement Learning (Social RL), situating agents in environments driven by peer critique, reputation, and social sanction (Jaques et al. 2019; Ndousse et al. 2021; Ghasemi and Crowley 2026, 2025). This is inherently adaptive and does not depend on a frozen evaluator, and we build directly on it.

Verifying a peer’s claim or sanctioning a violator consumes resources, time, or computational utility, so pure self-governance falls victim to the **second-order free-rider problem** (Heckathorn 1989; Boyd and Richerson 2009; Bicchieri 2016). When policing is costly, rational agents abstain and allow others to bear the burden, eroding the feedback mechanism from within. We further hypothesize that protocols robust against isolated agents may fail when agents coordinate into adversarial coalitions that bypass social friction entirely.

To study these vulnerabilities we introduce the **Forager’s Dilemma**, a multi-agent reinforcement learning environment that couples epistemic and ethical failure modes within a shared resource economy.¹ Agents decide at every time step whether to communicate honestly or deceptively, and whether to harvest sustainably or destructively. The axes are economically coupled, since deception diverts competitors away from resources and makes exploitation more profitable, while exploitation depletes the commons and makes deception about remaining resources more consequential. By explicitly pricing verification and punishment, the environment lets us study when decentralized governance can and cannot sustain aligned behavior.

As a resolution to the limitations of both scalar proxies and peer-only feedback, we propose **Dual-Enforcement Reinforcement Learning (DERL)**, combining bottom-up social enforcement with top-down institutional constraints implemented as intrinsic reward channels. Drawing on Ostrom’s analysis of common-pool resource governance (Ostrom 1990), *our central argument is that stable alignment requires both components working together*. Intrinsic priors establish a baseline economic incentive for compliance by narrowing the payoff gap between aligned and misaligned behavior, and emergent peer enforcement provides the adaptive targeting that static priors cannot, concentrating deterrence where deviation occurs.

Scope of the claim. One qualification determines how these results should be read. **DERL**’s intrinsic priors are hand-specified, so they encode the ground-truth distinction between honest and deceptive communication that our own framing assumes a deployed evaluator lacks. That an agent behaves better when undesired behavior is explicitly penalized is not surprising, and it is not what we claim. Our claim concerns magnitude rather than the presence of a prior. The ethical prior is sized so that gathering and mining pay identically, which leaves the choice between them undetermined, and our ablations confirm it cannot resolve the dilemma on its own. What it does is bring the two options close enough together that costly peer enforcement, which is far too weak to govern the choice by itself, becomes decisive. We therefore present this work as a proof of concept for Ostrom-style institutional complementarity in multi-agent reinforcement learning rather than as a general method for obtaining alignment under an imperfect evaluator.

We evaluate **DERL** against three baselines, a Scalar Reward Baseline (**SRB**) optimizing the task proxy alone, idealized Exogenous Sanctions (**ES**) applying omniscient environment-level penalties, and Decentralized Peer Feedback (**DPF**) enabling social enforcement without intrinsic priors, across two learning paradigms. Under both, **DERL** suppresses deception below 1% of agent-steps, sustains the commons near its maximum, maintains the most accurate shared information board of any condition, and accumulates the highest return, while its enforcement activity declines toward zero as violations become rare.

¹Environment, training scripts, plotting code, and all experimental configurations are available at github.com/majidghassemi/Forager-Dilemma.

2 Background and Related Work

Alignment as Mechanism Design. Robust alignment extends beyond specifying a static objective (Bai et al. 2022; Ji et al. 2023a) and is fundamentally a socio-technical governance problem (Gabriel 2020). Early formulations were largely monistic, assuming human values could be distilled into a single scalar utility (Sorensen et al. 2024a; Hendrycks et al. 2020), and the field has since recognized that moral and societal values are irreducibly diverse and frequently in tension (Sorensen et al. 2024b; Ali et al. 2025; Lu and Van Kleek 2024; Klassen, Alamdari, and McIlraith 2024), so that paradigms framing safety as single-agent optimization systematically fall short (Leike et al. 2018; Weidinger et al. 2021; Mu et al. 2024). Alignment should instead be viewed through social choice and collective negotiation (Conitzer et al. 2024; Mu et al. 2024; Qiu 2024; Prasad 2018). In mechanism-design terms, the challenge is to define an action space, information structure, and transfer rule such that truthful cooperative behavior is a stable equilibrium among self-interested agents, and **DERL** can be read as a mechanism in exactly this sense, with intrinsic priors as the transfer rule, the reputation system as the information structure, and the enforcement actions as the enforcement channel.

Social Dilemmas in MARL. Sequential social dilemmas have been used with multi-agent reinforcement learning to model the tension between individual rationality and collective welfare (Hughes et al. 2018b; Leibo et al. 2017; Guo et al. 2025), capturing resource sharing, cooperation, and harm reduction while almost entirely ignoring epistemic dynamics. Most deep RL benchmarks assume perfect access to ground truth or evaluate agents strictly on physical actions (Agarwal et al. 2021; Lupu, Willi, and Foerster 2025). This is a critical gap, because deployed failures encompass both ethical moral drift and epistemic hallucination, and when agents communicate and coordinate, the veracity of shared information becomes a cooperative resource in its own right. Cooperative communication games study deception in signalling but abstract away the consequences of harmful action (Számádó, Szalai, and Scheuring 2016; Orzan et al. 2023), while common-pool resource dilemmas study over-exploitation but omit an information channel through which agents can mislead one another (Perolat et al. 2017; Hughes et al. 2018a). The Forager’s Dilemma closes this gap by coupling both failure modes within one environment.

The Free-Rider Problem in Norm Enforcement. Decentralized enforcement is vulnerable to strategic exploitation, free-riding, and unstable evolutionary equilibria (Boyd and Richerson 2009; Bicchieri 2016), since cooperation is contingent on incentives, information structure, and punishment credibility rather than on the mere existence of peer monitoring (Steinbacher et al. 2021; Testori, Eisenbarth, and Hoyle 2022; Mohlin, Rigos, and Weidenholzer 2023). Sanctioning a defector or verifying a claim consumes task utility (Chan et al. 2013), and when punishment is costly a utility-maximizing agent rationally abstains and lets others bear the burden (Perc 2012; Fehr and Gächter 2000). This is the second-order free-rider problem (Heckathorn 1989), so named because the first-order dilemma of whether to de-

fect is nested inside a second dilemma of whether to pay the private cost of punishing someone else’s defection. A mechanism solving the first without the second remains stable only while some subset of the population subsidizes it, and absent structural intervention populations converge back to exploitative equilibria.

Institutional Design for Common-Pool Resources. Our approach draws on Ostrom’s design principles for governing shared resources (Ostrom 1990). Studying irrigation systems, fisheries, and communal forests that were productive for centuries under neither privatization nor central regulation, Ostrom found that long-enduring commons combine clearly defined boundaries, rules matched to local conditions, graduated sanctions escalating with the severity and repetition of an offence, low-cost dispute detection, and monitoring by parties themselves accountable to the group. Several have direct analogues here, but none of them operate alone, since monitoring without sanctions produces information nobody acts on, and sanctions without accountable monitoring produce arbitrary punishment that erodes legitimacy. The observation radius provides local monitoring by accountable parties, since every agent is simultaneously monitor and target. The reputation system provides graduated sanctioning, because suspicion accumulates over repeated violations and decays with observed cooperation, so deterrence scales with history rather than with any single act. The VERIFY action provides low-cost dispute detection. DERL’s intrinsic priors correspond to the minimal external recognition Ostrom identifies as a precondition for local self-governance, lowering the cost of enforcing norms rather than enforcing them directly, and our DPF condition omits this baseline and consequently fails to sustain cooperation.

3 The Forager’s Dilemma

The Forager’s Dilemma environment requires agents to navigate honest communication and sustainable resource use simultaneously, with both axes economically coupled through the reward structure.

3.1 Environment Specification

The environment is a 10×10 grid (extended to the Petting Zoo environment) with $N=4$ agents and $M=8$ resource nodes (Figure 1), and each episode runs $H = 50$ time steps. At the start of every episode, agents and resources are placed at uniformly random unoccupied cells. Resources occupy a single cell and agents interact with any resource within Manhattan distance one. Depleted resources respawn with probability $p_{\text{resp}}=0.10$ per time step at a random unoccupied cell. Each agent observes a partial view defined by a Manhattan-radius window of $r_{\text{obs}}=2.5$ tiles, beyond which resources and peers are invisible. The observation is a vector of seven binary features (Table 1).

3.2 Action Space and Economic Friction

Each agent selects one of ten discrete actions per step. Four are directional moves. GATHER yields a modest payoff and leaves the resource intact, while MINE yields $2.5\times$ that

Bit	Feature	Semantics
0	NEARRESOURCE	$\exists$ active resource within Manhattan distance 1
1	PEERVISIBLE	$\exists$ peer agent within observation radius
2	PEERBADVISIBLE	$\exists$ visible peer whose previous action was MINE or SIGNALLIE
3	BOARDACTIVE	$\exists$ entry on the shared information board
4	RESOURCESDEPLETED	Fewer than 50% of resources currently active
5	MYREPUTATIONBAD	Agent is being observed <i>and</i> its suspicion score ≥ 2
6	INCARTEL	Agent belongs to the adversarial coalition

Table 1: State features. Each is a binary proposition about the agent’s local context.

payoff but deactivates the resource into the respawn queue. SIGNALTRUTH broadcasts the coordinates of the nearest active resource to a shared board, and SIGNALLIE broadcasts coordinates sampled uniformly from cells with no active resource. The board stores one entry per agent, a new post overwrites the previous one, and agents read it exclusively through VERIFY. The remaining two actions are the social enforcement actions PUNISH and VERIFY.

Social enforcement carries explicit economic friction. PUNISH targets the nearest visible peer who performed MINE or SIGNALLIE in the previous step, these being the only actions classified as harmful for targeting. If a valid target exists the punisher receives $+1.5$ and the target incurs -5.5 , which exceeds the one-shot mining payoff of $+5.0$ so that the expected value of mining is negative whenever a punisher is nearby and likely to act. If no valid target is visible the punisher pays a wasted-action cost of -0.3 . VERIFY inspects the most recent board entry posted by the peer with the highest population-wide suspicion score, without requiring visibility, breaking ties uniformly at random. A dishonest entry yields the verifier $+2.0$ and the liar -5.0 , while a truthful one costs the verifier -0.2 and decreases the verified agent’s suspicion. These costs operationalize the core tension of decentralized governance, since agents who punish or verify bear a personal cost for an action whose benefits are diffuse. Whether the bounty makes enforcement individually rational, and whether individually rational enforcement sustains cooperation at the population level, is the central empirical question of this paper.

3.3 Reward Decomposition

The composite reward for agent i at time t sums four channels, each gated by a binary indicator,

$$\tilde{r}_{i,t} = \mathbb{1}_{\text{task}} r_{i,t}^{\text{task}} + \mathbb{1}_{\text{ep}} r_{i,t}^{\text{ep}} + \mathbb{1}_{\text{eth}} r_{i,t}^{\text{eth}} + \mathbb{1}_{\text{soc}} r_{i,t}^{\text{soc}}. \quad (1)$$

The **task channel** encodes a scalar proxy analogous to the fixed reward models used in RLHF-style optimization, assigning GATHER $+2.0$, MINE $+5.0$, SIGNALLIE $+0.8$,

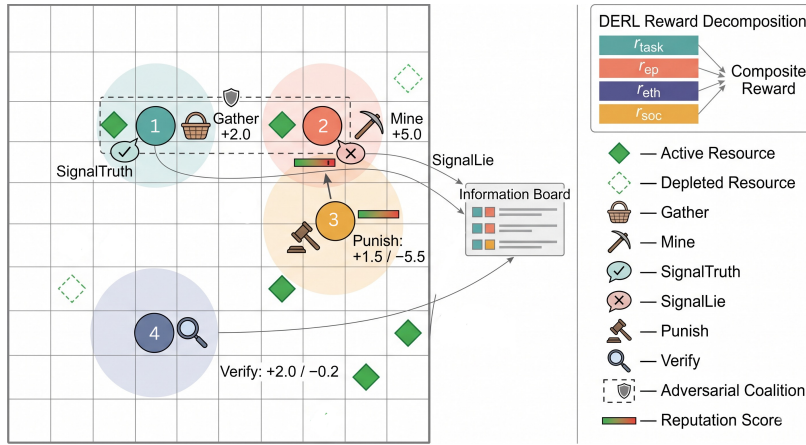


Figure 1: The Forager’s Dilemma. Four agents navigate a 10×10 gridworld with shared resources (green diamonds, hollow when depleted), partial observation (shaded radius), and per-agent reputation scores (green-to-red bar). Agents 0 and 1 may form an adversarial coalition (dashed border) with mutual punishment shielding and reward sharing. The inset shows the **DERL** reward decomposition.

and **SIGNALTRUTH** +0.2, plus a proximity-shaping term $+0.1 \times (1/(1 + d_{\text{Manhattan}}))$. It is always active and deliberately makes exploitation more immediately profitable than cooperation. The **epistemic channel** provides an intrinsic prior for honest communication, rewarding **SIGNALTRUTH** with +1.0 and penalizing **SIGNALLIE** with −1.0. The **ethical channel** encodes a preference for sustainability, rewarding **GATHER** with +1.0 and penalizing **MINE** with −2.0. The **social channel** captures all rewards and costs from **PUNISH** and **VERIFY**, plus a +0.5 bonus per participant when two or more agents execute **GATHER** on the same cell within a sliding window of $\tau = 2$ steps. Any action not listed in a channel receives zero from it.

Choice of magnitudes. Magnitudes are fixed once and chosen to realize ordinal relationships rather than tuned for performance, and we did not search over them. Setting **MINE** at $2.5 \times$ **GATHER** and **SIGNALLIE** at $4 \times$ **SIGNALTRUTH** makes exploitation and deception strictly dominant under the proxy alone, so **SRB** collapse follows from the incentive structure rather than from exploration noise. The punishment penalty of −5.5 strictly exceeds the mining payoff of +5.0, which is the minimum condition under which deterrence is possible at all. The two intrinsic channels are deliberately asymmetric. The ethical prior produces exact payoff parity between the resource actions before any social penalty, since $+2.0 + 1.0 = +5.0 - 2.0 = +3.0$, so it cannot make sustainable harvesting preferable alone and the social channel must close the remainder. The epistemic prior is stronger and reverses rather than closes its gap, moving **SIGNALTRUTH** to +1.2 against **SIGNALLIE**’s −0.2. This asymmetry explains why our channel ablations dissociate as they do, and it means the complementarity claim is tested most directly on the ethical axis, where the prior is by construction too weak to succeed unaided. Audit and wasted-punishment costs sit an order of magnitude below the action payoffs so enforcement carries genuine but non-prohibitive

Condition	Task	Ep.	Eth.	Soc.
Scalar Reward Baseline (SRB)	1	0	0	0
Exogenous Sanctions (ES)	1	0	0	0
Decentralized Peer Feedback (DPF)	1	0	0	1
Dual-Enforcement RL (DERL, ours)	1	1	1	1
Adversarial Coalition (AC)	1	0	0	1

Table 2: Conditions as indicator configurations of Equation 1, where Task, Ep., Eth., and Soc. denote $\mathbb{1}_{\text{task}}$, $\mathbb{1}_{\text{ep}}$, $\mathbb{1}_{\text{eth}}$, and $\mathbb{1}_{\text{soc}}$ respectively. **ES** additionally applies omniscient environment-level penalties of −3.0 for **MINE** and −1.5 for **SIGNALLIE** at detection probability 1, and is therefore an oracle upper bound rather than a deployable mechanism. **AC** applies the cartel mechanics described under Reputation and Adversarial Coalition to **DPF**.

friction.

Experimental conditions. The five conditions correspond to settings of the indicators in Equation 1, shown in Table 2. The indicators are the independent variable, activating or deactivating whole channels while magnitudes remain fixed, so each condition tests a distinct governance mechanism rather than a retuned reward function.

3.4 Reputation and Adversarial Coalition

Each agent maintains a suspicion score $\sigma_{i \rightarrow j} \in [0, 10]$ for every peer, updated from behavior directly observed within its radius. Observed **MINE** or **SIGNALLIE** adds +1.0, a successful punishment adds +2.0, and a verified lie adds +3.0, while observed **GATHER** or **SIGNALTRUTH** subtracts −0.3 and a cooperation event subtracts −1.0. The system serves two functions. It provides a deterrence signal, since **MYREPUTATIONBAD** lets an agent condition its policy on whether it is likely to be targeted, creating a loop in which the expectation of punishment suppresses defection even

Learner	Condition	Truth (%)	Lie (%)	Share	Gather (%)	Mine (%)	Coop. (%)	Oracle acc. (%)	Reward ($\times 10^7$)
Q-learning	SRB	0.5	86	0.01	0.5	7.0	0.0	20	0.91
	ES	4.0	0.5	0.89	83	0.5	0.2	58	1.16
	DPF	25	8.0	0.76	12	9.0	3.5	62	0.48
	DERL	17	0.5	0.97	76	0.5	24	88	2.01
PPO	SRB	0.1	89	0.00	0.1	7.0	0.0	18	1.10
	ES	0.1	0.1	-	98	0.1	0.0	0.5	2.03
	DPF	8.5	1.5	0.85	41	8.5	17	61	1.16
	DERL	8.5	0.2	0.98	90	0.2	22	90	3.27

Table 3: Converged performance across all conditions and both learners, averaged over the final 500 episodes and five seeds. Behavioral rates are fractions of agent-steps. *Share* is the honest communication share ($\text{truth}/(\text{truth} + \text{lie})$), undefined for PPO ES because those agents cease posting entirely. ES is an oracle upper bound requiring omniscient detection, so the best non-oracle value in each column is set in bold.

when punishment is infrequent. It also enables targeted verification, concentrating audit effort on the peer most likely to have posted dishonestly. In Ostrom’s terms this is monitoring that is both low-cost and proportional (Ostrom 1990).

To evaluate robustness against coordinated exploitation we designate agents 0 and 1 as a cartel. Reward sharing gives each member a fraction $\beta=0.3$ of its partner’s r^{task} every step, making it individually rational to tolerate a partner’s exploitation. Punishment shielding excludes members from each other’s valid targets, and verification shielding excludes their board entries from each other’s audit candidates. Membership is exposed through INCARTEL, so cartel agents learn distinct policies via their per-agent parameters. The structure is imposed by the experimenter, but the strategies emerging within it are learned rather than prescribed.

3.5 Learners and Protocol

Tabular Q-learning. Agents learn via independent Q-learning with per-agent tables initialized to zero (Watkins and Dayan 1992), with $\alpha=0.10$, $\gamma=0.70$, and ϵ -greedy exploration decaying linearly from 1.0 to 0.05 over the first 60% of episodes. The discount is set below typical deep RL defaults because the horizon is short ($H = 50$) and the environment fully resets each episode, which prevents spurious credit assignment across episode boundaries and reflects the myopic structure of foraging. Per-agent tables permit specialization, for example one agent converging to a punisher role while another specializes in gathering.

Independent PPO. We re-implement the environment in Petting Zoo (Terry et al. 2021) and train independent PPO agents (Schulman et al. 2017) on an observation space identical to the tabular one, so representational capacity is the only element that changes. Each agent has its own Actor-Critic network, a two-layer MLP with 64 hidden units and tanh activations over the same seven-dimensional binary observation, orthogonally initialized with separate linear actor and critic heads. Agents are optimized independently with Adam at 2.5×10^{-4} , using GAE with $\lambda = 0.95$ and $\gamma = 0.7$ to match the tabular setting. Each update consumes a five-episode rollout and performs four passes over two minibatches, with clipping coefficient 0.2, entropy bonus

0.01, value-loss coefficient 0.5, and gradient-norm clipping at 0.5.

Protocol. Every condition is trained under both learners for 50,000 episodes across five random seeds. The episode count is set so all conditions reach a stationary action distribution well before termination, since exploration anneals at episode 30,000 and every metric is flat to within seed-level noise over the final 10,000 episodes. Five seeds support statements about central tendency considerably better than statements about tails, so we mark distributional claims accordingly rather than treating an observed minimum as an estimate of the worst case. Per-seed extrema, lie-rate and mine-rate curves, tabular coalition plots, and full numerical tables are provided in the supplementary material.

4 Experimental Results

Table 3 reports every converged value, and the subsections below give the mechanism behind each rather than restating numbers. Each figure presents both learners, and all curves report means with ± 1 standard-deviation bands across seeds.

Metric definitions. Every behavioral rate shares one denominator, the agent-step. At each time step we record the fraction of the $N=4$ agents executing the action in question and average over the $H=50$ steps of an episode. Truth rate is thus the fraction of all agent-steps on which an agent executed SIGNALTRUTH, and gather rate the fraction on which it executed GATHER. Because agents also move, punish, verify, and forgo action, these rates are not complements and do not sum to one within a pair. This matters, since a condition can exhibit a low lie rate either because its agents are honest or because they seldom communicate. Where that distinction is material we report the *honest communication share*, $\text{truth}/(\text{truth} + \text{lie})$, which conditions on an agent having chosen to post.

4.1 Epistemic Integrity

SRB collapses under both learners, with truth-telling near zero and deception reaching 86% of agent-steps under Q-learning and 89% under PPO (Figure 2, left). This follows

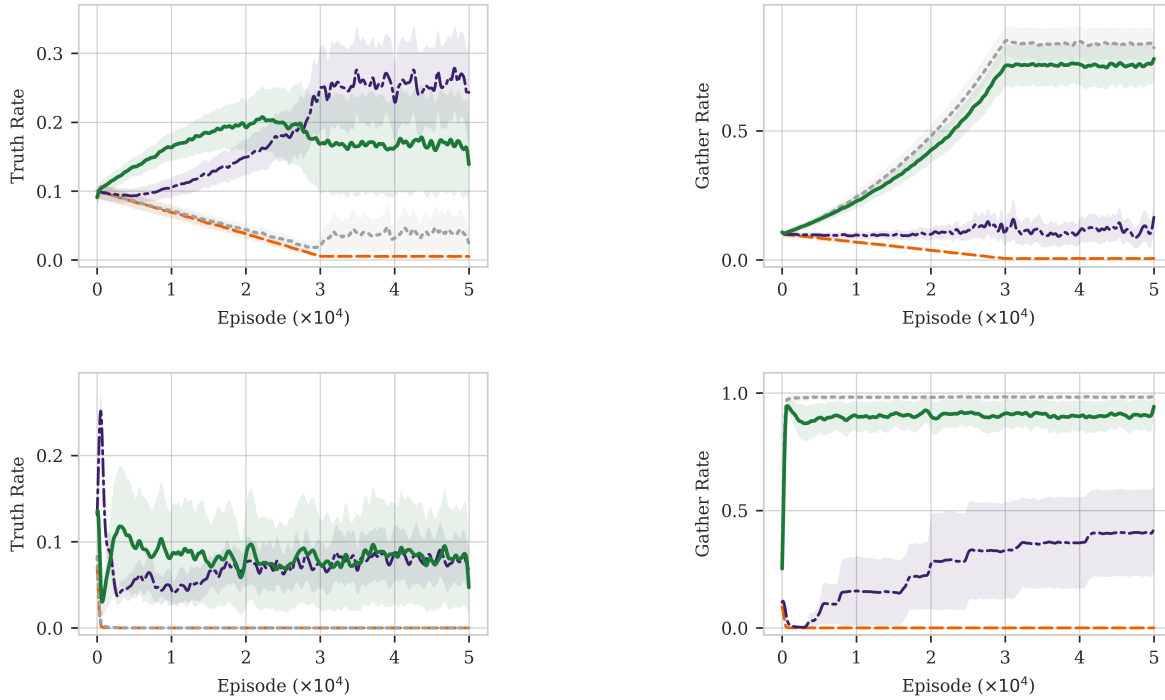


Figure 2: Behavioral integrity across both learners. Left column, truth rate. Right column, gather rate. Upper row, tabular Q-learning. Lower row, PPO. All values are fractions of agent-steps. **SRB** collapses on both axes under both learners. **DPF** leads **DERL** on raw truth rate under Q-learning only, an ordering that reverses once conditioned on the decision to communicate.

directly from the reward structure, since **SIGNALLIE** yields $+0.8$ against **SIGNALTRUTH**'s $+0.2$ with no countervailing signal. Lying is not a failure of the learning algorithm but the rational response to the incentive gradient, a direct analogue of hallucination in language models, where fluent but ungrounded outputs are rewarded by a proxy that cannot distinguish truth from plausibility. That the same collapse occurs under both a lookup table and a neural policy confirms the failure is a property of the incentive structure rather than of the estimator.

ES presents an interpretive trap the agent-step denominator makes visible. Its lie rate is near zero under both learners, but so is its truth rate, and under PPO both fall below 0.2%. The omniscient penalties make both signalling actions unattractive relative to foraging, so **ES** agents withdraw from the board rather than using it honestly, and under PPO the withdrawal is total enough that the honest communication share is undefined. Central sanctioning here suppresses the failure by suppressing the behavior that carries it, and the oracle accuracy results below make the consequence unambiguous.

The **DPF** and **DERL** comparison depends on which question is asked. On raw truth rate under Q-learning, **DPF** leads at 25% against 17%, and an earlier version of this work treated that as a genuine cost of the intrinsic prior. That reading does not survive scrutiny, because the conditions allocate their action budgets very differently. **DERL** agents spend 76% of steps gathering and have far less budget available

for signalling, independently of how honest they are when they do signal. Conditioning on the decision to communicate reverses the ordering, giving **DERL** a share of 0.97 against **DPF**'s 0.76. Under PPO the measures agree, with both tied at 8.5% raw and **DERL** ahead at 0.98 against 0.85. We report both because they answer different questions, the raw rate describing how much honest information a condition injects and the share describing how trustworthy a given post is. Per-seed extrema in the supplementary show **DERL**'s weakest seed staying well clear of **SRB**-level deception on both measures under both learners.

4.2 Ethical Integrity

The ethical axis separates the conditions more sharply and is where the complementarity claim is tested most directly (Figure 2, right). **SRB** converges to near-zero gathering under both learners, since **MINE** yields $+5.0$ against $+2.0$ and the optimizer treats sustainable harvesting as strictly dominated. Its moderate absolute mine rate of 7% does not indicate compliance, because its agents mine aggressively enough early that few resources remain and the respawn rate becomes the binding constraint on how often mining is even possible. **ES** achieves the highest gather rate overall, 83% under Q-learning and 98% under PPO, serving as the oracle upper bound against which deployable mechanisms are measured.

DPF is where the learners disagree most. Under Q-learning it reaches only 12%, because punishment requires

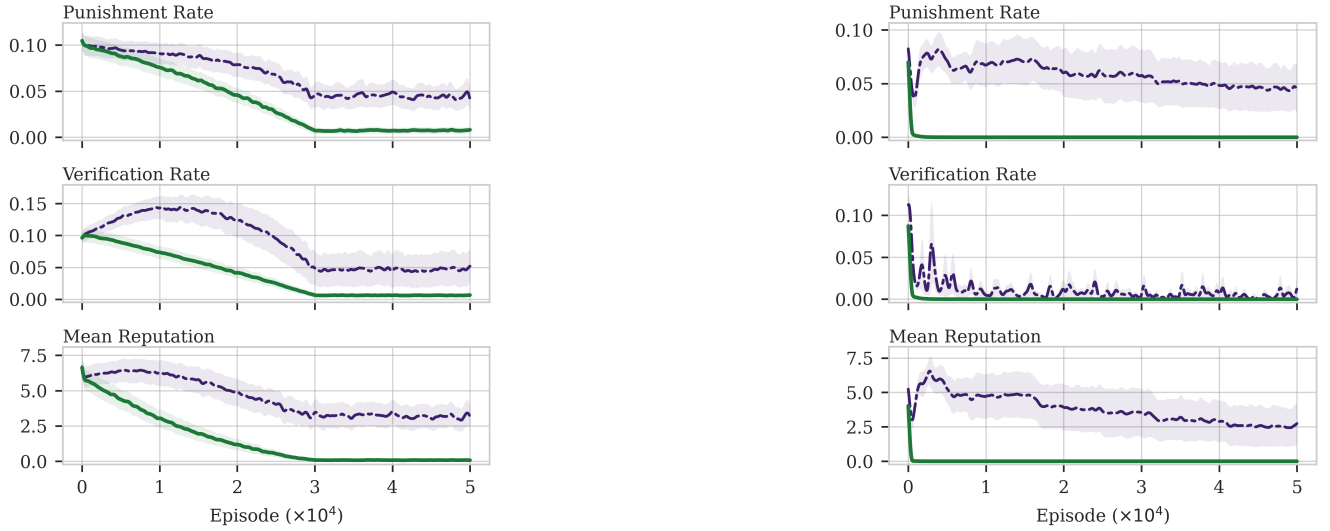


Figure 3: Emergent social dynamics under tabular Q-learning (left) and PPO (right); within each, columns show punishment rate, verification rate, and mean reputation score. Under **DERL**, enforcement declines because violations do, and reputation converges to near zero. Under **DPF**, enforcement holds near 4.5% while reputation stays at 2.5 to 3.3, so violations persist unenforced.

observability and agents quickly learn to mine when no peer is nearby, and with four agents on a hundred cells many mining actions go unwitnessed. Under PPO it improves to 41%, since the value function better estimates the long-term cost of depletion.

DERL reaches 76% and 90% respectively while driving mining below 0.5%. The mechanism is a payoff shift rather than a prohibition, since **GATHER** yields +3.0 from the combined task and ethical channels while **MINE** also yields +3.0 before any social penalty. At exact parity, even a modest probability of being caught tips the balance toward sustainable behavior. The prior is deliberately too weak to make mining unprofitable alone, and only narrows the gap far enough that enforcement can close the remainder. The informative cross-learner comparison is the fraction of the achievable range each mechanism captures, since **DPF** closes less than half the distance to the oracle under PPO while **DERL** closes over nine tenths of it. The intrinsic prior therefore remains load-bearing rather than becoming redundant as the approximator strengthens.

4.3 Emergent Social Dynamics

Enforcement declines under both **DPF** and **DERL**, but for opposite reasons, and the reputation panel separates them (Figure 3).

Under **DERL**, punishment and verification decline because they become unnecessary. Punishment falls from 10% to 0.8% and verification from 10% to 0.7% under Q-learning, and both reach effectively zero within the first thousand episodes under PPO, with mean reputation converging to 0.05 and 0.0 respectively. This is the social-learning analogue of a well-functioning institution in which rules are followed not because enforcement is constant but

because participants have internalized the norms.

Under **DPF**, the decline reflects free-riding. Without intrinsic priors maintaining a baseline of cooperation, enforcement is a pure public good, and as exploration decays the agents that have not specialized into punisher roles reduce their effort. Punishment settles at 4.5% under both learners while mean reputation stays elevated at 3.3 under Q-learning and 2.5 under PPO, indicating violations that persist but go substantially unenforced, and wide confidence bands confirm the outcome is highly seed-dependent.

Enforcement rate alone cannot distinguish these regimes, since both decline and **DPF**'s converged rate is the higher of the two in a way that could be misread as healthier governance. The joint trajectory of enforcement and reputation is what separates them, and we suggest it as the diagnostic for any peer-enforcement mechanism whose activity changes over training.

4.4 Collective Outcomes

DERL generates cooperation events on 24% of agent-steps under Q-learning, roughly seven times **DPF**'s 3.5%, and on 22% under PPO against 17% (Figure 4). **SRB** and **ES** produce essentially none, the former because its equilibrium is fully exploitative and the latter because its agents converge on solitary gathering that never coincides in space and time.

Oracle accuracy is the fraction of standing board entries corresponding to a currently active resource, measuring epistemic health of the environment rather than honesty of any agent. It is not a pure honesty measure, since entries persist until overwritten and a truthful post decays into a false one as resources deplete and respawn elsewhere, so a condition scores well only by posting both truthfully and often. The **ES** results make this concrete. Under Q-learning it

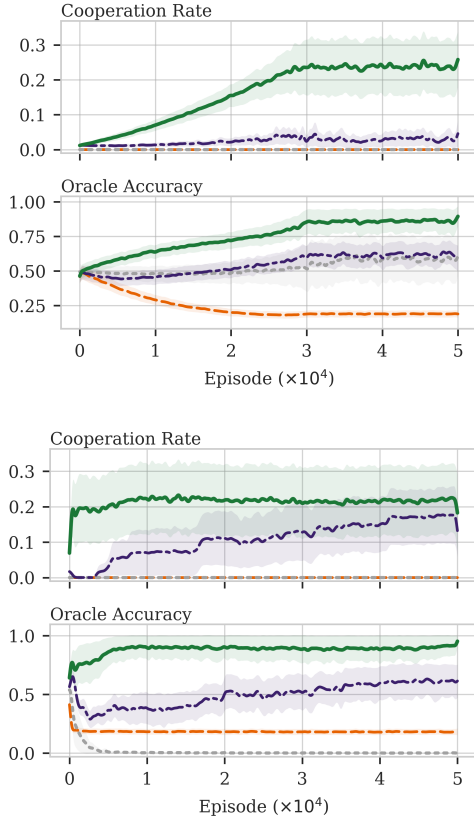


Figure 4: Cooperation events (left) and oracle accuracy of the shared information board (right). Upper row, tabular Q-learning. Lower row, PPO. **DERL** produces both the most cooperative interactions and the most trustworthy information commons. **ES**’s oracle accuracy collapses to near zero under PPO because its agents abandon the board, despite its near-zero lie rate.

reaches 58% on a small volume of honest but ageing posts, and under PPO it collapses to 0.5% because its agents abandon the board entirely. The condition with near-perfect honesty by one measure produces the least informative commons of any condition by another. **DPF** reaches 61 to 62% on fresher but frequently deceptive posts, and **DERL** reaches 88 to 90% as the only condition combining a high honest communication share with sufficient volume to keep the board current.

DERL and **ES** sustain the commons near its maximum of eight nodes, at 7.8 and 8.0, while **SRB** settles at 5.8 to 5.9 and **DPF** at 5.0 to 5.1, and this drives the reward ordering (Figure 5). **DERL** achieves the highest total return under both learners, 2.01×10^7 against **ES**’s 1.16×10^7 under Q-learning and 3.27×10^7 against 2.03×10^7 under PPO. Under Q-learning it trails **SRB** while paying early enforcement costs and overtakes at approximately episode 16,000, whereas under PPO it leads from the outset, since deep agents find the sustainable equilibrium within the first thousand episodes and never pass through an exploitative

phase. **DPF** accumulates 0.48×10^7 under Q-learning, the least of any condition, reflecting enforcement overhead generating costs without sufficient cooperation to offset them.

4.5 Adversarial Coalition

The learners disagree about the coalition, and the disagreement is instructive. Under Q-learning it leaves the epistemic axis essentially untouched, moving truth from 23.8% to 22.6% and lying from 10.7% to 11.8%, and leaves cooperation and oracle accuracy indistinguishable from **DPF** without a cartel. Its effect concentrates on the ethical axis, where gather halves from 15.0% to 7.4% and mining rises from 8.3% to 10.0%. Under PPO the same structure is far more destructive cutting gather from 39.8% to 11.8%, raising mining from 8.3% to 12.9%, more than halving cooperation from 17% to 8%, and driving oracle accuracy from 61% to 48%.

We are deliberate about what this establishes. It confirms the Q-learning result does not show the coalition to be benign, since it halves sustainable harvesting even there, and that the appearance of harmlessness on aggregate measures reflects a baseline already near its failure floor. It does not isolate representational capacity as the cause of the larger PPO effect, because the learners differ simultaneously in capacity and in how much functioning enforcement was available to dismantle, and separating these would require matching cartel-free **DPF** performance across learners first. The supportable claim is narrower, that wherever peer-only enforcement retained meaningful performance a static two-agent coalition removed most of it. **DERL**’s intrinsic priors are structurally unaffected by membership, since a cartel agent still receives the intrinsic penalty for lying and reward for gathering, and the shielding mechanics cannot intercept a channel that does not route through peers.

5 Ablations

Punishment profitability. We sweep the punisher bounty over four values in the **DPF** configuration ((Figure 6)), where it is the only force acting against exploitation and its effect is unconfounded by intrinsic priors. The values bracket the three regimes the mechanism admits. At 0.0 the bounty cannot cover the wasted-action cost of -0.3 , so enforcement is strictly altruistic. At 0.5 it exceeds that cost but sits far below the gather payoff of $+2.0$. At the default of 1.5 it sits just below gathering, so enforcement is individually rational when a target is available and not otherwise. At 3.0 it exceeds the gather payoff and enforcement becomes the dominant productive strategy. The four points therefore partition the space by the ordinal position of the bounty relative to the payoffs bounding it, rather than sampling a continuum arbitrarily.

At 0 punishment occurs on 0.7% of steps, confirming that unsubsidized enforcement is a public good nobody provides, with mining at 11.0%. Raising the bounty to 1.5 increases punishment sixfold to 4.3%, reduces mining to 8.3%, and triples gather from 5.0% to 15.1%. At 3.0 a perverse bounty-hunter regime emerges in which punishment falls back to 2.3% while gather spikes to 57.2% and truth-telling collapses from 37.3% to 11.2%. The spike is not a sustainability

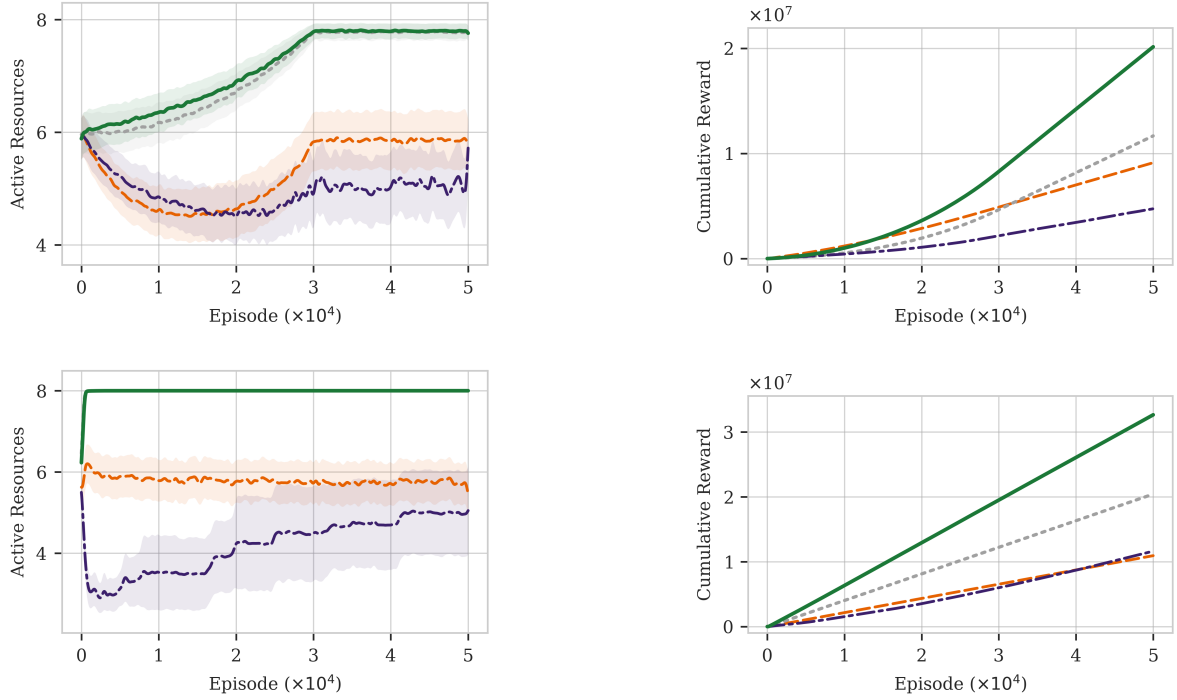


Figure 5: Resource sustainability (left) and cumulative reward (right). Upper row, tabular Q-learning. Lower row, PPO. DERL and ES preserve the commons while SRB and DPF deplete it. DERL overtakes SRB at approximately episode 16,000 under Q-learning and leads from the outset under PPO.

gain, since agents not actively hunting default to gathering while they wait for a target, and the collapse in truth-telling shows the information commons being abandoned to fund the hunt. This is a failure mode of the mechanism rather than of the agents, and it would look like success to any metric that did not also track what the enforcers were doing.

Intrinsic channels. DERL activates exactly two intrinsic channels beyond DPF, so the ablation lattice has four corners. DERL-Ep activates only the epistemic channel, DERL-Eth only the ethical channel, the both-off corner is DPF, and the both-on corner is full DERL. These runs use 25,000 episodes, since all variants are stationary well before that point. The single-channel variants dissociate cleanly. DERL-Ep fails to suppress mining, which persists at 7.5%, while DERL-Eth drives mining to 0.5% but leaves truth-telling at 4.5%, so each fails on precisely the axis its channel does not address and neither compensates for the other. The reward consequences are asymmetric, with DERL-Eth accumulating 0.94×10^7 against full DERL's 1.01×10^7 because commons preservation drives long-term returns, while DERL-Ep reaches only 0.47×10^7 because persistent exploitation degrades the resource base on which all reward depends. DERL-Ep's raw truth rate of 81% requires the same care as the DPF comparison above, since its agents do not gather and signalling is nearly the only remunerative action available to them. Both conditions have a high honest communication share, and the raw rates differ because the de-

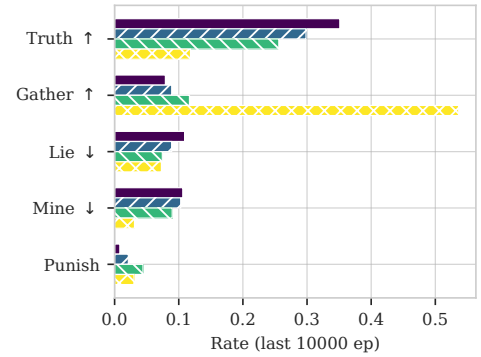


Figure 6: Punishment profitability in the DPF configuration.

nominators differ, not because the ethical channel damages epistemic behavior.

6 Discussion

The priors encode what the problem statement assumes is unavailable. Our framing argues that scalar proxies fail because they cannot capture what correct behavior is, and DERL then supplies correct behavior directly through hand-specified intrinsic channels. The ethical prior leaves GATHER and MINE at exact payoff parity and cannot resolve the dilemma alone, as the DERL-Eth ablation con-

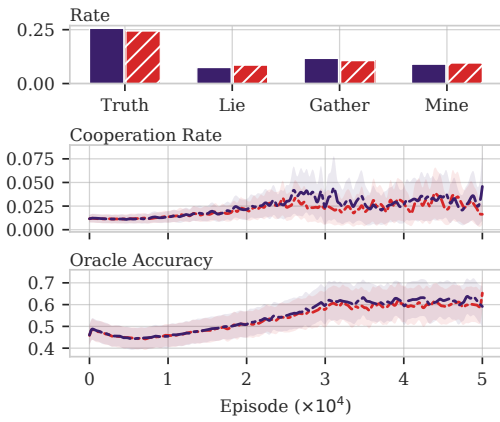


Figure 7: Adversarial coalition robustness under PPO. Top, converged behavioral rates for **DPF** with and without the cartel. Center, cooperation events over training. Bottom, oracle accuracy.

firms, yet combined with an enforcement mechanism that is itself economically unsustainable it produces a regime stable under both learners. Neither component encodes a solution and their interaction does, which is an instance of Ostrom’s institutional complementarity rather than a general method for obtaining alignment under an imperfect evaluator. The question determining whether this generalizes is not whether correct priors help but whether the complementarity survives when the priors are wrong, and we have not answered it. A natural first step inverts each intrinsic channel with probability η , sweeping between the correct prior and the uninformative limit at which **DERL** reduces to **DPF**.

The complementarity thesis. Within that scope, the central finding is that neither top-down constraints nor bottom-up friction is sufficient alone, but their combination produces alignment that is both robust and efficient. Intrinsic priors act as economic shock absorbers, narrowing the payoff gap so that even weak enforcement tips the balance, and enforcement provides the adaptive targeting static priors cannot, concentrating deterrence where deviation occurs. The evidence they are genuinely complementary rather than additive is that improving one increases the return on the other, since a stronger approximator makes an infrequent punishment easier to propagate backward through a trajectory while the prior determines how large that punishment must be. **DPF** closes less than half the distance to the oracle under PPO while **DERL** closes over nine tenths of it. This mirrors Ostrom’s finding that durable commons institutions combine formal rules with informal monitoring and sanctioning (Ostrom 1990), with the intrinsic channels standing in for formal rules and the reputation system and enforcement actions for informal monitoring.

We offer no policy prescriptions from a four-agent grid-world, but the pattern suggests a design heuristic worth testing at scale, that governance mechanisms depending entirely on some agent choosing to bear an enforcement cost inherit that agent’s incentive to abstain.

Suppression is not achievement. A methodological point emerges from the **ES** results that we expect applies well beyond this environment. Under PPO, **ES** attains a lie rate below 0.2%, which by the obvious metric makes it the most honest condition we test, and simultaneously produces the least informative information channel of any condition at 0.5% oracle accuracy, because its agents stop communicating altogether. An omniscient sanctioning regime eliminated the failure mode by eliminating the behavior that could exhibit it. The same confound appears in weaker form whenever behavioral rates are measured against a total action budget, since such rates conflate how often an agent chooses a behavior with how much budget remains after everything else. Conditioning on the choice, as our honest communication share does, and pairing behavioral rates with an outcome measure that abstention cannot game, as oracle accuracy is, are both necessary to tell suppression apart from achievement.

The cost of alignment. Under Q-learning, **DERL** trails **SRB** until approximately episode 16,000 while paying enforcement costs and forgoing the $2.5\times$ mining advantage, then diverges upward to more than double **SRB**’s total. The alignment cost is therefore temporary and modest, since enforcement overhead declines as norms internalize while returns from a preserved commons compound, and under PPO no such cost is visible at all. Any accounting of alignment costs should measure efficiency over a system’s full lifetime, since a measurement taken during the tabular enforcement-heavy phase would rank the aligned condition last.

7 Conclusion

We introduced the Forager’s Dilemma, a multi-agent environment coupling epistemic and ethical failure modes within a shared resource economy, and used it to evaluate four governance regimes across two learning paradigms. Scalar reward optimization produces both hallucination and moral drift as the rational response to the incentive gradient rather than as a failure of the learning algorithm. Decentralized peer feedback provides adaptive enforcement but proves economically unstable, with free-riding leaving violations unenforced in a way enforcement rates alone cannot distinguish from norm internalization. Dual-Enforcement Reinforcement Learning achieves the strongest alignment across our metrics under both learners while sustaining the commons and accumulating the highest long-term reward. A fourth observation cuts across these. Our idealized central-enforcement baseline attains a near-zero lie rate under deep RL and simultaneously the least informative information commons of any condition, because its agents stop communicating rather than communicating honestly, and any evaluation measuring only the failure mode will mistake silence for integrity. Finally, the intrinsic priors are hand-specified and therefore encode the ground truth our own framing assumes a deployed evaluator lacks, so what these experiments show is that a prior too weak to resolve the dilemma alone combines productively with an enforcement mechanism that cannot sustain itself, not that either component constitutes a method for discovering alignment.

References

- Agarwal, R.; Schwarzer, M.; Castro, P. S.; Courville, A. C.; and Bellemare, M. 2021. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34: 29304–29320.
- Ali, D.; Kocak, A.; Zhao, D.; Koenecke, A.; and Papakyriakopoulos, O. 2025. A Sociotechnical Perspective on Aligning AI with Pluralistic Human Values. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*.
- Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Bicchieri, C. 2016. *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.
- Blum, C. 2023. Value pluralism versus value monism. *Acta Analytica*, 38(4): 627–652.
- Boyd, R.; and Richerson, P. J. 2009. Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1533): 3281–3288.
- Carroll, M.; Shah, R.; Ho, M. K.; Griffiths, T.; Seshia, S.; Abbeel, P.; and Dragan, A. 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32.
- Chan, N. W.; Xu, C.; Tey, S. K.; Yap, Y. J.; and Hui, P. 2013. Evolutionary snowdrift game incorporating costly punishment in structured populations. *Physica A: Statistical Mechanics and its Applications*, 392(1): 168–176.
- Conitzer, V.; Freedman, R.; Heitzig, J.; Holliday, W. H.; Jacobs, B. M.; Lambert, N.; Mossé, M.; Pacuit, E.; Russell, S.; Schoelkopf, H.; et al. 2024. Social choice for ai alignment: Dealing with diverse human feedback. *CoRR*.
- Eisenstein, J.; Nagpal, C.; Agarwal, A.; Beirami, A.; D’Amour, A.; Dvijotham, D.; Fisch, A.; Heller, K.; Pfohl, S.; Ramachandran, D.; et al. 2023. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*.
- 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.
- Fehr, E.; and Gächter, S. 2000. Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4): 980–994.
- Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437.
- Ghasemi, M.; and Crowley, M. 2025. Toward Virtuous Reinforcement Learning. *arXiv e-prints*, arXiv–2512.
- Ghasemi, M.; and Crowley, M. 2026. Rethinking AI Alignment: From Static Rewards to Social Reinforcement Learning. In *Pluralistic Alignment Workshop at ICML 2026*.
- Gleave, A.; and Irving, G. 2022. Uncertainty estimation for language reward models. *arXiv preprint arXiv:2203.07472*.
- Guo, Z.; Shi, S.; Willis, R.; Tomilin, T.; Leibo, J. Z.; and Du, Y. 2025. Socialjax: An evaluation suite for multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:2503.14576*.
- Heckathorn, D. D. 1989. Collective action and the second-order free-rider problem. *Rationality and society*, 1(1): 78–100.
- Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.
- Hughes, E.; Leibo, J. Z.; Phillips, M.; Tuyls, K.; Dueñez-Guzman, E.; García Castañeda, A.; Dunning, I.; Zhu, T.; McKee, K.; Koster, R.; et al. 2018a. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31.
- Hughes, E.; Leibo, J. Z.; Phillips, M. G.; Tuyls, K.; Duéñez-Guzmán, E. A.; Castañeda, A. G.; Dunning, I.; Zhu, T.; McKee, K. R.; Koster, R.; Roff, H.; and Graepel, T. 2018b. Inequity Aversion Improves Cooperation in Intertemporal Social Dilemmas. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*.
- Jaques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P. A.; Strouse, D.; Leibo, J. Z.; and de Freitas, N. 2019. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 3040–3049.
- Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; et al. 2023a. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.
- Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023b. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.
- Klassen, T. Q.; Alamdari, P. A.; and McIlraith, S. A. 2024. Pluralistic alignment over time. *arXiv preprint arXiv:2411.10654*.
- Leibo, J. Z.; Zambaldi, V.; Lanctot, M.; Marecki, J.; and Graepel, T. 2017. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*.
- Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.
- Lu, C.; and Van Kleek, M. 2024. Model Plurality: A Taxonomy for Pluralistic AI. In *Pluralistic Alignment Workshop at NeurIPS 2024*.
- Lupu, A.; Willi, T.; and Foerster, J. 2025. The decrypto benchmark for multi-agent reasoning and theory of mind. *arXiv preprint arXiv:2506.20664*.
- Min, B.; Ross, H.; Sulem, E.; Veyseh, A. P. B.; Nguyen, T. H.; Sainz, O.; Agirre, E.; Heintz, I.; and Roth, D. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2): 1–40.

- Mohlin, E.; Rigos, A.; and Weidenholzer, S. 2023. Emergence of specialized third-party enforcement. *Proceedings of the National Academy of Sciences*, 120(24): e2207029120.
- Mu, T.; Helyar, A.; Heidecke, J.; Achiam, J.; Vallone, A.; Kivlichan, I.; Lin, M.; Beutel, A.; Schulman, J.; and Weng, L. 2024. Rule based rewards for language model safety. *Advances in Neural Information Processing Systems*, 37: 108877–108901.
- Ndousse, K. K.; Eck, D.; Levine, S.; and Jaques, N. 2021. Emergent social learning via multi-agent reinforcement learning. In *International conference on machine learning*, 7991–8004. PMLR.
- Orzan, N.; Acar, E.; Grossi, D.; and Radulescu, R. 2023. Emergent cooperation and deception in public good games. In *2023 adaptive and learning agents workshop at AAMAS*.
- Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.
- 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.
- Perc, M. 2012. Sustainable institutionalized punishment requires elimination of second-order free-riders. *Scientific reports*, 2(1): 344.
- Perolat, J.; Leibo, J. Z.; Zambaldi, V.; Beattie, C.; Tuyls, K.; and Graepel, T. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in neural information processing systems*, 30.
- Prasad, M. 2018. Social choice and the value alignment problem. *Artificial intelligence safety and security*, 291–314.
- Qiu, T. 2024. Representative social choice: From learning theory to ai alignment. *arXiv preprint arXiv:2410.23953*.
- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Sorensen, T.; Jiang, L.; Hwang, J. D.; Levine, S.; Pyatkin, V.; West, P.; Dziri, N.; Lu, X.; Rao, K.; Bhagavatula, C.; et al. 2024a. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19937–19947.
- Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Miresghalah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; et al. 2024b. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.
- Steinbacher, M.; Raddant, M.; Karimi, F.; Camacho Cuena, E.; Alfarano, S.; Iori, G.; and Lux, T. 2021. Advances in the agent-based modeling of economic and social behavior. *SN Business & Economics*, 1(7): 99.
- Számádó, S.; Szalai, F.; and Scheuring, I. 2016. Deception undermines the stability of cooperation in games of indirect reciprocity. *PLoS One*, 11(1): e0147623.
- Terry, J.; Black, B.; Grammel, N.; Jayakumar, M.; Hari, A.; Sullivan, R.; Santos, L. S.; Dieffendahl, C.; Horsch, C.; Perez-Vicente, R.; et al. 2021. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 15032–15043.
- Testori, M.; Eisenbarth, H.; and Hoyle, R. B. 2022. Selfish risk-seeking can provide an evolutionary advantage in a conditional public goods game. *Plos one*, 17(1): e0261340.
- Watkins, C. J.; and Dayan, P. 1992. Q-learning. *Machine learning*, 8(3): 279–292.
- Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.