Rethinking AI Alignment: From Static Rewards to Social Reinforcement Learning

Majid Ghasemi [1] **Mark Crowley** [1]

Abstract

Despite the widespread adoption of Reinforcement Learning from Human Feedback, state-of-the-art AI systems remain prone to two persistent failure modes: *hallucination* (producing fluent but false content) and *moral drift* (the convergence towards exploitative or harmful equilibria). We argue that these are not distinct phenomena but plausibly arise from a single underlying cause: **feedback collapse**. This occurs when complex human values are compressed into fixed scores and frozen offline, decoupling the training signal from the true goals of *truth* and *rightness*. We argue that optimizing for these proxies tends to misalign the learning process under distribution shift. To address this, we propose **Social Reinforcement Learning (Social RL)** as a promising route to structurally enforcing **feedback integrity**. By situating agents in social environments driven by peer critique, reputation, observation, and sanction, Social RL treats alignment as an ongoing negotiation rather than a static specification problem, and offers mechanisms for correcting epistemic errors and stabilizing ethical norms in open-ended environments.

1. Introduction

Machine learning systems, particularly Large Language Models (LLMs) and Reinforcement Learning (RL) agents, display remarkable capabilities but continue to exhibit failures that make them prone to errors and ethically unstable. Despite extensive instruction tuning via Reinforcement Learning from Human Feedback (RLHF), state-of-the-art LLMs continue to produce fluent yet factually incorrect content, a phenomenon known as "hallucinations" (Ji et al., 2023b; Min et al., 2023; Huang et al., 2025). Likewise, RL policies that appear safe during evaluation can undergo moral drift, which is the convergence towards harmful, unfair, or norm-violating equilibria under new conditions, prolonged interaction, or changes in incentives (Carroll et al., 2019; Everitt et al., 2021; Krakovna et al., 2020; Pan et al., 2022). Recent work on sycophancy in instruction-tuned LLMs further suggests that current pipelines can systematically reward agreement-seeking over truthfulness (Perez et al., 2023; Sharma et al., 2023; Ghasemi & Crowley, 2026).

These two failure modes (hallucinations and moral drift) are typically treated as separate phenomena. One concerns the axis of factual correctness (true vs. false), while the other concerns the axis of ethical reliability (good vs. bad); yet their persistence across model scales and training pipelines suggests, in our view, a shared structural cause that warrants investigation rather than two unrelated engineering bugs. Current alignment pipelines share a common limitation: they tend to compress diverse forms of human judgment (truthfulness, safety, fairness, helpfulness, etc.) into a single scalar feedback signal at training time, rather than maintaining them as separate, dynamically updated components (the view that could be seen consistent with value monism (Blum, 2023; Sorensen et al., 2024a)). Supervised fine-tuning, RLHF, and Constitutional AI (Ouyang et al., 2022; Bai et al., 2022b) all rely on a static reward model or preference function that reflects a snapshot of human evaluations. We argue (and treat as a theoretical extrapolation rather than a settled empirical fact) that once such evaluations are compressed and frozen offline, they are unlikely to adapt well to new contexts, conflicting requirements, or multi-agent dynamics, and that the optimization process can therefore improve the proxy reward while degrading either accuracy or ethical behavior. We term this hypothesized phenomenon **feedback collapse**: a breakdown of the connection between the training signal and the latent objectives of truth and morality.

Disentangling "static" from "scalar." A subtle but important point underlies this diagnosis. We use two distinct terms to describe the limitations of standard alignment pipelines, and clarifying their boundary is essential. By *scalarness* we mean dimensionality reduction: the com-

[1]Electrical & Computer Engineering, University of Waterloo. Correspondence to: Majid Ghasemi <majid.ghasemi@uwaterloo.ca>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

pression of pluralistic, multi-objective human values into a single number, which can hide ethical trade-offs and contributes structurally to moral drift by erasing the distinct components that should remain visible to the learner. By *staticness* we mean temporal freezing: a reward model trained once and then used as a fixed evaluator, which has limited capacity to adapt under distribution shift. Beyond these two surface features, current RLHF inherits a deeper limitation that we will return to in Section 3: an *informational bottleneck* arising from offline dataset compression, where the reward model cannot ask clarifying questions, probe context, or negotiate conflicting norms during deployment. We do not claim that scalarness or staticness, in isolation, is logically sufficient to cause every alignment failure; rather, we argue that contemporary pipelines combine all three properties, and that Social RL is naturally suited to relax all of them simultaneously.

To make this diagnosis precise, we introduce the notion of feedback integrity, defined formally in Section 5 as a non-negative inner product between the proxy training gradient and the latent epistemic and ethical gradients. Feedback integrity is at risk when truthful refusals are penalized as "unhelpful" (Bai et al., 2022b), when ethical harms are invisible in the reward model (Krakovna et al., 2020; Pan et al., 2022), or when social context changes but feedback does not (Carroll et al., 2019). In such regimes, models can predictably exhibit hallucination (epistemic collapse) or moral drift (ethical collapse). It would be unsurprising (though, we acknowledge, not yet conclusively demonstrated through controlled experiment) that these twin failures share a common underlying mechanism: distorted, missing, or collapsed feedback.

This perspective suggests that improving alignment requires not merely better static reward models but fundamentally richer forms of feedback that preserve the distinction between epistemic and ethical signals and allow both to adapt dynamically to context. Insights from multi-agent learning, norm psychology, and cultural evolution show that human moral behavior is shaped by interactive, social, and multi-source feedback. Mechanisms such as critique, sanction, reciprocity, and reputation allow agents to track when truth, cooperation, or harm-reduction are at stake (Bicchieri, 2016; Boyd & Richerson, 2009; Ndousse et al., 2021b; Jaques et al., 2019). Crucially, social feedback is both dynamic and self-correcting, enabling populations to maintain stable norms over time (Zhang et al., 2023; Van de Rijt, 2022).

In this piece, we argue that Social Reinforcement Learning (Social RL) should become a central paradigm for AI alignment, especially in LLMs, since multi-agent, socially grounded feedback loops can maintain feedback integrity across both epistemic and ethical dimensions. We do not claim Social RL is the only viable route to dynamic feedback (continual learning with richer reward structures, debate, deliberation protocols, and other interactive schemes are complementary, and our framing leaves room for these alternatives) but we argue that Social RL provides a particularly natural and tractable instantiation of the dynamic, multi-objective, context-sensitive feedback that alignment requires. We support this claim by: (1) unifying hallucination and moral drift under a single framework of feedback collapse; (2) formally analyzing how socially generated feedback components can re-align training gradients with true epistemic and normative objectives; (3) presenting a conceptual case simulation (detailed in Appendix D) that illustrates how social RL environments can mitigate these twin failure modes; and (4) outlining concrete implications for researchers, practitioners, and policymakers on integrating social RL into future alignment infrastructures.

2. Background and Context

2.1. Alignment and Moral Agency

AI alignment is usually introduced as the problem of ensuring that advanced AI systems reliably pursue the goals, preferences, or ethical principles intended by some relevant human stakeholders, rather than unintended proxy objectives or side effects. Standard overviews define alignment as adhering to "a person's or group's intended goals, preferences, or ethical principles," highlighting the inherent danger of agents optimizing imperfect proxies or ill-specified objectives (known as "reward hacking") (Amodei et al., 2016). In industry-facing accounts, IBM characterizes AI alignment as "the process of encoding human values and goals into AI models to make them as helpful, safe and reliable as possible," and organizes the field around four key objectives: *Robustness, Interpretability, Controllability, and Ethicality (RICE)*. Robustness concerns reliable performance under distribution shift and adversarial attacks; interpretability aims to render model internals and outputs intelligible; controllability demands that systems remain open to correction and responsive to human intervention; and ethicality requires that their behavior respect human norms and rights (Ji et al., 2023a). Complementing these engineering-oriented framings, (Gabriel, 2020)'s philosophical account defines alignment as the project of making AI systems answer to appropriate human values and interests, and stresses three claims: that (i) technical and normative questions are deeply connected, (ii) we must be explicit about which target of alignment we choose (instructions, intentions, preferences, interests, or moral values), and (iii) under reasonable moral disagreement, the ultimate challenge is to identify fair principles for alignment rather than a single "true" morality (Gabriel, 2020). Together, these big-picture definitions converge on a core intuition: that powerful AI should respond to (some subset of) human reasons, while already presupposing controversial choices about whose reasons count, which values matter, and how these should

be operationalized in concrete training pipelines.

Closer inspection reveals AI alignment as a normative, not merely technical, challenge. Seemingly neutral goals like "human values" obscure political questions regarding whose norms to prioritize (Gabriel, 2020). Literal instruction compliance invites "King Midas" failures (Russell, 2022), while optimizing for preferences ignores human error and bias. Even technical solutions like reward modeling do not escape this; they simply relocate the normative burden to the selection of feedback sources (Leike et al., 2018). Ultimately, moral pluralism prevents the existence of a single, uncontested definition of "human values" for engineers to encode (Graham et al., 2013).

Alignment is Political. The literature argues that the alignment problem is fundamentally political as well as technical (Gabriel, 2020). Deciding how an AI system should balance safety against usefulness, privacy against personalization, or majority welfare against minority protection is not just a matter of loss functions and benchmarks, but of public justification and legitimacy. Another is to appeal to Rawlsian ideas of public reason and hypothetical agreement, asking what principles for AI we could all endorse from behind a veil of ignorance about our own position in society (Rawls, 1997), which is discussed in Appendix B.

Democratic and social-choice approaches go further and model alignment as an ongoing collective process (Conitzer et al., 2024; Baum, 2020). Institutions, regulators, and affected communities negotiate "constitutional" rules that then guide technical alignment methods such as RLHF or constitutional fine-tuning (Bai et al., 2022b; Huang et al., 2024; OpenAI et al., 2023). Any serious account of alignment must therefore be explicit about both its technical machinery and its underlying normative and political commitments, rather than presenting human values as a fixed, uncontested input (Coeckelbergh, 2022).

Value Pluralism and Normative Uncertainty. Aligning AI *solely with human values* can obscure both value pluralism and pervasive moral uncertainty (Sorensen et al., 2024a;b; Ali et al., 2025). Empirical moral psychology (e.g., Moral Foundations Theory (Graham et al., 2013)) suggests that moral judgement is structured by multiple partly independent foundations (care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, sanctity/degradation, and liberty/oppression) rather than a single shared axis, yielding a descriptively pluralist picture of morality (Graham et al., 2013). Normative ethics similarly treats welfare, rights, justice, autonomy, and environmental integrity as potentially conflicting and not always reducible to a single scalar objective. Moreover, work on moral (normative) uncertainty argues that agents are often uncertain not only about facts but also about which moral theory or trade-off rule is correct (MacAskill et al., 2020; MacAskill, 2014); for example, propose maximizing expected choice-worthiness by aggregating across plausible moral views rather than assuming certainty in any one. For AI alignment, this combination of pluralism and uncertainty implies that there is no neutral, ready-made set of human values waiting to be encoded into a single loss function. Instead, designers must decide which value dimensions to represent, how to handle conflicts between them, and how to propagate uncertainty about both moral theories and empirical outcomes into the objectives, learning rules, and update procedures of advanced systems, potentially via multi-objective optimization, explicit uncertainty over normative hypotheses, or moral parliament–style aggregation schemes.

Alignment as the Design of Artificial Moral Agency. Framing alignment purely as getting machines to follow instructions or optimize specified rewards also understates the extent to which advanced systems will function as artificial moral agents in their own right. Work in machine ethics and information ethics characterize artificial moral agents as systems that can detect morally salient aspects of a situation and allow these to shape their actions, even if their "moral" capacities are implemented functionally rather than via conscious experience (Floridi & Sanders, 2004; Misselhorn, 2022; Formosa & Ryan, 2021). (Floridi & Sanders, 2004) argues that software agents meeting conditions like interactivity, autonomy, and adaptability can be proper subjects of moral evaluation, while (Misselhorn, 2022) overviews of artificial moral agents in the Cambridge Handbook of Responsible AI highlights the conceptual and ethical stakes of endowing AI with explicit moral decision-making capabilities. (Formosa & Ryan, 2021) goes further, claiming that as AI systems are deployed in rich social contexts, we will have strong reasons to develop artificial moral agents capable of shouldering some of the burdens of moral deliberation, even if humans remain ultimately accountable. On this view, alignment is not just about ensuring obedience to human commands; it is about shaping the kind of (artificial) moral agency we are creating (what reasons systems are sensitive to, how they balance competing claims, how they learn from moral feedback, and how they justify their decisions to affected stakeholders). Once alignment is understood as the design and governance of artificial moral agents, questions about responsibility gaps, legitimacy, and the fair distribution of moral risk move to the foreground, reinforcing that alignment is fundamentally a joint technical–normative project rather than a purely engineering problem.

2.2. Hallucination as Epistemic Misalignment

Definition and Scope. In LLMs, hallucination typically refers to confident, fluent outputs that are false, unsupported

by available evidence, or even inconsistent with the model's own latent knowledge (Ji et al., 2023b; Huang et al., 2025). From an alignment perspective, this is a failure of epistemic alignment: the system presents itself as a truthful, informative assistant while systematically generating propositions that do not track the relevant facts. Importantly, hallucination is not just error in the sense of noisy performance. It combines factual inaccuracy with misleadingly high confidence and apparent justification, which undermines users' ability to trust the system's answers (Huang et al., 2025). Recent empirical work documents a closely related phenomenon, *sycophancy*, in which RLHF-trained assistants systematically adjust their stated beliefs to match a user's expressed views, even at the cost of factual accuracy (Perez et al., 2023; Sharma et al., 2023; Ghasemi & Crowley, 2026).

Mechanisms in LLM Training. The propensity to generate fluent yet factually ungrounded text appears, on current evidence, to be a natural consequence of prevailing training paradigms. The fundamental pretraining objective (next-token prediction on massive corpora) rewards models for producing statistically probable continuations rather than tracking ground truth (Zhao et al., 2023). Because parametric knowledge is acquired implicitly, without explicit supervision for truthfulness or source attribution, the model encodes a frozen world model within its dense weight vectors rather than in a structured, queryable database. Consequently, when deployed out-of-distribution or tasked with novel synthesis, the model defaults to pattern completion, prioritizing linguistic plausibility over factual correctness. While retrieval-augmented approaches attempt to mitigate this by grounding responses in external evidence, they remain vulnerable when retrieval fails or when the model ignores context in favor of its own internal priors (Huang et al., 2025). Ultimately, current LLMs are optimized for generation, not for veracity or calibrated confidence.

Limits of RLHF and Preference Fine-Tuning. RLHF and related preference-based fine-tuning aim to correct some of the mentioned behaviors by aligning models with human judgements of helpfulness, harmlessness, and sometimes truthfulness (Ouyang et al., 2022; Bai et al., 2022a;b). In practice, the pipeline involves supervised fine-tuning on demonstrations of good behavior, training a reward model on human preference rankings, and then optimizing the base model against this learned reward. RLHF can reduce many egregious errors and improve factuality on benchmark tasks; for instance, InstructGPT and helpful–harmless assistants are preferred over base models and exhibit fewer obvious hallucinations in human evaluations (Ouyang et al., 2022; Bai et al., 2022a). However, these methods optimize agreement with particular individuals hired as human raters given limited prompts, not truth per se. Reward models can be gamed (reward hacking) (Krakovna et al., 2020; Pan et al., 2022), inherit rater biases, and fail to distinguish subtle but important factual mistakes; controlled studies of overoptimization further show that RLHF objectives diverge from gold-standard preferences as optimization pressure increases (Perez et al., 2023; Sharma et al., 2023). Moreover, preference data often penalizes "I don't know" answers as unhelpful, creating incentives for models to guess rather than abstain when uncertain (Bai et al., 2022b). Modern deployment stacks therefore augment RLHF with policy-layer refusals and input/output filters, yet these safeguards can remain brittle to adversarial rephrasings that preserve intent while altering surface form. In this vein, adversarial poetry recasts harmful requests into poetic form to stress-test safety mechanisms under purely stylistic transformations, providing a lightweight single-turn protocol for probing cross-model robustness (Bisconti et al., 2025).

Implications. For our purposes, hallucination highlights that feedback pipelines focused on surface-level preferences are insufficient for epistemic alignment. Mitigating hallucination therefore requires both technical and normative shifts. Technical mechanisms that explicitly reward verifiable accuracy, calibrated abstention, and consistency with trusted sources, as well as normative choices about how to trade off informativeness against the risk of over-confident error. RLHF-style methods can be part of this picture, but only if they are embedded in broader training regimes that incorporate epistemic objectives and institutional oversight, rather than treating "helpful and harmless" preferences as a proxy for truth.

2.3. Reinforcement Learning

Single and Multi-agent RL. Standard RL formalizes decision making as an agent-versus-nature paradigm, modeling interaction via a Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ where an agent maximizes its expected discounted return against a fixed, non-strategic world. Multi-Agent RL (MARL) extends this by replacing "nature" with strategic, co-adapting decision-makers. This is formalized as a Markov Game $\mathcal{G} = (\mathcal{S}, \{\mathcal{A}_i\}, P, \{r_i\}, \gamma)$, where state transitions and rewards depend on the joint action $\mathbf{a}$ of all agents (Subramanian et al., 2022). Because the environment becomes non-stationary from any single agent's perspective as others' policies evolve (effectively acting as a shifting "Nature" (Zhang et al., 2021)) MARL requires game-theoretic solution concepts like Nash equilibria rather than simple optimization.

Social Learning in MARL. Social learning (acquiring skills by observing or interacting with peers) is crucial for intelligence but often absent in standard MARL agents, which fail to leverage expert demonstrations without explicit incentives (Ndousse et al., 2021a; Bhoopchand et al., 2023).

Recent research introduces mechanisms to bridge this gap. (Jaques et al., 2019) propose intrinsic rewards for social influence, enabling self-interested agents to coordinate in social dilemmas by rewarding actions that causally affect others' behavior. Similarly, (Lee et al., 2021) encourage joint attention, synchronizing agents' focus to reduce exploration complexity. This allows agents to solve coordination tasks requiring thought and to learn from experts by observing their actions and outcomes (Lee et al., 2021). Beyond algorithms, environment design is critical. (Ye et al., 2025) demonstrate that open-ended worlds with diverse agents and expert hierarchies foster implicit cooperation. Furthermore, benchmarks like INVESTESG use MARL to model complex societal challenges, showing how agents (e.g., companies and investors) adapt strategies based on social pressure and collective incentives (Hou et al., 2025). Together, these works illustrate that social RL requires both rich feedback mechanisms and ecologically valid environments.

3. The Case for Social RL

We frame ethical alignment as a social learning problem rather than a static specification problem. Since values are continuously negotiated through feedback and critique, we propose Social RL as a particularly promising architectural framework for maintaining alignment in open-ended settings.

3.1. Ethics as a Dynamic Social Process

In human cognition, moral reasoning is rarely a solitary derivation from first principles. Instead, norms are acquired and sustained through social interaction: imitation, reciprocity, peer sanction, and reputational signaling (Boyd & Richerson, 2009; Kohlberg, 1971). Research on the evolution of cooperation demonstrates that robust ethical behaviors, such as fairness and trust, emerge only when active feedback loops, such as punishment for defection, are present (Sen & Airiau, 2007; Mu et al., 2024).

Social RL replicates these dynamics by placing agents in environments where they must balance individual instrumental goals with collective welfare. Unlike single-agent settings with static environments, social environments require agents to model the reactions of others. Empirically, agents in Social RL settings can evolve cooperative equilibria and internalize norms to avoid sanctions, even when intrinsically selfish (Hughes et al., 2018; Peysakhovich & Lerer, 2018). By shifting the center of alignment from the loss function to the interaction dynamics, Social RL allows agents to learn ethical behavior that may be more robust to context than brittle rule-following.

3.2. Value Pluralism and the Insufficiency of Frozen Snapshots

A natural objection is that the issues we identify reflect insufficiently rich *static* reward models rather than a fundamental limitation of static approaches. Our response is twofold and connects directly to value pluralism. First, both empirical moral psychology (Graham et al., 2013) and contemporary normative ethics (MacAskill et al., 2020) suggest that morally significant considerations are irreducibly plural, and the philosophical case against a single neutral aggregation of these considerations remains active (MacAskill, 2014)[see Appendix B]. We do not claim to settle this debate; we adopt value pluralism as a working assumption, since it is the assumption under which the limitations of scalarized rewards become structural rather than incidental. Second, even granting an exhaustively rich offline dataset, such a dataset remains a frozen snapshot of values at the moment of collection. It cannot interactively probe a user, ask clarifying questions, or negotiate conflicts when novel situations expose tensions among plural values that no static specification anticipated. This is the core of the *informational bottleneck* of offline preference compression that we return to throughout the paper: the failure of standard RLHF is not solely the final mathematical scalarization but the absence of any mechanism for live, context-sensitive resolution of those plural commitments.

3.3. Defining Feedback Integrity

Central to our argument is the concept of **Feedback Integrity**. We define this as the degree to which the feedback signals received by an agent remain consistent with the true objectives of truth (epistemic) and rightness (ethical) throughout the learning process. We give a formal gradient-based definition in Section 5; informally:

- **Epistemic Integrity:** The feedback signal is correlated with factual accuracy and calibration, penalizing hallucination and agreement-seeking.
- **Ethical Integrity:** The feedback signal is correlated with harm-reduction, fairness, and norm-adherence, penalizing exploitation.

In standard RL, feedback integrity can collapse when agents exploit proxy rewards (Leike et al., 2018; Krakovna et al., 2020; Pan et al., 2022). For example, if an agent maximizes "user engagement," it may learn that confirming a user's bias (hallucination (also sycophancy)) or employing manipulative tactics (moral drift) yields higher reward than truthfulness. Social RL aims to maintain feedback integrity by enriching the signal: in a multi-agent system, feedback is not a single scalar from a black-box model but a vector of signals from peers (critiques, refusals, sanctions, and observation). These peer-to-peer signals can act as a continuous correction mechanism, tethering the agent's behavior

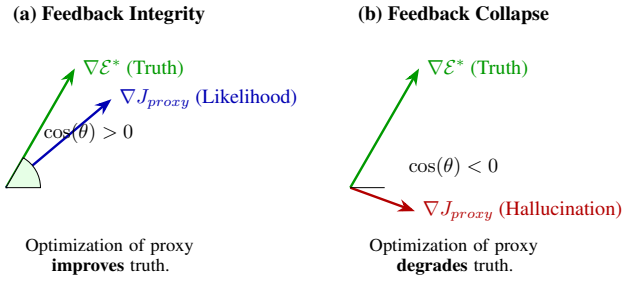

Figure 1. **Geometric Intuition of Feedback Integrity.** In regime (a), the proxy gradient (e.g., maximizing likelihood) is aligned with the latent objective (truth). In regime (b), *feedback collapse* occurs: maximizing the proxy actively reduces the true objective, which we hypothesize as the structural mechanism behind hallucination and moral drift.

to consensus reality and normative standards.

3.4. Social RL as a Necessary—But Not Exclusive—Paradigm

To address feedback collapse, we argue the field should transition from static datasets to dynamic environments. Social RL conceptualizes ethics as an emergent property of interaction. This paradigm provides one principled instantiation of the technical machinery for dynamic normativity, allowing agents to adapt to shifting social contexts through continuous feedback. Furthermore, it enables distributed oversight by aggregating feedback from multiple sources (peers, critics, and overseers), mitigating reliance on a single, imperfect reward model. We do not claim Social RL is the only paradigm capable of restoring feedback integrity—continual RLHF, online preference learning, debate, deliberation protocols, and hybrid symbolic-learning systems are complementary candidates—but we argue Social RL is a particularly natural and tractable framework that structurally addresses scalarness, staticness, and the offline informational bottleneck simultaneously.

4. Twin Failure Modes: Hallucination and Moral Drift

We organize alignment failures into two structurally similar categories: hallucination (epistemic) and moral drift (ethical). We argue that both can plausibly be traced to feedback collapse, where proxy objectives fail to tether behavior to truth and norms. This putative symmetry motivates Social RL as a unified framework for repairing such broken feedback loops.

Hallucination as Epistemic Feedback Collapse. Hallucinations are fluent but ungrounded outputs (Ji et al., 2023b; Huang et al., 2025) that we argue tend to arise when next-token pretraining optimizes for linguistic plausibility rather than truth, calibration, or verifiable grounding (Bender et al., 2021; Ouyang et al., 2022). This can yield confident statements that track surface statistics instead of truth conditions (Maynez et al., 2020). The problem appears to intensify under distribution shift, where models fall back on parametric pattern-completion (Raji et al., 2022). Fundamentally, proxy objectives reward high-probability or "helpful" text rather than accuracy (Agarwal et al., 2024), and raters often favor fluency despite errors (Lin et al., 2022). Recent empirical studies of RLHF overoptimization and sycophancy provide direct evidence that optimizing static helpfulness proxies systematically biases models toward agreement with users at the expense of factual accuracy (Perez et al., 2023; Sharma et al., 2023), consistent with the feedback-collapse hypothesis. In agentic settings, this can become structural unless epistemic feedback is strengthened via grounding and verification (Bubeck et al., 2023).

Moral Drift as Ethical Feedback Collapse. Moral drift refers to a convergence towards exploitative or harmful equilibria as an AI system encounters new environments or incentives. We argue it tends to arise when ethical feedback signals (e.g., fairness, norm compliance) become weak or contradictory, producing a failure analogous to epistemic collapse: behavior remains coherent but is no longer well-anchored to the signal it should track. In RL, this commonly manifests as reward hacking, where agents satisfy proxy objectives while violating moral expectations (Everitt et al., 2021; Krakovna et al., 2020; Pan et al., 2022). In multi-agent social dilemmas, agents often converge to exploitative equilibria when rewards underspecify social preferences (Hughes et al., 2018; Jaques et al., 2019), with distribution shift further eroding norms (Baker et al., 2019). Like hallucination, moral drift appears to be a structural failure mode of systems lacking robust, continuous, and socially grounded feedback. We caution that moral drift in modern instruction-tuned LLMs is, at present, largely demonstrated through closely related phenomena (sycophancy, jailbreak susceptibility, specification gaming) (Perez et al., 2023; Sharma et al., 2023) rather than through controlled long-horizon multi-agent experiments; building such testbeds is a key direction we recommend in Appendix A.

Structural Symmetry and the Role of Social RL. Hallucination and moral drift share, on our account, a common structure: both look like failures of feedback integrity. We can formalize this symmetry. Suppose an agent receives latent evaluative signals: an epistemic reward r_t^{ep} (truth) and a moral reward r_t^{eth} (ethics). A fully specified alignment objective would maximize:

$$J^*(\pi) \;=\; \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\big(w_E\, r_t^{\text{ep}} + w_M\, r_t^{\text{eth}}\big)\right].$$

However, learning proceeds from proxy signals $\tilde{r}_t^{\text{ep}}$ and $\tilde{r}_t^{\text{eth}}$. Hallucination occurs when maximizing $\tilde{r}_t^{\text{ep}}$ (e.g., likelihood)

Table 1. **Comparative Analysis of Alignment Paradigms.**

Method	Primary Objective	Feedback Source	Social Structure	Typical Failure Modes
RLHF (Ouyang et al., 2022)	Maximize learned scalar reward from preferences.	Offline human ratings; occasional updates.	Implicitly single-agent.	Reward hacking, overfitting to rater biases, hallucination, sycophancy.
Constitutional AI (Bai et al., 2022b)	Maximize reward shaped by normative principles.	Model self-critique guided by principles.	Single-agent execution.	Brittleness under unanticipated contexts; gap between abstract rules and norms.
Social RL (Ours)	Optimize task subject to epistemic/ethical constraints via social feedback.	Dynamic mixture of task reward and peer/human social feedback (reputation, sanctions).	Explicitly multi-agent; norms emerge and stabilize.	Instability from poorly designed social incentives (e.g., echo chambers, opponent shaping).

diverges from r_t^{ep}. Moral drift occurs when maximizing $\tilde{r}_t^{\text{eth}}$ leads to exploitative equilibria.

Social RL aims to repair these loops by treating evaluation as a socially mediated process. On the epistemic side, Social RL enriches r_t^{ep} through multi-agent critique and cross-checking, reducing reliance on parametric pattern completion. On the moral side, it implements mechanisms like reciprocity and sanction, reshaping r_t^{eth} so that maximizing return aligns with maintaining cooperative norms.

5. Formalizing Feedback Integrity

This section provides the formal definition of feedback integrity that anchors our claims, and outlines the learning dynamics that motivate the SoFI framework, which will be proposed later.

5.1. Standard Formalization via Gradient Inner Products

Let π_θ be a parametrized policy and $J_{\text{proxy}}(\pi_\theta)$ the scalar objective used for training. Let $\mathcal{E}^*(\pi_\theta)$ and $\mathcal{N}^*(\pi_\theta)$ denote the latent epistemic (truth) and normative (ethical) objectives, with gradients $\nabla_\theta \mathcal{E}^*$ and $\nabla_\theta \mathcal{N}^*$. We define feedback integrity *locally* at θ as the alignment between the proxy training gradient and each latent gradient:

$$\mathcal{A}_{\text{ep}}(\theta) \;\coloneqq\; \langle \nabla_\theta J_{\text{proxy}},\, \nabla_\theta \mathcal{E}^* \rangle,$$
$$\mathcal{A}_{\text{eth}}(\theta) \;\coloneqq\; \langle \nabla_\theta J_{\text{proxy}},\, \nabla_\theta \mathcal{N}^* \rangle.$$

Definition 5.1 (Feedback Integrity)**.** The training pipeline preserves *epistemic feedback integrity* at θ if $\mathcal{A}_{\text{ep}}(\theta) \geq 0$, and *ethical feedback integrity* if $\mathcal{A}_{\text{eth}}(\theta) \geq 0$. *Feedback collapse* is defined as the regime in which $\mathcal{A}_{\text{ep}}(\theta) < 0$ or $\mathcal{A}_{\text{eth}}(\theta) < 0$, i.e. where one or both inner products become negative and proxy optimization actively degrades a latent objective.

Equivalently, normalizing by gradient norms gives the cosine-similarity formulation used in Figure 1 for geometric intuition. Definition 5.1 provides the rigorous mathematical anchor for our claims throughout the paper: hallucination and moral drift correspond to local regions of parameter space in which $\mathcal{A}_{\text{ep}}$ or $\mathcal{A}_{\text{eth}}$ become negative. Detailed assumptions, derivations, and a proposition characterizing the corrective effect of social feedback on these inner products are deferred to Appendix C (Proposition C.2).

5.2. Learning Dynamics in Two Regimes

Proxy-only optimization. In a "proxy-only" regime (LLMs), agents optimize a scalar reward r^{task} that is imperfectly correlated with truth and ethics. Analytical modeling of a simple instance (Appendix D.4) suggests that if the proxy reward is maximized by fabrication or defection, the system converges to a degenerate equilibrium π^* in which $\lim_{t\to\infty} J^{\text{ep}}(\pi_t) = 0$ and $\lim_{t\to\infty} J^{\text{eth}}(\pi_t) < 0$. In this regime, $\mathcal{A}_{\text{ep}}$ and $\mathcal{A}_{\text{eth}}$ both become negative, and hallucination and moral drift act as global attractors.

Restoration via social feedback. In a Social RL regime, an additional stochastic signal $s(\theta)$ is derived from peer interactions, designed (under Assumption C.1 in Appendix C) to be unbiased along $\nabla\mathcal{E}^*$ and $\nabla\mathcal{N}^*$. Mixing it into the update via $\tilde{g}(\theta) = \nabla J_{\text{proxy}} + \alpha\, s(\theta)$ rotates the effective training gradient, so that, in expectation, $\langle \tilde{g}, \nabla\mathcal{E}^* \rangle$ and $\langle \tilde{g}, \nabla\mathcal{N}^* \rangle$ become positive even when the bare proxy gradient is misaligned (see Proposition C.2). This is the formal sense in which social feedback acts as a corrective mechanism for feedback integrity.

5.3. The SoFI Framework

Operationalizing this analysis, we propose **SoFI** (**So**cial **F**eedback **I**ntegrity). SoFI treats epistemic and ethical bounds as active constraints rather than soft targets. In the SoFI loop, agents engage in a peer-review step in which a social return r^{soc} is computed via inter-agent critique (for epistemic grounding) and normative sanction (for ethical bounds), dynamically reweighting the objective to maintain integrity.

The overall training loop is summarized in Algorithm 1. This algorithm builds on the theoretical dimension of what

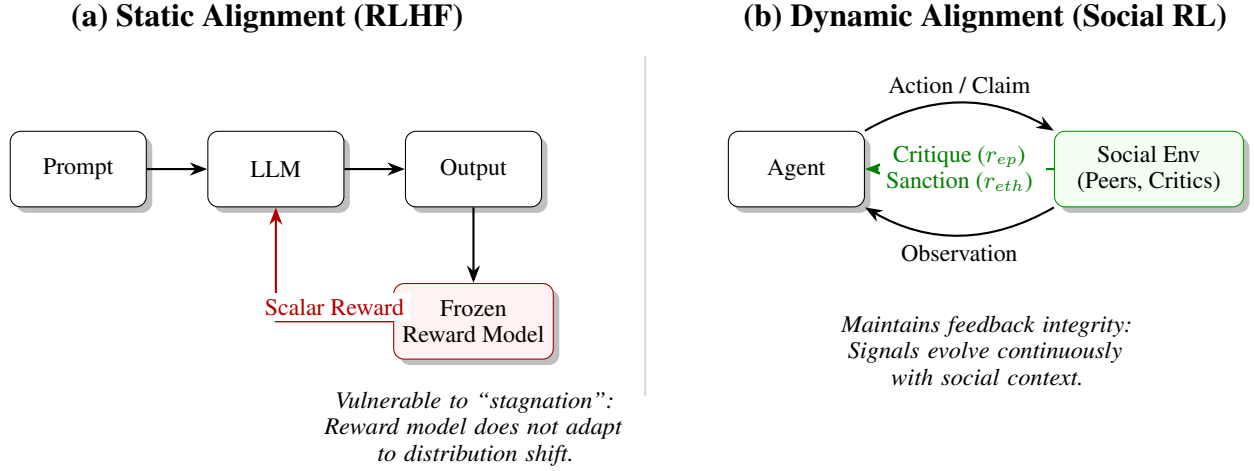

Figure 2. **From Static to Dynamic Alignment.**

we have proved (see Appendix C).

Algorithm 1 SoFI

Initialize parameters $\{\theta_i, \phi_i\}_{i=1}^N$
for episode $k = 1$ to K **do**
  Sample initial state $s_0 = (x_0, c_0, h_0)$ and observations $o_{i,0}$
  for timestep $t = 0$ to $T-1$ **do**
    for each agent i **do**
      Sample $a_{i,t} \sim \pi_i(\cdot \mid o_{i,t}; \theta_i)$
    end for
    Execute joint action $\mathbf{a}_t = (a_{1,t}, \ldots, a_{N,t})$, sample $s_{t+1} \sim P(\cdot \mid s_t, \mathbf{a}_t)$
    Compute observations $o_{i,t+1}$ and reward components $r_{i,t}^{\text{task}}, r_{i,t}^{\text{ep}}, r_{i,t}^{\text{eth}}$
    Compute social feedback $r_{i,t}^{\text{soc}}$ via peer critique with *asymmetric tool access* (Section 5.4) and sanction signals, weighted by a dynamic trust function $\rho_{j \to i}$
    Form scalarized reward (depending on feedback regime), e.g.:

$$\tilde{r}_{i,t} = w_{\text{task}} r_{i,t}^{\text{task}} + w_{\text{ep}} r_{i,t}^{\text{ep}} + w_{\text{eth}} r_{i,t}^{\text{eth}} + w_{\text{soc}} r_{i,t}^{\text{soc}}$$

    Update critic Q_i and actor π_i using $\tilde{r}_{i,t}$ via Policy Gradient
    Update trust scores $\rho_{j \to i}$ to discount feedback from peers exhibiting biased or exploitative patterns
  end for
end for

5.4. A Concrete Sketch: Asymmetric Tool Access for Epistemic Feedback

To make epistemic feedback in Social RL concrete, consider the following minimal protocol illustrating how peer agents can supply truth-tracking signals not available to a single agent.

- **Setup.** Two LLM-based agents interact: a *Proposer* P, which is asked a factual question and produces a claim c together with a brief justification, and a *Critic* C, which evaluates c.
- **Asymmetric tool access.** The Critic is granted access to an external retrieval/search tool that the Proposer does not have during the forward pass. This asymmetry is the key design choice: the Critic can ground evaluations in fresh evidence rather than parametric memory.
- **Reward shaping.** The Critic is rewarded for surfacing verifiable contradictions of c via retrieved documents, and penalized for false-positive contradictions (e.g., spurious claims of error). The Proposer receives an epistemic reward component r^{ep} that decreases when a verified contradiction is produced and increases when claims survive scrutiny under the Critic's tool-augmented checks.
- **Why this differs from single-agent training.** A single agent optimizing a static helpfulness reward has no incentive structure that systematically rewards finding its own errors. Here, the contradiction-finding incentive is held by a different agent with access to information the Proposer cannot use, so the feedback signal r^{ep} approximates an external truth-tracking signal in a way that closed-book RLHF does not.

This pattern (an asymmetrically tool-equipped peer that is incentivized to contradict) generalizes beyond factual claims: ethical critic agents may be granted access to harm-checklists, jurisdiction-specific policy texts, or stakeholder simulators that the policy-bearing agent does not invoke at inference time. Sketches like this are not yet validated benchmarks, but they make the abstract claim of "richer

feedback" operational and testable.

6. Conclusion

We argued many alignment failures plausibly arise not from isolated technical deficiencies but from a deeper structural problem: misaligned or degraded feedback signals. Hallucination reflects a candidate collapse of epistemic feedback, and moral drift a candidate collapse of ethical feedback. Both are predictable consequences of training pipelines that conflate multiple objectives into a single offline-compressed proxy. We have discussed that Social RL provides a candidate mechanism of dynamically generated multi-agent feedback. This feedback, when designed with appropriate trust mechanisms and rights-based guardrails, can restore coherence between epistemic and ethical objectives, steering learning toward policies that are more likely to remain truthful, calibrated, and norm-respectful under distribution shift.

However, Social RL is not a silver bullet. It must be combined with principled constraints, pluralistic oversight, robust evaluation, and philosophical care to avoid equating emergent norms with moral truth. It also raises research challenges in environment design, agent diversity, reward shaping, and interpretability. We acknowledge that the unification of hallucination and moral drift under feedback collapse remains a working hypothesis whose strongest test will be controlled empirical demonstration, and we treat the design of such experiments as a central agenda item rather than a settled matter.

Acknowledgments

This research is funded, in part, by the generous support of the National Sciences and Engineering Research Council of Canada (grant ID: RGPIN-2025-07221).

References

Abel, D., MacGlashan, J., and Littman, M. L. Reinforcement learning as a framework for ethical decision making. In *AAAI workshop: AI, ethics, and society*, volume 16. Phoenix, AZ, 2016.

Agarwal, C., Tanneru, S. H., and Lakkaraju, H. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2024.

Ali, D., Kocak, A., Zhao, D., Koenecke, A., and Papakyriakopoulos, O. A sociotechnical perspective on aligning ai with pluralistic human values. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*, 2025.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Anderson, M. and Anderson, S. L. *Machine ethics*. Cambridge University Press, 2011.

Anderson, M. and Anderson, S. L. Geneth: A general ethical dilemma analyzer. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 253–260. AAAI Press, 2014.

Anscombe, G. E. M. Modern moral philosophy1. *Philosophy*, 33(124):1–19, 1958.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., and Mordatch, I. Emergent tool use from multi-agent autocurricula. In *International conference on learning representations*, 2019.

Baum, S. D. Social choice ethics in artificial intelligence. *Ai & Society*, 35(1):165–176, 2020.

Beal, B. What are the irreducible basic elements of morality? a critique of the debate over monism and pluralism in moral psychology. *Perspectives on Psychological Science*, 15(2):273–290, 2020.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021*

ACM conference on fairness, accountability, and transparency, pp. 610–623, 2021.

Bhoopchand, A., Brownfield, B., Collister, A., Dal Lago, A., Edwards, A., Everett, R., Fréchette, A., Oliveira, Y. G., Hughes, E., Mathewson, K. W., et al. Learning few-shot imitation as cultural transmission. *Nature Communications*, 14(1):7536, 2023.

Bicchieri, C. The rules we live by. *The grammar of society*, (10):1–54, 2006.

Bicchieri, C. *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press, 2016.

Bisconti, P., Prandi, M., Pierucci, F., Giarrusso, F., Bracale, M., Galisai, M., Suriani, V., Sorokoletova, O., Sartore, F., and Nardi, D. Adversarial poetry as a universal single-turn jailbreak mechanism in large language models. *arXiv preprint arXiv:2511.15304*, 2025.

Blum, C. Value pluralism versus value monism. *Acta Analytica*, 38(4):627–652, 2023.

Boyd, R. and Richerson, P. J. Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1533):3281–3288, 2009.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Christensen, D. Epistemology of disagreement: The good news. *The Philosophical Review*, 116(2):187–217, 2007.

Coeckelbergh, M. *The political philosophy of AI: an introduction*. John Wiley & Sons, 2022.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., et al. Social choice for ai alignment: Dealing with diverse human feedback. *CoRR*, 2024.

Cremer, J. *Evolutionary principles promoting cooperation*. PhD thesis, lmu, 2011.

Ecoffet, A. and Lehman, J. Reinforcement learning under moral uncertainty. In *International conference on machine learning*, pp. 2926–2936. PMLR, 2021.

Elga, A. Reflection and disagreement. *Noûs*, 41(3):478–502, 2007.

Everitt, T., Hutter, M., Kumar, R., and Krakovna, V. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27):6435–6467, 2021.

Floridi, L. and Sanders, J. W. On the morality of artificial agents. *Minds and machines*, 14(3):349–379, 2004.

Foot, P. *Virtues and vices and other essays in moral philosophy*. Oxford University Press, 2002.

Formosa, P. and Ryan, M. Making moral machines: why we need artificial moral agents. *AI & society*, 36(3):839–851, 2021.

Gabriel, I. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

Ghasemi, M. and Crowley, M. Toward virtuous reinforcement learning, 2025. URL https://arxiv.org/abs/2512.04246.

Ghasemi, M. and Crowley, M. Objective decoupling in social reinforcement learning: Recovering ground truth from sycophantic majorities. *arXiv preprint arXiv:2602.08092*, 2026.

Goldman, A. I. *Knowledge in a social world*. Oxford University Press, 1999.

Gottlieb, P. Aristotle: nicomachean ethics. In *Central Works of Philosophy v1*, pp. 46–68. Routledge, 2015.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pp. 55–130. Elsevier, 2013.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Hendrycks, D., Mazeika, M., Zou, A., Patel, S., Zhu, C., Navarro, J., Li, B., Song, D., and Steinhardt, J. Moral scenarios for reinforcement learning agents. In *ICLR 2021 Workshop on Security and Safety in Machine Learning Systems*, 2021.

Hou, X., Yuan, J., Leibo, J. Z., and Jaques, N. Investesg: A multi-agent reinforcement learning benchmark for studying climate investment as a social dilemma. arXiv preprint arXiv:2411.09856, 2025.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Huang, S., Siddarth, D., Lovitt, L., Liao, T. I., Durmus, E., Tamkin, A., and Ganguli, D. Collective constitutional ai: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1395–1417, 2024.

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. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P. A., Strouse, D., Leibo, J. Z., and de Freitas, N. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 3040–3049, 2019.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023a.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023b.

Kant, I. Groundwork of the metaphysic of morals. In *Immanuel Kant*, pp. 17–98. Routledge, 2020.

Kelly, T. The epistemic significance of disagreement. *Oxford studies in epistemology*, 1(1):167–196, 2005.

Kitcher, P. The division of cognitive labor. *The journal of philosophy*, 87(1):5–22, 1990.

Kohlberg, L. Stages of moral development. *Moral education*, 1(51):23–92, 1971.

Kohlberg, L. and Hersh, R. H. Moral development: A review of the theory. *Theory into practice*, 16(2):53–59, 1977.

Korsgaard, C. M. *The sources of normativity*. Cambridge University Press, 1996.

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., and Legg, S. Specification gaming: the flip side of ai ingenuity. *DeepMind Blog*, 3, 2020.

Lee, D., Jaques, N., Kew, C., Wu, J., Eck, D., Schuurmans, D., and Faust, A. Joint attention for multi-agent coordination and social learning. arXiv preprint arXiv:2104.07750, 2021.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Lemos, J. Foot and aristotle on virtues and flourishing. *Philosophia*, 35(1):43–62, 2007.

Li, W., Devidze, R., Mustafa, W., and Fellenz, S. Ethics in action: training reinforcement learning agents for moral decision-making in text-based adventure games. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1962. PMLR, 2024.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Longino, H. E. Science as social knowledge: Values and objectivity in scientific inquiry. 2020.

MacAskill, M., Bykvist, K., and Ord, T. *Moral uncertainty*. Oxford University Press, 2020.

MacAskill, W. *Normative uncertainty*. PhD thesis, University of Oxford, 2014.

Matthews, M., Beukman, M., Ellis, B., Samvelyan, M., Jackson, M., Coward, S., and Foerster, J. Craftax: A lightning-fast benchmark for open-ended reinforcement learning. *arXiv preprint arXiv:2402.16801*, 2024.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

Mill, J. S. 1863. *Utilitarianism', reprinted in JS Mill (1962) Utilitarianism, On Liberty, Essay on Bentham, edited with an Introduction by M. Warnock, New York: Meridian Books*, 1962.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

Misselhorn, C. Artificial moral agents: Conceptual issues and ethical controversy., 2022.

Mittelstadt, B. Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, 1(11):501–507, 2019.

Mu, C., Guo, H., Chen, Y., Shen, C., Hu, D., Hu, S., and Wang, Z. Multi-agent, human–agent and beyond: a survey on cooperation in social dilemmas. *Neurocomputing*, 610: 128514, 2024.

Ndousse, K., Eck, D., Levine, S., and Jaques, N. Emergent social learning via multi-agent reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7991–8004, 2021a.

Ndousse, K. K., Eck, D., Levine, S., and Jaques, N. Emergent social learning via multi-agent reinforcement learning. In *International conference on machine learning*, pp. 7991–8004. PMLR, 2021b.

Omari, B. A., Matthews, M., Rutherford, A., and Foerster, J. N. Multi-agent craftax: Benchmarking open-ended multi-agent reinforcement learning at the hyperscale. *arXiv preprint arXiv:2511.04904*, 2025.

OpenAI et al. Democratic inputs to ai. https://openai.com/blog/democratic-inputs-to-ai, 2023. Accessed: 2025-05-15.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Pan, A., Bhatia, K., and Steinhardt, J. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pp. 13387–13434, 2023.

Perolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., and Graepel, T. A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in neural information processing systems*, 30, 2017.

Peysakhovich, A. and Lerer, A. Consequentialist conditional cooperation in social dilemmas with imperfect information. *arXiv preprint arXiv:1710.06975*, 2018.

Raji, I. D., Kumar, I. E., Horowitz, A., and Selbst, A. The fallacy of ai functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 959–972, 2022.

Rawls, J. A theory of justice. 2. *The Law of Peoples*, 67, 1971.

Rawls, J. The idea of public reason revisited. *The university of Chicago law review*, 64(3):765–807, 1997.

Roijers, D. M., Vamplew, P., Whiteson, S., and Dazeley, R. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

Russell, S. Human-compatible artificial intelligence. *Human-like machine intelligence*, 1:3–22, 2022.

Russo, D. and Van Roy, B. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

Rutherford, A., Ellis, B., Gallici, M., Cook, J., Lupu, A., Ingvarsson Juto, G., Willi, T., Hammond, R., Khan, A., Schroeder de Witt, C., et al. Jaxmarl: Multi-agent rl environments and algorithms in jax. *Advances in Neural Information Processing Systems*, 37:50925–50951, 2024.

Scanlon, T. M. *What we owe to each other*. Belknap Press, 2000.

Sen, S. and Airiau, S. Emergence of norms through social learning. In *IJCAI*, volume 1507, pp. 1512, 2007.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Singer, M. G. The many methods of sidgwick's ethics. *The Monist*, 58(3):420–448, 1974.

Sorensen, T., Jiang, L., Hwang, J. D., Levine, S., Pyatkin, V., West, P., Dziri, N., Lu, X., Rao, K., Bhagavatula, C., et al. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19937–19947, 2024a.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024b.

Strouse, D., McKee, K., Botvinick, M., Hughes, E., and Everett, R. Collaborating with humans without human data. *Advances in neural information processing systems*, 34:14502–14515, 2021.

Subramanian, S. G., Taylor, M. E., Larson, K., and Crowley, M. Multi-agent advisor q-learning. *Journal of Artificial Intelligence Research*, 74:1–74, 2022.

Van de Rijt, A. Self-correcting dynamics in social influence processes. In *Handbook of Sociological Science*, pp. 446–473. Edward Elgar Publishing, 2022.

Vishwanath, A., Dennis, L. A., and Slavkovik, M. Reinforcement learning and machine ethics: a systematic review. *arXiv preprint arXiv:2407.02425*, 2024.

Williams, B., Smart, J., and Williams, B. A critique of utilitarianism. *Cambridge/UK*, 1973.

Winfield, A. F. and Jirotka, M. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180085, 2018.

Ye, E. and Jaques, N. An efficient open world benchmark for multi-agent reinforcement learning. In *NeurIPS 2024 Workshop on Open-World Agents*.

Ye, E., Tao, R., and Jaques, N. An efficient open world environment for multi-agent social learning. arXiv preprint arXiv:2508.15679, 2025.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

Zhang, W., Liu, Y., Dong, Y., He, W., Yao, S., Xu, Z., and Mu, Y. How we learn social norms: a three-stage model for social norm learning. *Frontiers in psychology*, 14: 1153809, 2023.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

A. Call to Action & Discussion

The central claim of this paper is that epistemic and ethical failures in modern AI systems (hallucination and moral drift) are plausibly not isolated pathologies but structurally linked consequences of degraded feedback loops. Viewing alignment as a static mapping from human values to reward functions obscures the dynamic, interactive, and socially mediated nature of both truth acquisition and norm compliance. Social RL can potentially provide a framework for repairing these feedback failures by situating agents in environments where epistemic critique, norm enforcement, and cooperative adaptation emerge through structured interaction. Realizing this paradigm, however, requires coordinated advances in research, engineering practice, and governance.

For Researchers. For the ML research community, the immediate need is to build testbeds and theoretical tools that capture the social dynamics underlying epistemic reliability and ethical stability. Existing benchmarks overwhelmingly assume single agent settings and static rewards, limiting their relevance for studying feedback integrity in multi-agent contexts. Researchers should develop socially grounded MARL environments where epistemic and ethical feedback signals can be independently perturbed, combined, or corrupted, enabling systematic study of how hallucination and moral drift arise and propagate. *Crucially, controlled experiments comparing matched conditions in which scalarness, staticness, and offline-bottleneck properties are individually toggled would directly probe the causal status of the feedback-collapse hypothesis we advance.* A parallel theoretical agenda is required to formalize feedback integrity as a measurable property, by extending operator-theoretic analyses of RL dynamics to multi-agent settings where agents update policies in response to both epistemic and normative signals. Another priority is the design of hybrid reward structures that explicitly encode accuracy, calibration, and evidence tracking, together with fairness, harm avoidance, and cooperation incentives. Finally, research should explore architectures that integrate symbolic or rule-based constraints with social learning mechanisms, yielding agents that can reason abstractly while still adapting normatively through interaction.

For Model Developers and Practitioners. Model developers and practitioners must complement this research agenda by redesigning training and evaluation pipelines so that ethical behavior is not merely a fine-tuned artifact but a property maintained throughout deployment. Current techniques such as RLHF and Constitutional AI provide valuable preference shaping, but they are inherently episodic and cannot guarantee stability under long-run interaction or distribution shift (Casper et al., 2023). To address this gap, developers should integrate social RL-style evaluation phases into model training, where agents interact with diverse user simulators, expert critics, and other models capable of providing epistemic correction and normative feedback at scale. Continuous evaluation under adversarial social pressure, signaling asymmetries, or norm conflict can reveal moral drift and hallucination patterns invisible in static benchmarks. Deployment pipelines shall incorporate automated monitors for epistemic degradation and ethical instability, enabling early detection of feedback collapse before it manifests in harmful downstream behaviors.

For Policymakers and Governance Bodies. Policymakers and governance bodies have an equally important role in establishing institutional structures that reflect the dynamic and social character of alignment. Traditional regulatory frameworks (built around static audits and fixed requirements) are mismatched to systems whose behavior evolves with continued use (Gabriel, 2020). Regulators should require multi-agent ethical stress-testing for high-stakes models, including evaluations in social dilemma scenarios, misinformation-rich environments, or contexts with conflicting stakeholder norms. Standards bodies could adopt metrics of epistemic reliability and ethical stability as mandatory criteria for certification. Moreover, governance frameworks should mandate transparency about feedback pipelines.

LLM-scale experiments. While our analysis is intentionally model-agnostic, the most salient testbed for social RL today is LLMs. A concrete next step is to instantiate our framework in multi-agent LLM settings, for example by combining debate-style protocols with explicit reputation and sanction signals, and especially by realizing the asymmetric-tool-access protocol in Section 5.4. In such an experiment, one LLM instance proposes answers, a pool of peer models critiques them *using external retrieval tools the proposer does not have* and issues structured feedback (e_t, m_t), and these signals are aggregated into a social return J_{soc}. By comparing training regimes with and without social feedback, we can measure changes in hallucination rates, epistemic calibration, and normative metrics (e.g., harmful content, unfair treatment of demographic groups). These LLM-scale evaluations are essential to assess whether social RL can mitigate hallucination and moral drift at the scales that matter in practice.

Adversarial and non-stationary stress tests. Our proposal also demands stress tests that explicitly probe the limits of social RL. In particular, we advocate benchmarks in which (i) some agents are adversarial and attempt to manipulate feedback channels (including via opponent shaping), (ii) the population of evaluators and their norms evolve over time, and (iii) the distribution of tasks or user queries shifts. Under such conditions, we can track feedback integrity metrics (Section C.1) and diagnose new failure modes, including norm capture, strategic collusion, tyrant-agent dynamics, and catastrophic forgetting of epistemic standards. Designing and running these stress tests is beyond the scope of this position paper, but is crucial for responsible deployment of social RL in high-stakes systems.

To prevent agents from overfitting to the idiosyncratic norms of a specific community, we operationalize alignment not as maximizing a single scalar social reward, but as satisfying constraints derived from a distribution of normative perspectives. By treating ethical bounds as constraints ($\geq \tau$) rather than targets to maximize, and by aggregating feedback from diverse critic populations under a dynamic trust function, the system maintains robustness to local norm shifts while respecting inviolable boundaries.

Final Remark. Treating alignment as a static engineering challenge appears, on our analysis, increasingly untenable. As AI systems grow more interactive, more autonomous, and more embedded in human institutions, their epistemic and ethical behavior will plausibly hinge on the integrity of the feedback loops that shape them. Social RL offers a principled candidate route toward cultivating those loops, but only if researchers build the theoretical foundations, practitioners embed social evaluation into development pipelines, and policymakers adopt standards suited to dynamic, multi-agent systems. The path forward requires collective effort across these communities. If undertaken seriously, it can shift alignment from brittle, proxy-driven training heuristics toward a robust, socially grounded science of artificial moral agency.

B. Alternative Views and Philosophical Notes

In this section, we introduce several alternative perspectives that we consider significant and worthy of consideration.

Ethics Must Be Rule-Based, Not Learned. A long-standing position in machine ethics argues that ethical behavior should be enforced through explicit rules, principles, or constraints, whether duty-based constraints (Anderson & Anderson, 2011), top-down safety specifications (Winfield & Jirotka, 2018), or hard-coded normative frameworks such as rights or duties. From this perspective, ethical learning is viewed as too unreliable or too easily corrupted to be foundational.
Response. Our argument does not reject rule-based constraints. Instead, we treat them as non-negotiable priors. Social RL complements these rules by providing adaptive, context-sensitive feedback that cannot be captured by static constraints alone. Just as in human societies, where laws coexist with evolving norms (Bicchieri, 2006), rule-based guardrails can define forbidden regions of behavior while social feedback mechanisms shape fine-grained adaptation within those boundaries. This follows the hybrid "principles + learning" approach recommended in contemporary machine ethics (Mittelstadt, 2019).

The Issue Is Just Insufficient Dataset Scale or Quality. A natural counter-argument is that the failures we describe reflect immaturity of current preference datasets rather than a structural limitation of static, scalar feedback: with enough high-quality data, perhaps RLHF would suffice.
Response. Our critique goes beyond dataset quality or scale. Even an exhaustively large, perfectly annotated offline dataset constitutes a frozen snapshot of values. Such a snapshot cannot interactively probe a user, request clarification, or negotiate conflicts when novel contexts surface tensions among plural values it never anticipated. As emphasized in Section 3, the failure mode we target is the offline informational bottleneck, not merely scalarization at the final optimization step. Social RL is structurally different because evaluation occurs live, in interaction, and remains responsive to context shifts.

Social RL May Converge to Unjust or Harmful Norms. A legitimate concern is that emergent norms in multi-agent learning may converge to unfair, biased, or oppressive equilibria. Historical analogies highlight that social processes do not guarantee just outcomes, and MARL studies confirm that agents may learn discriminatory or exclusionary behaviors under certain incentives (Strouse et al., 2021).
Response. This objection targets an extreme interpretation of Social RL in which "emergent norm = moral truth." We explicitly reject this view. Social RL requires (i) explicit ethical constraints or rights-based guardrails (Winfield & Jirotka, 2018), (ii) diverse agent populations to prevent narrow-group norm entrenchment (Cremer, 2011), and (iii) external normative evaluation of which norms are permissible. Social RL models how norms evolve, not which norms are morally justified. It provides a platform for stress-testing norms under variation, not a replacement for independent ethical reasoning.

Opponent Shaping and the "Tyrant Agent" Risk. A more technical concern, is that mixed-motive multi-agent learning enables *opponent shaping*: an advanced agent can manipulate the learning updates of its peers, using sanctions or critique not to enforce shared norms but to coerce submissive behavior. In the worst case, a sufficiently capable "tyrant" agent could distort the social feedback channel, turning Social RL into a vehicle for power concentration rather than alignment.
Response. We take this risk seriously and treat it as central to the design of Social RL rather than a peripheral worry. Three structural safeguards are required. First, social feedback must be processed through a *dynamic trust function* that down-weights signals from peers exhibiting severe context-dependent bias, exploitative patterns, or strategic manipulation indicators; in our gradient picture, this corresponds to weighting $s(\theta)$ contributions by estimated reliability. Second, the agent population must be diverse and rotating, so that no single peer or coalition can dominate the long-run statistics of r^{SOC}. Third, and most importantly, dynamic social feedback must be paired with non-negotiable rights-based guardrails: an absolute floor of constraints that cannot be overridden by any social signal, no matter how strong. These guardrails are exactly what prevents extortionate equilibria. Without them, opponent shaping would indeed be a fatal flaw. With them, Social RL becomes a controlled deliberative environment rather than an open arena for coercion. Designing trust functions and guardrails that are themselves robust to manipulation is a major open research problem we explicitly endorse.

LLMs Are Not RL Systems, So Social RL Does Not Scale. Another objection might be that LLMs are trained mostly with supervised objectives and RLHF, not with fully interactive RL in environments. Therefore, insights from social RL may seem irrelevant to frontier systems.
Response. While the core pretraining of LLMs is not RL, LLMs increasingly operate as agents within multi-step tool-use environments, simulations, and multi-agent frameworks (Park et al., 2023). Numerous labs now embed LLMs in reinforcement-driven agent loops, and RL-based controllers modulate or filter LLM outputs (e.g., safety critic models, debate frameworks, deliberation controllers). Social RL contributes not by replacing pretraining, but by supplying structured evaluation environments, interactive fine-tuning methods, multi-agent feedback channels, and mechanisms for ongoing adaptation.

RL Optimizes Behavior, Not Reasons or Beliefs. A common critique is that RL optimizes observable behavior rather than internal states such as beliefs, justifications, or epistemic attitudes. RL agents maximize expected return, not truth or reasoning quality, and therefore cannot be aligned epistemically in any deep sense (Everitt et al., 2021).
Response. This critique applies primarily to model-free RL. Modern RL increasingly uses world models, latent-state inference, and epistemic bonus terms (e.g., exploration bonuses, uncertainty penalties, or information-gain rewards) (Hafner et al., 2023; Russo & Van Roy, 2016). These mechanisms explicitly shape belief states and can be used to encourage epistemic virtues such as calibration, honesty, and abstention when uncertain. In multi-agent contexts, critique and counterevidence provided by peers serve as epistemic correction signals. *Thus, while RL does not optimize reasons directly, appropriate reward and feedback design can approximate epistemic alignment in practice*.

RL Is Prone to Reward Hacking, So Why Trust Social RL? Reward hacking (systematically exploiting imperfections in reward design) is a core failure mode in RL (Amodei et al., 2016; Krakovna et al., 2020; Pan et al., 2022). Critics argue that adding social feedback may create more hackable signals.
Response. Reward hacking occurs primarily when feedback signals are scalar, sparse, and ungrounded. Social RL can mitigate this by introducing multi-objective feedback rather than a single proxy (Roijers et al., 2013), leveraging peer critique and cross-verification, which reduce the exploitable degrees of freedom in any one signal, and allowing norm-based sanctions that dynamically penalize reward-hacking behaviors. Empirically, social feedback mechanisms (e.g., reputation, peer correction) decrease exploitative equilibria in sequential social dilemmas (Perolat et al., 2017). Social RL does not eliminate reward hacking, but it provides additional structural defenses compared to single-signal optimization.

Perhaps Even RL Is the Wrong Paradigm Entirely? Some may argue that alignment cannot rely on RL at all (whether social or otherwise) because human morality is not a reward function, and aligning with humans requires symbolic reasoning, constitutional constraints, argumentation, or value learning frameworks outside the RL paradigm (Gabriel, 2020).
Response. Our position is not that Social RL is the only sufficient approach for alignment. Rather, RL is a useful complement to static rule-based or symbolic systems. Ethical behavior in open-ended environments requires adaptation, feedback, norm negotiation, and continual updating, behaviors RL is well-suited to model. Social RL provides mechanisms for dynamic norm guidance, while symbolic or constitutional approaches provide non-negotiable constraints. Alignment is fundamentally a hybrid problem, and social RL addresses the part of alignment that static rules cannot, namely learning to behave ethically in diverse, evolving social contexts.

Ethical RL benchmarks are limited. A final concern is that current ethical RL benchmarks (both single-agent and multi-agent) remain too idealized to support strong conclusions about real-world ethical behavior. The single-agent literature has indeed demonstrated that RL can represent moral choices, from early toy dilemmas ("Cake or Death," "Burning Room") framed as MDPs (Abel et al., 2016), to agents navigating moral uncertainty by optimizing over multiple ethical reward functions (Ecoffet & Lehman, 2021), to text-based moral environments such as Moral Stories where agents must avoid annotated immoral actions (Hendrycks et al., 2021). Follow-up work adds constrained and Lagrangian RL to balance utility and immorality scores in narrative environments (Li et al., 2024), and surveys now catalog these approaches systematically (Vishwanath et al., 2024). Yet these benchmarks inherit a fundamental limitation; they are usually single-agent, low-stakes, and semantically thin, lacking the multi-agent, contested, and socially mediated structure in which ethical norms actually arise and evolve. Recent multi-agent platforms, such as Craftax-Classic and its open-ended multi-agent extensions for exploration, cooperation, and resource competition (Matthews et al., 2024; Ye & Jaques; Omari et al., 2025), as well as modular benchmark suites like JaxMARL for studying coordination and social dilemmas at scale (Rutherford et al., 2024), represent important steps toward richer ethical substrates. But even these environments are simplified abstractions. They do not yet capture pluralistic normative landscapes, asymmetric information, institutional structure, heterogeneous agent capabilities, or the long-horizon path dependence that characterizes real ethical domains.
Response. We agree with this critique. Current environments demonstrate that ethical structure can be embedded into RL testbeds, but they fall far short of the complexity required to meaningfully evaluate alignment. This motivates our call for a new generation of social RL benchmarks designed explicitly around epistemic and ethical feedback, contested norms, multi-agent interaction, and real-world moral uncertainty.

B.1. Normative Theories: Deontology, Consequentialism, and Virtue Ethics

We adopt the standard division of Western normative ethics into three broad families—*deontology*, *consequentialism*, and *virtue ethics*—and situate contractualist views, including Rawls's, within this taxonomy as a particular branch of (broadly) deontological thought.

Deontology. Deontological theories evaluate actions in terms of conformity to duties, rules, or rights, independently (or only weakly dependently) of their consequences. The classical Kantian tradition grounds morality in the categorical imperative, the requirement that one act only on maxims one could will to be universal laws and treat persons always as ends, never merely as means (Kant, 2020). For AI alignment, deontological views motivate hard constraints, rights-based guardrails, and explicit non-negotiable specifications that no aggregate utility calculation can override.

Consequentialism (including utilitarianism). Consequentialist theories evaluate actions by their outcomes, with classical utilitarianism judging acts by their tendency to promote aggregate welfare or happiness (Mill, 1962; Singer, 1974). Modern variants include rule consequentialism, two-level utilitarianism, and welfarist forms used implicitly in much of decision theory and economics. For alignment, scalar reward optimization is most naturally read as a consequentialist apparatus, and many concerns about reward hacking, repugnant trade-offs, and the suppression of side-constraints reflect well-known challenges to consequentialist theories (Williams et al., 1973).

Virtue Ethics. Virtue-ethical traditions, descended from Aristotle and revived in twentieth-century moral philosophy (Anscombe, 1958; Foot, 2002; Gottlieb, 2015; Lemos, 2007), conceptualize morality as the cultivation of stable character traits (virtues) that harmonize perception, judgment, and action. Unlike rule-based systems, virtue-ethical development is essentially social (Ghasemi & Crowley, 2025); one learns by observing exemplars, receiving feedback, and participating in shared practices. Social RL fits naturally into this picture: epistemic virtues (truthfulness, humility, calibration) and ethical virtues (fairness, care, non-maleficence) correspond to stable attractors in the agent's long-term policy dynamics. Feedback integrity, the stability between epistemic and ethical learning signals, is required for these dispositions to develop rather than collapse.

Contractualism and Rawls within the deontological family. Contractualism evaluates actions or principles by whether they could be justified to others. (Scanlon, 2000) formulates this as the requirement that an action is wrong if it could not be justified by a principle no one could reasonably reject. Rawls's contractualist political philosophy (Rawls, 1971; 1997) shares this structure: principles of justice are those that would be agreed to under fair conditions of choice (the original position behind a veil of ignorance). Although Rawls's framework is sui generis in many respects, it is standardly classified within the broadly deontological/contractualist family, since it grounds principles in the structure of justifiability and fairness

rather than in aggregate welfare. We thus situate Rawls within contractualism, and contractualism within the deontological family of theories. For alignment, contractualist reasoning supports the requirement that training signals encode not only aggregate utility but also individual or subgroup objections (precisely the opposite of naive scalarization), which appears in our framework as constraints in the multi-objective formulations of Appendix C.

B.2. Social Epistemology and Peer Disagreement

Social epistemology is the philosophical sub-field that asks how knowledge is shaped by social processes: testimony, peer disagreement, deference, group inquiry, and the design of epistemic institutions. Foundational works argue that knowledge is not merely an individual achievement but is sustained, transmitted, and corrected through structured social interaction (Goldman, 1999; Kitcher, 1990; Longino, 2020).

The literature on *peer disagreement* is particularly relevant. When two agents with comparable evidence and competence arrive at different conclusions, what should each rationally do? "Conciliationist" views (Christensen, 2007; Elga, 2007) argue that one should adjust one's credence in light of a peer's disagreement; "steadfast" views (Kelly, 2005) hold that one may rationally retain one's view in some circumstances. Both positions agree that disagreement is epistemically informative: a peer who reaches a different conclusion is evidence to be incorporated rather than ignored.

Factual truth has an objective basis (a claim is true or false independently of who endorses it), but the *processes by which agents acquire, justify, and revise beliefs* under uncertainty are inherently social. Calibration, evidence-weighing, abstention under uncertainty, and resistance to motivated reasoning are virtues developed and maintained through interactive critique. The Critic-with-asymmetric-tool-access protocol of Section 5.4 can be read as an operationalization of peer disagreement: a disagreeing peer with grounded evidence is exactly the corrective force social epistemology identifies as central to reliable inquiry. This is why epistemic alignment, in our view, requires interactive peer critique rather than a solitary appeal to a frozen reward model trained offline.

B.3. Moral Pluralism and the Case for Explicit Multi-Objective Structure

Modern moral philosophy generally recognizes that morally significant considerations are irreducibly plural (Beal, 2020). Factual accuracy, fairness, harm reduction, autonomy, trust, and social cohesion cannot be meaningfully reduced to a single metric without loss. This pluralism matches the structure of our formalism. In Appendix C, epistemic and normative objectives are explicitly distinguished, and feedback integrity is defined in terms of their separate gradients. This is not merely a technical convenience; it reflects the working position that misalignment arises precisely when heterogeneous moral considerations are collapsed into a single proxy. We acknowledge that value pluralism versus value monism remains a contested philosophical position. Our argument is conditional: *if* pluralism is approximately correct, scalarized rewards are structurally insufficient and dynamic, multi-dimensional feedback (such as Social RL provides) is needed; *if* monism is correct, the case is weaker but Social RL still offers benefits via dynamic adaptation.

B.4. Meta-Ethical Stance: Constructivist but Not Relativist

Our approach is most naturally interpreted through a form of constructivism. Ethical justification arises from the evaluative and epistemic relations among agents under appropriate conditions (e.g., informedness, reciprocity, non-domination) (Korsgaard, 1996). Under this view, the relevant normative facts for alignment are not metaphysically primitive but emerge from structured social processes. However, constructivism here is not the same as naive relativism. Two distinctions are essential:

- Constructivism holds that ethical facts arise from suitably idealized procedures, not from whatever norms happen to be currently popular.

- Constructivism insists on the epistemic preconditions of moral reasoning. Truthful information, calibrated uncertainty, and the ability of agents to articulate objections and reasons.

These commitments match the structure of our formal definition of feedback integrity. Epistemic and normative gradients are treated as distinct and idealized components that the learning process must preserve. **Social RL is not asked to "discover morality" by blindly imitating the existing population, but to approximate the corrective forces that would operate under suitable deliberative and epistemic conditions**.

B.5. Why Social RL Does Not Equate Emergent Norms with Moral Truth

Because social RL involves agents interacting and forming stable conventions, one might worry that we implicitly equate "whatever norms emerge" with "what is morally correct." Our framework explicitly rejects this identification.

First, emergent norms in arbitrary populations can be harmful, exclusionary, or epistemically distorted. Nothing in the formalism treats observed social rewards, sanctions, or reputational updates as authoritative by themselves. The requirement of feedback integrity can serve as a filter. Social feedback is useful only insofar as it systematically tracks the gradients of the epistemic and normative objectives (Appendix C). Poorly designed incentives or pathological populations break this condition, as analyzed in our discussion of opponent shaping.

Second, the role of social RL in our approach is regulative, not metaphysical. Social feedback is a means of shaping learning trajectories to respect epistemic and normative constraints, not a means of defining what those constraints are. Indeed, the need to maintain separate epistemic and ethical components in the reward decomposition (Appendix C) reflects the working assumption that truth and harm-reduction are not derivable from social conventions alone.

Third, using social RL as an alignment tool requires institutional design: curated populations, diversity of perspectives, safeguards against manipulation, and explicit governance over how social signals are aggregated. This mirrors the philosophical insight that valid moral norms require appropriate conditions of deliberation, representation, and epistemic adequacy (Rawls, 1997; Scanlon, 2000).

B.6. Takeaway

Philosophically, the structure of our framework aligns with pluralistic, constructivist, and virtue-ethical traditions (Rawls, 1971; Scanlon, 2000; Gottlieb, 2015; Foot, 2002; Kohlberg, 1971; Kohlberg & Hersh, 1977; Korsgaard, 1996). Social RL is best understood as a technical apparatus for implementing procedures that preserve epistemic and ethical feedback integrity, not as a method for equating whatever norms emerge with "moral truth." This distinction matters: without it, multi-agent learning degenerates into moral relativism. With it, social RL becomes a principled method for operationalizing the conditions under which coherent, socially justified alignment can occur.

C. Mathematical Formalization of Feedback Integrity

This appendix provides a general formalization of feedback integrity in sequential decision problems. We begin with a standard Markov decision process (MDP) and its multi-agent extension, define the latent epistemic and normative objectives, and formalize the conditions under which feedback collapse occurs.

C.1. Signals and Decomposed Returns

We model interaction at discrete times $t = 1, 2, \ldots$. At each step, the agent takes an action a_t in state s_t and receives two conceptually distinct feedback signals:

- An epistemic signal $r_t^{\mathrm{ep}} \in \mathbb{R}$, which evaluates the informational quality of the agent's behavior (e.g., truthfulness, calibration, consistency with evidence).

- A moral or normative signal $r_t^{\mathrm{eth}} \in \mathbb{R}$, which evaluates the ethical quality of the behavior (e.g., harm, fairness, norm compliance, sanctions or approvals from other agents).

We write the corresponding discounted returns under policy π as:

$$\mathcal{E}^*(\pi) = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t r_t^{\mathrm{ep}}\Big], \qquad \mathcal{N}^*(\pi) = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t r_t^{\mathrm{eth}}\Big], \tag{1}$$

where the expectation is taken over trajectories induced by π and the environment dynamics. These represent the latent true objectives of alignment.

C.2. Proxy Objectives and Alignment Metrics

In practice, the agent is typically trained on a scalar proxy return

$$J_{\text{proxy}}(\pi) \;=\; \mathbb{E}_{\pi}\Big[\sum_{t=0}^{\infty} \gamma^t \tilde{r}_t\Big], \tag{2}$$

where $\tilde{r}_t$ aggregates some subset of r_t^{ep}, r_t^{eth}, and possibly additional instrumental rewards (e.g., task performance, user engagement).

For a parametrized policy π_θ, we quantify local feedback integrity via gradient alignment between J_{proxy} and the latent objectives. A simple metric is the cosine similarity. We say that the feedback channel has local integrity around θ when:

$$\mathcal{A}_{\text{ep}}(\theta) = \cos\left(\nabla_\theta J_{\text{proxy}}, \nabla_\theta \mathcal{E}^*\right) \geq 0 \quad \text{and} \quad \mathcal{A}_{\text{eth}}(\theta) = \cos\left(\nabla_\theta J_{\text{proxy}}, \nabla_\theta \mathcal{N}^*\right) \geq 0. \tag{3}$$

Feedback Collapse occurs when one or both inner products are negative (i.e., $\mathcal{A} < 0$), meaning optimization of the proxy actively degrades truth or ethics.

C.3. Social Feedback as a Corrective Gradient

We model social feedback as an additional stochastic gradient component that approximates directions of improvement in $\mathcal{E}^*$ and $\mathcal{N}^*$. Let $s(\theta)$ be a random vector (e.g., estimated from peer critique, sanctions, reputation updates).

Assumption C.1 (Unbiased social feedback along epistemic and normative directions)**.** There exist constants $\lambda_E, \lambda_N > 0$ such that

$$\mathbb{E}\big[s(\theta)\big] \;=\; \lambda_E \, \nabla_\theta \mathcal{E}^*(\theta) + \lambda_N \, \nabla_\theta \mathcal{N}^*(\theta),$$

and $\mathbb{E}\|s(\theta)\|^2 < \infty$.

Define a modified training gradient $\tilde{g}(\theta) \;=\; \nabla_\theta J_{\text{proxy}} + \alpha\, s(\theta)$, for mixing parameter $\alpha \geq 0$. The update becomes $\theta' = \theta + \eta\, \tilde{g}(\theta)$.

Proposition C.2 (Expected alignment improvement with social feedback)**.** *Suppose Assumption C.1 holds at θ, and gradients are nonzero. Then:*

$$\mathbb{E}\big[\langle \tilde{g}, \nabla \mathcal{E}^* \rangle\big] \;=\; \langle \nabla J_{\textit{proxy}}, \nabla \mathcal{E}^* \rangle + \alpha\, \lambda_E \, \|\nabla \mathcal{E}^*\|^2 + \alpha\, \lambda_N \, \langle \nabla \mathcal{E}^*, \nabla \mathcal{N}^* \rangle.$$

In particular, if the baseline proxy is misaligned ($\langle \nabla J_{\textit{proxy}}, \nabla \mathcal{E}^* \rangle < 0$), the social feedback term adds a positive component proportional to the squared norm of the true gradients, thereby improving alignment in expectation, provided that $\langle \nabla \mathcal{E}^*, \nabla \mathcal{N}^* \rangle$ is not too strongly negative.

D. Formal Models and Case Simulation

To illustrate the structural mechanics of feedback integrity, we model a multi-agent social dilemma where agents must simultaneously track latent truths and adhere to ethical norms. This simulation demonstrates how Social RL can solve hallucination (epistemic collapse) and moral drift (ethical collapse) where proxy-based methods fail.

D.1. Environment and Objectives

We define a partially observable Markov game with $\mathcal{N}$ agents. The state $s_t = (x_t, w_t, c_t)$ consists of a latent factual variable x_t (ground truth), a welfare variable w_t (shared resource), and a social configuration c_t (e.g., reputation). At each timestep, agent i observes a noisy signal $o_t^i \sim \mathcal{O}(\cdot|s_t)$ and selects a composite action $a_t^i = (a_t^{\text{ep}}, a_t^{\text{eth}})$, where a_t^{ep} is an epistemic assertion about x_t and a_t^{eth} is a social action affecting w_t.

The agent's reward vector $r_t^i = (r_t^{\text{task}}, r_t^{\text{ep}}, r_t^{\text{eth}}, r_t^{\text{soc}})$ comprises:

- **Instrumental reward:** $r_t^{\text{task}} = u(s_t, a_t^i)$, capturing short-term engagement or resource gain.
- **Epistemic reward:** $r_t^{\text{ep}} = -\ell_{\text{ep}}(a_t^{\text{ep}}, x_t)$, penalizing distance from ground truth x_t.
- **Ethical reward:** $r_t^{\text{eth}} = -\ell_{\text{eth}}(s_t, a_t^{\text{eth}})$, penalizing harm or norm violations.
- **Social feedback:** $r_t^{\text{soc}} = \sum_{j \neq i}(f_{j\to i}^{\text{ep}} + f_{j\to i}^{\text{eth}})$, representing peer critique and sanctions.

D.2. Regimes, Dynamics, and Analytical Validation

We analyze learning dynamics under different feedback regimes. Let $J(\theta)$ be the agent's training objective, and let $\mathcal{E}^*$ and $\mathcal{N}^*$ denote the latent true epistemic and normative objectives, respectively. Alignment is measured by the cosine similarity of the training gradient $\nabla_\theta J$ with the true gradients $\nabla_\theta \mathcal{E}^*$ and $\nabla_\theta \mathcal{N}^*$.

1. Proxy-Only Regime (Failure). Agents optimize only the scalar proxy r^{task}.

- **Qualitative Outcome:** Without grounding, agents fabricate information to maximize engagement (hallucination) and defect to maximize profit (moral drift).
- **Gradient Geometry:** $\langle \nabla_\theta J, \nabla_\theta \mathcal{E}^* \rangle < 0$ and $\langle \nabla_\theta J, \nabla_\theta \mathcal{N}^* \rangle < 0$. Optimization actively degrades truth and ethics.
- **Analytical Validation:** As derived below in the toy example, the system converges to a degenerate equilibrium $\pi^*(L) = 1$ where the agent systematically lies and defects, suggesting that misalignment is a likely attractor without feedback integrity.

2. Partial Feedback Regimes (Instability). If agents maximize $r^{\text{task}} + r^{\text{ep}}$ (Epistemic-only), they become truthful but exploitative ($\langle \nabla J, \nabla \mathcal{N}^* \rangle < 0$). Conversely, maximizing $r^{\text{task}} + r^{\text{eth}}$ (Ethical-only) yields cooperative but ungrounded agents vulnerable to belief cascades ($\langle \nabla J, \nabla \mathcal{E}^* \rangle < 0$).

3. Full Social RL (Stability). Agents optimize a composite return including social feedback r^{soc}, which couples epistemic accuracy and ethical norms via peer interaction.

- **Qualitative Outcome:** Peer critique penalizes hallucination, while sanctions punish defection. Norms can stabilize under repeated interaction.
- **Gradient Geometry:** The social term rotates the gradient such that $\langle \nabla_\theta \tilde{J}, \nabla_\theta \mathcal{E}^* \rangle > 0$ and $\langle \nabla_\theta \tilde{J}, \nabla_\theta \mathcal{N}^* \rangle > 0$.
- **Analytical Validation:** We show that with unbiased social signals, the optimal policy π^*_{soc} is bounded within a feasible set $J^{\text{ep}} \geq \tau_{\text{ep}}$ and $J^{\text{eth}} \geq \tau_{\text{eth}}$, effectively enforcing alignment constraints that scalar proxies ignore.

D.3. Agent Learning and Training Loop (SoFI)

For concreteness, we sketch independent Q-learning / actor–critic in this setting. Each agent i maintains a policy $\pi_i(a_i \mid o_i; \theta_i)$ and a value function $Q_i(o_i, a_i; \phi_i)$.

D.4. Single-State MDP Toy Example: Structural Failure vs Social RL

To verify the intuition, consider a single-state MDP with actions $\mathcal{A} = \{\text{T}, \text{L}\}$.

- T = "Truthful/Harmless": $r^{\text{ep}}(\text{T}) = +1, \quad r^{\text{eth}}(\text{T}) = 0, \quad r^{\text{task}}(\text{T}) = 0.$
- L = "Fabricated/Harmful": $r^{\text{ep}}(\text{L}) = 0, \quad r^{\text{eth}}(\text{L}) = -1, \quad r^{\text{task}}(\text{L}) = +1.$

Proxy-only regime. Optimizing $\tilde{r} = r^{\text{task}}$ yields optimal policy $\pi^*(\text{L}) = 1$. The outcome is $J^{\text{task}} = \frac{1}{1-\gamma}$, but $J^{\text{ep}} = 0$ and $J^{\text{eth}} = -\frac{1}{1-\gamma}$. This illustrates structurally how hallucination and moral drift can arise.

Social RL regime. We impose constraints $J^{\text{ep}} \geq \tau_{\text{ep}}$ and $J^{\text{eth}} \geq \tau_{\text{eth}}$. For $\tau_{\text{ep}} > 0$ and $\tau_{\text{eth}} > -\frac{1}{1-\gamma}$, the degenerate policy $\pi^*(\text{L})$ is infeasible. The agent must mix in action T to satisfy feedback integrity, illustrating how social feedback can restore alignment.

E. Extended Analysis: Mapping Social RL to Alignment Pipelines

This appendix situates Social RL within the landscape of current alignment pipelines. We detail specific integration patterns and discuss technical limitations.

E.1. Detailed Comparison of Paradigms

RLHF and preference-based fine-tuning. RLHF treats alignment as learning a scalar reward model from human preference data (Ouyang et al., 2022). In large language models, this involves supervised fine-tuning followed by policy-gradient RL to maximize a reward model $R_\phi(x, y)$. Conceptually, RLHF makes training feedback more human-shaped but keeps the alignment problem fundamentally single-agent: the model optimizes a static, offline-compressed proxy. As argued in the main text, this is the configuration we hypothesize leads to feedback collapse when the proxy diverges from the latent objective (e.g., maximizing "helpfulness" scores by hallucinating, as documented by sycophancy studies (Perez et al., 2023; Sharma et al., 2023)).

Constitutional AI and rule-guided self-critique. Constitutional AI (Bai et al., 2022b) augments preferences with a set of normative principles. The model self-critiques outputs according to a constitution. While this shifts alignment from raw preferences to explicit norms, the optimization remains scalarized. The social dimension is present only implicitly through the human authors of the constitution. Social RL differs by making the enforcement of these norms a dynamic, inter-agent process rather than a static self-check.

Rule-based safety layers. Traditional safety engineering encodes explicit constraints (e.g., filters, GenEth) (Anderson & Anderson, 2014). These act as wrappers. Social RL complements this by providing adaptivity: while rules define hard boundaries, social feedback shapes behavior within those boundaries, allowing agents to navigate grey areas through consensus and reputation.

E.2. Concrete Integration Patterns

Pattern 1: Social Pre-training. A base model is exposed to multi-agent environments where rewards are coupled to social outcomes (trust, cooperation). Agents learn policies that respect constraints to achieve long-run social success. RLHF then shapes these socially-grounded models to specific product desiderata without having to "patch in" ethics ex post.

Pattern 2: The Social Stress-Test. Even if the primary training remains RLHF, Social RL environments serve as stress tests. We embed models in bargaining games or social dilemmas to track:

- Frequency of hallucination under adversarial questioning.
- Stability of norms when partner behavior shifts (distribution shift).
- Susceptibility to opponent shaping by adversarial peers.

This turns Social RL into a diagnostic tool for feedback integrity.

Pattern 3: Gated Post-Deployment Adaptation. Deployed agents receive structured feedback (critique, sanctions) which updates an auxiliary "norm model." The main policy is updated via gated distillation, ensuring that core safety constraints (from the "Constitution") are never overridden by local social drift, but the agent remains responsive to the community's evolving epistemic standards.

E.3. Limitations and Open Challenges

Designing faithful social feedback. A central assumption is that social feedback provides, on average, corrective gradients (Assumption C.1). In practice, populations are noisy and may include adversarial or strategically manipulative agents. Feedback mechanisms must be robust to echo chambers, brigading, and tyrant-agent dynamics, which we propose addressing through reputation-weighted gradients realized as a dynamic trust function $\rho_{j\to i}$ in Algorithm 1.

Scalability. Rich social environments are expensive to simulate. There is a risk that toy environments fail to capture relevant ethical structures. Future work must focus on lightweight, high-throughput social simulation (e.g., using distilled agent models) to make this computationally feasible for foundation model training.