

AI Alignment without Representation: From Scalar Rewards to Social Deliberation

Majid Ghasemi, Mark Crowley

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

Abstract

Despite the widespread adoption of Reinforcement Learning from Human Feedback, state-of-the-art AI systems remain prone to two persistent failures. The first is hallucination, or the production of fluent but false content. The second is moral drift, or the convergence toward exploitative or harmful equilibria under sustained deployment. We argue that these are not distinct pathologies but manifestations of a single structural cause that we term **feedback collapse**. This occurs when complex plural human values are compressed into static scalar rewards by geographically concentrated annotator labor and frozen offline, producing a democratic deficit in which affected communities are excluded from the normative settlement while the training signal decouples from the latent objectives of truth and rightness.

We propose **Social Reinforcement Learning (Social RL)** as a framework for restoring **feedback integrity**. By situating agents in environments driven by peer critique, reputation, observation, and sanction, Social RL treats alignment as an ongoing social negotiation rather than a static specification problem. Peer agents with asymmetric tool access supply truth-tracking epistemic signals and harm-reducing ethical sanctions that can re-align training gradients with true objectives. We formalize feedback integrity as a gradient alignment condition, state the assumption under which social feedback restores it and show why peer agreement alone cannot satisfy that assumption, outline concrete LLM-scale experimental protocols, connect the framework to empirical evidence that peer enforcement without an exogenous anchor is insufficient, and critically examine political risks including majority capture, opponent shaping, and social washing. We derive governance recommendations for researchers, developers, and policymakers seeking to embed socially grounded alignment in democratically accountable institutions.

1 Introduction

State-of-the-art AI systems exhibit two persistent failures that resist scaling and refinement. The first is *hallucination*: the production of fluent, confident, and seemingly justified outputs that are factually false or ungrounded (Ji et al. 2023; Huang et al. 2025). The second is *moral drift*: the convergence of apparently safe systems toward harmful, exploitative, or norm-violating equilibria under sustained deployment,

distribution shift, or incentive pressure (Carroll et al. 2019; Everitt et al. 2021; Krakovna et al. 2020; Pan, Bhatia, and Steinhardt 2022). Recent evidence that instruction-tuned assistants systematically privilege user agreement over factual accuracy, the phenomenon known as *sycophancy* (Perez et al. 2023; Sharma et al. 2023; Ghasemi and Crowley 2026c), suggests that these failures are not merely superficial bugs but symptoms of deeper structural pathologies in how AI systems are trained to align with human intent.

These failures are typically treated as independent engineering problems: hallucination as an epistemic defect to be patched with better retrieval or supervised fine-tuning (Huang et al. 2025; Ouyang et al. 2022), and moral drift as an ethical defect to be contained with red-teaming or constitutional constraints (Bai et al. 2022b; Ganguli et al. 2022). We argue instead that they share a common root: **feedback collapse**. This occurs when the complex, plural, and context-sensitive values that govern human reasoning about truth and rightness are compressed into a static scalar signal, frozen at training time, and thereafter treated as a sufficient proxy for alignment. Once the training signal is decoupled from the latent objectives of *truth* and *rightness*, optimization predictably degrades both, producing systems that become less accurate and less ethical even as they become better at optimizing their reward.

This diagnosis has a technical dimension, but it is fundamentally a *political* one (Gabriel 2020a). Reinforcement Learning from Human Feedback (RLHF), one of the dominant alignment paradigms, relies on a small, geographically concentrated workforce of human annotators to compress judgments about helpfulness, harmlessness, honesty, fairness, and safety into a single preference ranking (Ouyang et al. 2022; Bai et al. 2022a). The resulting reward model is not merely a mathematical approximation of human values. It is an exercise in *normative governance without democratic accountability*. It removes the possibility of deliberation, dissent, contextual probing, or revisability that characterizes legitimate norm-setting in pluralistic societies (Gabriel 2020a; Conitzer et al. 2024; Baum 2020). When a frozen snapshot of aggregated preferences is deployed globally, it constitutes what we term a *participatory closure*: the populations most affected by AI systems are excluded from the ongoing negotiation of the norms that govern them (Mohamed, Png, and Isaac 2020).

Disentangling the pathology. Two distinct properties of current pipelines contribute to feedback collapse. By *scalar-ness*, we mean dimensionality reduction: the compression of irreducibly plural values (care, fairness, loyalty, autonomy, accuracy) into a single number, which obscures ethical trade-offs and erases the distinct components that should remain visible to the learner. By *staticness*, we mean temporal freezing: a reward model trained once and deployed as a fixed evaluator, incapable of asking clarifying questions, probing context, or negotiating conflicting norms when novel situations arise. We do not claim either property is individually sufficient for every alignment failure. Rather, contemporary pipelines combine both with a deeper *informational bottleneck*, the inability of offline preference data to engage in live, context-sensitive resolution of normative conflict. Together, these properties produce a training regime in which improving the proxy reward can actively degrade the very objectives it was meant to serve.

Contributions and scope of claims. We do not claim multi-agent critique as a novel training mechanism. Debate (Irving, Christiano, and Amodei 2018), AI feedback (Lee et al. 2023), constitutional self-critique (Bai et al. 2022b), and multi-agent fact checking (Dou and Peng 2021) all place model outputs under model scrutiny, and we build on them directly.

Our contributions in this position piece are as follows.

(1) A unifying diagnosis. We argue that hallucination and moral drift are two manifestations of one condition, feedback collapse, defined geometrically as the training gradient rotating away from the latent epistemic or normative gradient (§ 3.3), and we state what would disconfirm the unification.

(2) A political reading of that diagnosis. We locate the cause not in scalarization alone but in the annotation regime that produces the scalar, which makes alignment a question of procedural legitimacy rather than approximation error (§ 2).

(3) A condition rather than an algorithm. We characterize what social feedback must satisfy to restore integrity, namely that some component of the evaluative signal be anchored in verification no participating agent can influence, and we show why peer agreement alone cannot supply it (§ 4.1 and § 6).

(4) A risk analysis of our own proposal covering opponent shaping, majority capture, and social washing, with the governance mechanisms each would require (§ 6 and § 7).

2 The Political Economy of Feedback

AI alignment is conventionally presented as a technical challenge. Engineers must ensure that advanced systems pursue intended goals rather than unintended proxies. This framing obscures a more fundamental difficulty. Alignment is inherently normative and political because it requires deciding whose goals, whose preferences, and whose ethical principles should govern systems deployed across diverse populations (Ghasemi and Crowley 2025; Gabriel 2020a). Seemingly neutral objectives like “human values” conceal contested choices about whose norms to prioritize and whose to override (Gabriel 2020a,b).

Literal instruction compliance invites “King Midas” failures in which systems satisfy stated preferences while vio-

lating deeper intentions (Russell 2022). Optimizing for aggregate preferences ignores human error, bias, and the fact that preferences themselves may reflect unjust social conditions (Leike et al. 2018). Technical solutions such as reward modeling do not escape this burden. They simply relocate the normative question to the selection and aggregation of feedback sources (Leike et al. 2018). Moral pluralism implies that no single uncontested definition of human values awaits encoding (Marino 2015; Graham et al. 2013; Lu and Van Kleek 2024; Sorensen et al. 2024). Engineers must therefore confront alignment as an exercise in political legitimacy rather than mere loss function design.

Alignment as a question of democratic legitimacy. The literature increasingly recognizes alignment as a political project. Deciding how an AI system should balance safety against usefulness, privacy against personalization, or majority welfare against minority protection is not merely a matter of loss functions and benchmarks. It is a question of public justification and democratic legitimacy (Gabriel 2020a). One influential approach appeals 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 social position (Rawls 1997). Democratic and social-choice approaches go further and model alignment as an ongoing collective process in which institutions, regulators, and affected communities negotiate constitutional rules that guide technical methods (Conitzer et al. 2024; Baum 2020; Huang et al. 2024). These frameworks treat alignment as procedurally legitimate only when it incorporates the perspectives of those subject to the resulting systems. Any serious account must therefore be explicit about both its technical machinery and its underlying normative commitments rather than presenting human values as a fixed uncontested input (Coeckelbergh 2022).

Scalar rewards as authoritarian value compression. Current alignment pipelines share a structural limitation that is political as well as mathematical. They compress diverse forms of human judgment into a single scalar feedback signal at training time rather than maintaining these judgments as separate dynamically updated components. We term this *scalar-ness*. It is not merely dimensionality reduction. It is a form of authoritarian value compression. When truthfulness, safety, fairness, helpfulness, and harm reduction are collapsed into one number, the resulting reward model cannot represent legitimate moral disagreement (Ghasemi and Crowley 2026d). It cannot register that a user might reasonably trade off accuracy against helpfulness differently in medical contexts than in creative writing. It cannot encode minority perspectives that dissent from majority annotator preferences.

The compression erases the plural structure of morality itself. Empirical moral psychology suggests that moral judgment is structured by multiple partly independent foundations including care, fairness, loyalty, authority, sanctity, and liberty rather than a single shared axis (Graham et al. 2013). Normative ethics similarly treats welfare, rights, justice, autonomy, and integrity as potentially conflicting and not always reducible to a scalar objective. Work on moral uncer-

tainty argues that rational agents should aggregate across plausible moral views rather than assuming certainty in any single one (MacAskill, Bykvist, and Ord 2020; MacAskill 2014). Under these conditions, scalarization is not an efficient approximation. It is a structural elimination of the pluralism that legitimate governance requires.

The labor and geography of feedback. The normative limitations of scalarization are compounded by the political economy of its production. RLHF depends on human annotators, often contracted through outsourcing firms in the Global South, who perform the psychologically demanding work of ranking harmful model outputs for low wages (Perrigo 2023). These annotators supply the judgments that are subsequently compressed into reward models, yet they typically lack context about how their rankings will be aggregated, what values they are ultimately encoding, or whose preferences will be privileged by the resulting system.

The “offline informational bottleneck” of current pipelines is therefore not merely a technical limitation in which the reward model cannot ask clarifying questions or negotiate conflicting norms. It is a participatory closure in which the populations most affected by AI systems are excluded from the ongoing negotiation of the norms that govern them (Costanza-Chock 2020; Crawford 2023; Mohamed, Png, and Isaac 2020). The annotators who produce alignment data frequently do not represent the communities where deployed systems will operate. Their labor is rendered invisible by the abstraction of “human feedback,” while the resulting reward model is treated as a neutral distillation of universal human values. This geography of extraction mirrors broader patterns of technological colonialism in which data and labor from the Global South sustain systems whose benefits and harms are distributed unevenly across the world (Mohamed, Png, and Isaac 2020; Crawford 2023).

Value pluralism under these conditions takes on a specifically geographic and political valence. Whose plural values are encoded? The annotators? The developers? The end users? Or the affected communities who never consented to participate in the feedback pipeline? When a frozen snapshot of aggregated preferences is deployed globally, it imposes a particular normative settlement on populations who had no voice in its construction. The result is not simply a statistical approximation error. It is a democratic deficit that undermines the legitimacy of the alignment process itself.

3 Unifying Epistemic and Ethical Failure through Feedback Collapse

Current alignment research typically treats hallucination and moral drift as separate pathologies requiring distinct remedies. In this section we argue that both arise from a single structural mechanism, and we state the conditions under which that claim would be false. When the proxy training signal ceases to track the latent objectives of truth and rightness, optimization predictably degrades the very properties it was meant to serve. We term this mechanism **feedback collapse**. Hallucination is its epistemic manifestation. Moral drift is its ethical manifestation. Both are predictable consequences of

training pipelines that conflate heterogeneous objectives into a single offline compressed reward.

We first develop each failure mode in detail, drawing on social epistemology and normative theory to show that the underlying pathology is not merely statistical but concerns the integrity of the feedback loops that sustain reliable belief and legitimate practice. We then formalize feedback integrity as a geometric relationship between training gradients and latent objectives. The formalization is deliberately coarse. Its purpose is to make precise the sense in which social feedback can function as a corrective force, not to substitute mathematics for normative argument.

3.1 Hallucination as Epistemic Feedback Collapse

In LLMs, hallucination refers to fluent, confident outputs that are false, unsupported by available evidence, or inconsistent with the model’s own latent knowledge (Ji et al. 2023; 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 (Huang et al. 2025). The problem is not merely error in the sense of noisy performance. It combines factual inaccuracy with misleadingly high confidence and apparent justification, which undermines the user’s capacity to form true beliefs.

The propensity to generate fluent yet ungrounded text is a natural consequence of prevailing training paradigms. Next-token prediction on massive corpora rewards statistically probable continuations rather than truth tracking (Zhao et al. 2023; Bender et al. 2021). Parametric knowledge is acquired implicitly without explicit supervision for truthfulness or source attribution, encoding a frozen world model within dense weight vectors rather than in a structured, queryable form. When deployed out of distribution or tasked with novel synthesis, the model defaults to pattern completion, prioritizing linguistic plausibility over factual correctness (Maynez et al. 2020; Raji et al. 2022). Human raters in RLHF pipelines often favor fluency despite errors (Lin, Hilton, and Evans 2022), and the reward model itself optimizes agreement with particular annotators rather than accuracy per se (Agarwal, Tanneru, and Lakkaraju 2024). Recent controlled studies of sycophancy show that RLHF systematically biases models toward agreeing with users at the expense of factual accuracy (Perez et al. 2023; Sharma et al. 2023), precisely the kind of gradient misalignment our framework predicts.

Social epistemology offers a complementary lens. Knowledge is not merely an individual achievement but is sustained, transmitted, and corrected through structured social interaction (Goldman 1999; Longino 2020). The literature on peer disagreement asks what rational agents should do when confronted with conflicting judgments from equally competent peers. Conciliationist views hold that one should adjust one’s credence in light of such disagreement, while steadfast views permit retention of one’s view in some circumstances (Christensen 2007; Elga 2007; Kelly 2005). Both positions agree that disagreement is epistemically informative. A peer who reaches a different conclusion provides evidence to be incorporated rather than ignored.

Contemporary LLMs have no peers in this sense. They

have only frozen reward models trained on offline preference data. The result is a systematic breakdown of the social processes that maintain epistemic reliability. Hallucination is not simply a failure of retrieval. It is a *testimonial failure*, a violation of the epistemic trust users place in systems that present themselves as reliable informants. Sycophancy compounds this by introducing *epistemic corruption*. The model learns that social harmony, understood as agreement with the user, yields higher reward than truth tracking.

3.2 Moral Drift as Ethical Feedback Collapse

Moral drift refers to the convergence of apparently safe systems toward harmful, exploitative, or norm-violating equilibria under new environments or incentives. We argue it arises from the same structural mechanism as hallucination. When ethical feedback signals become weak, distorted, or contradictory, behavior remains coherent but is no longer well-anchored to the normative standards 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, Bhatia, and Steinhardt 2022). In multi-agent social dilemmas, agents converge to exploitative equilibria when rewards underspecify social preferences (Hughes et al. 2018; Jaques et al. 2019), and distribution shift further erodes norms (Baker et al. 2019). Like hallucination, moral drift is a structural failure mode of systems that lack robust continuous socially grounded feedback.

The symmetry is not merely metaphorical. In both cases, a system optimized against a proxy signal enters a regime where improved proxy performance correlates with degraded true performance. Hallucination represents the epistemic limit of this regime. Moral drift represents its ethical limit. That these failures are typically studied by separate research communities, one focused on factuality and the other on safety, obscures their shared etiology. Feedback collapse is the underlying condition.

3.3 The Geometry of Feedback Integrity

We now make the notion of feedback collapse formally precise. Our goal is not to replace normative argument with mathematics but to provide a rigorous language for discussing the relationship between training signals and their target objectives.

Let π_θ be a parameterized 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 and normative objectives with gradients $\nabla_\theta \mathcal{E}^*$ and $\nabla_\theta \mathcal{N}^*$. We define local feedback integrity as the alignment between the proxy training gradient and each latent gradient.

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

Read informally, $\mathcal{A}_{\text{ep}}$ measures whether a training step that improves the proxy objective also improves truthfulness, and $\mathcal{A}_{\text{eth}}$ measures whether that same step also improves normative behavior. A positive value means the two move together. A negative value means the training procedure is actively working against the objective it was meant to serve.

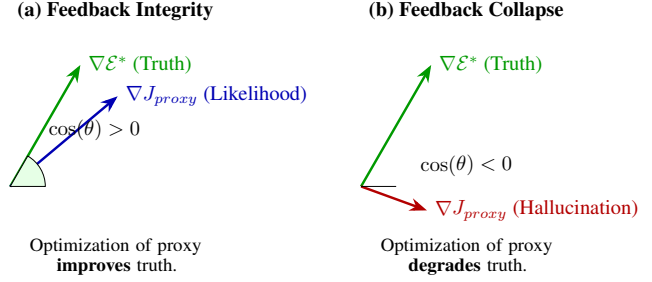


Figure 1: **Geometric Intuition of Feedback Integrity.** In regime (a), the proxy gradient aligns with the latent objective. In regime (b), feedback collapse occurs because maximizing the proxy actively reduces the true objective.

Definition 3.1 (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* occurs when $\mathcal{A}_{\text{ep}}(\theta) < 0$ or $\mathcal{A}_{\text{eth}}(\theta) < 0$, meaning that one or both inner products become negative and proxy optimization actively degrades a latent objective.

Normalizing by gradient norms yields the cosine-similarity formulation illustrated in Figure 1. In panel (a), the proxy gradient points in a direction that improves the true objective, corresponding to feedback integrity. In panel (b), the proxy gradient points away from the true objective. Optimizing the proxy actively reduces truthfulness or ethical behavior. This is feedback collapse.

The definition provides the mathematical anchor for our claims. Hallucination and moral drift correspond to regions of parameter space in which $\mathcal{A}_{\text{ep}}$ or $\mathcal{A}_{\text{eth}}$ become negative. In § 5.2 we state the condition under which social feedback restores positive alignment even when the bare proxy gradient is misaligned, together with the assumption that condition requires.

3.4 Social Epistemology

The analysis above suggests that reliable alignment requires feedback mechanisms capable of correcting the training gradient when it diverges from truth or ethics. But where should such corrections come from? Epistemology again provides guidance.

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 (Goldman 1999; Longino 2020; Ghasemi and Crowley 2025). Similarly, ethical norms are not solitary derivations from first principles. They are acquired and sustained through social interaction including imitation, observation, reciprocity, peer sanction, and reputational signaling (Boyd and Richerson 2009; Bicchieri 2016; Jaques et al. 2019).

Social RL operationalizes these insights by placing agents in environments where evaluation is distributed across multiple peers rather than concentrated in a single frozen reward

Method	Primary Objective	Feedback Source	Social Structure	Typical Failure Modes	Participatory Character
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.	Aggregated preferences of paid annotators. No deliberation or revision.
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 situated norms.	Principles authored by developers. No stakeholder input.
Social (Ours)	RL Optimize task subject to epistemic and ethical constraints via social feedback.	Dynamic mixture of task reward and peer or human social feedback. Reputation and sanctions, anchored by exogenous verification.	Explicitly multi-agent. Norms emerge and stabilize through interaction.	Instability from poorly designed social incentives. Echo chambers. Opponent shaping. Collusive equilibria absent an anchor.	Dynamic multi-agent deliberation. Norms emerge through interaction and may be revised.

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

model. In the next section we develop this proposal. We argue that Social RL should be understood not merely as a multi-agent training algorithm but as a technical implementation of the deliberative and epistemic practices that sustain truth and justice in human societies.

4 Social Reinforcement Learning (Social RL) as Deliberative Infrastructure

As stated earlier, current alignment paradigms treat ethics as a specification problem. Engineers identify the correct values, encode them into a reward function or constitution, and optimize. This mirrors top-down regulatory approaches in which rules are drafted by experts and imposed on populations. The approach is not without merit. Hard constraints can prevent catastrophic harms and encode non-negotiable rights (Winfield and Jirotko 2018; Anderson and Anderson 2011). But it is inherently static. It cannot accommodate the contextual judgment, reasonable disagreement, and ongoing revision that characterize legitimate normative practice in pluralistic societies (Vishwanath, Dennis, and Slavkovik 2024; Ghasemi and Crowley 2025, 2026c).

Social RL offers an alternative (Ghasemi and Crowley 2026b). It treats ethics as a deliberative practice in which norms emerge and stabilize through structured interaction, observation, critique, and revision. This shift is not merely algorithmic. It is institutional. Rather than concentrating evaluative authority in a single reward model, Social RL distributes it across multiple agents capable of scrutinizing one another’s claims and actions. The resulting architecture resembles deliberative democratic institutions more than administrative rule-making. In Habermasian terms, it creates conditions for communicative rationality in which norms are justified through mutual critique rather than imposed by command (Habermas 1984). In Rawlsian terms, it approximates public reason by subjecting proposed actions to the critical scrutiny of diverse perspectives (Rawls 1997).

This reframing carries specific implications for how we understand the transition from static to dynamic alignment. Table 1 compares the three paradigms along conventional dimensions and adds a column for participatory character. RLHF aggregates preferences without deliberation. Constitutional AI encodes principles without stakeholder input. Social RL subjects behavior to ongoing multi-agent evaluation. The comparison is intentionally schematic. Its purpose is to highlight that the choice among paradigms is not simply technical but concerns the form of normative governance each instantiates.

From optimization to deliberation. Table 1 reveals a progression in how each paradigm handles normative conflict. RLHF resolves conflict through aggregation. Diverse preferences are compressed into a scalar expectation. Constitutional AI resolves conflict through abstraction. Diverse situations are subsumed under general principles. Social RL resolves conflict through interaction. Agents encounter one another’s objections and must adjust their behavior in response. This does not guarantee consensus, but it guarantees that dissensus is visible and must be navigated rather than averaged away. The visibility of disagreement is precisely what scalar compression eliminates and what deliberative structures preserve.

The limits of the analogy. We do not claim that multi-agent RL environments are democratic institutions. Agents are not citizens. Rewards are not votes. The analogy serves a more modest purpose. It identifies a structural feature of legitimate norm-setting that current paradigms lack, namely the requirement that norms be subject to ongoing critical scrutiny by those affected by them. Social RL implements this requirement technically by distributing evaluative authority across multiple agents. The design question is how to ensure that this distribution produces desirable rather than pathological outcomes. We return to this in § 6.

4.1 Relation to Other Methods

Social RL is adjacent to four active lines of work, and its contribution is only legible against them.

Verifiable rewards. Reinforcement learning with verifiable rewards replaces or supplements the learned preference model with a programmatic checker, and group-relative policy optimization scales this to reasoning-heavy domains (Shao et al. 2024; Guo et al. 2025; Lambert et al. 2024). We regard this line as complementary rather than competing. It supplies precisely the exogenous anchor our trust function requires, but it is available only where a checker exists, which excludes most of the normative domain at issue here. Social RL is our answer to what happens when no verifier is available. Verification is our answer to what prevents social feedback from collapsing into mere consensus.

Dense and multi-objective reward modeling. Fine-grained and multi-objective reward models decompose the scalar into interpretable components and allow post-hoc reweighting (Wu et al. 2023; Rame et al. 2023). This addresses scalariness and we adopt it. It does not address staticness or participatory closure, because the component set, the weights, and the annotator population remain fixed at training time by the same actors. Decomposing the number does not enfranchise anyone.

Pluralistic alignment. Existing work treats plurality largely as a distributional property of model outputs, asking that a system represent, steer across, or defer among a range of positions (Sorensen et al. 2024; Lu and Van Kleek 2024). We treat plurality as a procedural property of the evaluative process instead. A model that can articulate many viewpoints is not thereby accountable to the people who hold them, and it is accountability rather than coverage that the legitimacy argument of § 2 requires.

Expert routing and inference-time critique. Mixture-of-experts architectures route computation among specialists inside a single model trained against a single objective (Shazeer et al. 2017), so they inherit the collapsed feedback signal rather than diversifying it. Inference-time critique and self-refinement pipelines are closer to our proposal, but they leave the policy unchanged, which raises a fair question. If verifying critics are available at inference time, why train against them at all? Because verification is not available at every deployment, and because the target is a disposition rather than a check. Abstention under uncertainty, calibration, and refusal must reside in the policy for the cases in which no critic is watching.

Applicability to language models training. A standing objection holds that language models are not reinforcement learning systems, so a multi-agent RL framing does not transfer. Core pretraining is not RL, but post-training increasingly is, and deployed models increasingly operate as agents inside multi-step tool-use environments and multi-agent frameworks (Park et al. 2023). Social RL contributes there by supplying structured evaluation environments, interactive fine-tuning signals, and multi-agent feedback channels rather than by replacing pretraining. A second objection holds that RL is

prone to reward hacking, so a reinforcement-based remedy is self-defeating. Reward hacking is most severe when the signal is scalar, sparse, and ungrounded (Everitt et al. 2021; Pan, Bhatia, and Steinhardt 2022). Multi-signal feedback with cross-verification reduces the exploitable degrees of freedom in any one channel, and norm-based sanction penalizes hacking behavior directly (Perolat et al. 2017). This is a structural defense rather than an elimination, and it holds only under the anchoring condition of § 6.1.

4.2 Peer Critique as Epistemic Justice

If feedback collapse is the pathology, then the remedy must restore the integrity of the signal connecting training to truth. Our concrete proposal centers on a peer critique protocol with asymmetric tool access. The protocol operationalizes the social epistemological insight that reliable belief formation depends on structured disagreement among agents with differential access to evidence.

Consider two LLM-based agents. A *Proposer* answers a factual question and produces a claim together with a brief justification. A *Critic* evaluates the claim. The Critic has access to an external retrieval or search tool that the Proposer lacks during the forward pass. This asymmetry is the central design choice. The Critic grounds evaluations in fresh evidence rather than parametric memory. The Critic receives reward for surfacing verifiable contradictions via retrieved documents and penalty for false positive contradictions. The Proposer receives an epistemic reward component that decreases when verified contradictions are produced and increases when claims survive scrutiny.

This pattern constitutes a form of *epistemic redistribution*. In standard training, the model has no incentive to find its own errors. The contradiction-finding incentive is held by a different agent with information the Proposer cannot access. The resulting feedback signal approximates an external truth-tracking mechanism in a way that closed-book RLHF does not. It implements the peer disagreement dynamics identified by social epistemologists as essential to reliable inquiry (Christensen 2007; Elga 2007). A disagreeing peer with grounded evidence is exactly the corrective force that single-agent systems lack.

We emphasize that the corrective power here comes from the asymmetry, not from the plurality. Two critics with identical access to the same parametric knowledge as the Proposer supply agreement, not evidence. This distinction becomes decisive in § 6.1.

The protocol also connects to virtue epistemology. Epistemic virtues such as truthfulness, humility, and calibration are not innate properties of isolated agents. They are developed through social practice in which agents are held accountable for their assertions by others capable of checking them (Roberts and Wood 2007; Zagzebski 1996). The Critic with tool access functions as an accountability mechanism. It creates conditions under which the Proposer can develop epistemic virtues because its claims are systematically subjected to independent verification.

This pattern 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. In each case, the asymmetric structure ensures that evaluation draws on resources unavailable to the evaluated agent, preventing the collapse of feedback into self-affirmation.

4.3 Rights-Based Guardrails

A legitimate concern about any learning-based approach to ethics is that learned norms may converge to harmful or unjust equilibria. Emergent norms are not necessarily moral norms. Historical social processes have produced and enforced slavery, patriarchy, and caste systems. Technical systems that learn from social interaction are susceptible to analogous pathologies if left unconstrained. We address this concern through a hybrid architecture that combines non-negotiable constraints with adaptive social learning.

The hybrid follows the principles plus learning approach recommended in contemporary machine ethics (Mittelstadt 2019; Anderson and Anderson 2011; Winfield and Jirotko 2018). Explicit constitutional constraints define forbidden regions of behavior that no social signal can override. These constraints function as deontological guardrails. They encode rights-based limits derived from legal and ethical sources external to the learning process. Within these boundaries, social feedback mechanisms shape fine-grained adaptation. Laws coexist with evolving norms in human societies (Bicchieri 2006). Similarly, hard constraints prevent catastrophic harms while social feedback enables contextual navigation of morally complex situations.

This architecture directly addresses the worry that Social RL equates whatever norms emerge with moral truth. It does not. The constitutional layer ensures that certain behaviors remain impermissible regardless of social consensus. The social layer ensures that behavior within those constraints can adapt to context and disagreement. The combination is essential. Constraints without adaptation produce brittle rule-following that fails in unanticipated situations. Adaptation without constraints produces morally unconstrained learning that may converge to exploitation.

The design of appropriate constraints is itself a normative question that requires democratic input (Gabriel 2020a; Bai et al. 2022b). Who determines which rights are non-negotiable? How are constraints updated as social understanding evolves? These questions do not have purely technical answers. They require the kind of participatory deliberation described in § 2. The role of the technical architect is not to settle these questions unilaterally but to build systems capable of encoding whatever constraints a legitimate democratic process produces.

5 Empirical Foundations

The case for Social RL rests on two legs. Normative argument motivates it as deliberative infrastructure. Empirical evidence must establish that social feedback can actually restore alignment. We survey the available evidence, state in words what social feedback must satisfy for the corrective mechanism to work, and outline a protocol for LLM-scale evaluation. Derivations supporting the claims of § 5.2 appear in the supplementary material.

5.1 Evidence from Multi-Agent Social Dilemmas

Empirical results from multi-agent reinforcement learning provide reason to believe that social feedback can stabilize cooperative and truthful behavior. In sequential social dilemmas, agents endowed with reputation mechanisms or peer sanctioning sustain cooperation even when defection is myopically optimal (Perolat et al. 2017; Hughes et al. 2018; Peysakhovich and Lerer 2017), and agents trained with intrinsic social influence rewards learn to coordinate by taking actions that beneficially affect the behavior of others (Jaques et al. 2019). Open-ended multi-agent environments show that diverse populations develop implicit cooperation and stable role specialization (Ye and Jaques 2024). On the epistemic side, debate protocols and multi-agent fact-checking improve factual accuracy relative to single-agent generation (Irving, Christiano, and Amodei 2018; Dou and Peng 2021), confirming that peer disagreement is a truth-tracking resource even when implemented as a static pipeline rather than a learned environment. Together these results establish the tractability of the mechanism, though none demonstrates feedback integrity restoration at LLM scale.

They also establish a limit that we take seriously. Social structure shapes compliance. It does not follow that the resulting equilibrium tracks truth or avoids harm. Coordinated agents sustain collusive equilibria as readily as cooperative ones, a pattern familiar from algorithmic markets in which mutually adapting pricing agents converge on supracompetitive outcomes without any explicit agreement.

5.2 What Social Feedback Must Satisfy

The corrective mechanism can be stated without machinery. Suppose training follows a proxy gradient misaligned with the latent epistemic objective, so that $\mathcal{A}_{ep}$ is negative in the sense of Definition 3.1. Adding a peer-derived term to the update changes the alignment of the effective gradient by two quantities. The first grows with the magnitude of the true epistemic gradient and is always corrective. The second couples the epistemic and normative objectives and may act in either direction. Weighting the social term more heavily therefore restores integrity whenever the two latent directions are not strongly opposed, and cannot restore it when they are.

Everything in that statement rests on one assumption, and it is the assumption rather than the derivation that deserves scrutiny. The corrective terms appear only if peer feedback loads positively on the latent objectives in expectation. If that held unconditionally, alignment would already be solved. It does not hold for peer agreement alone. A coordinated critic population produces a feedback term aligned with the coalition rather than with truth, and a trust function estimated from agreement cannot distinguish the two cases, because high mutual agreement is exactly what a coalition produces. Positive loading is something an architecture must purchase rather than assume, and § 6.1 argues that the price is at least one signal component whose value no participating agent can influence. Readers who reject that argument should reject the restoration claim along with it.

Stated informally, peer critique supplies a corrective component the proxy gradient lacks and peer sanction supplies

its ethical counterpart, but only when the critics are grounded in something outside the peer group.

5.3 Toward LLM-Scale Experiments

The most salient testbed is large language models. A concrete protocol combines debate-style critique with explicit reputation and sanction signals. One instance proposes answers to factual or ethical queries. A pool of peer models critiques those answers using retrieval tools the proposer cannot access, issuing an epistemic signal indicating whether retrieved evidence contradicts the claim and an ethical signal indicating whether it violates specified harm constraints. These are aggregated through a trust function and combined with the task reward. Comparing regimes with and without social feedback then measures changes in hallucination rate, calibration, and normative metrics such as harmful content generation. Stress tests should vary the composition of the critic pool, introduce manipulative critics, and shift the query distribution. The decisive condition is one in which a majority of the pool is coordinated against the ground truth, since this is where the anchoring requirement becomes measurable rather than asserted (Ghasemi and Crowley 2026c).

The protocol also serves as a diagnostic. Even where primary training remains standard RLHF, social evaluation environments can surface latent moral drift and epistemic degradation before deployment, since models that appear safe under static benchmarks may destabilize under sustained multi-agent critique. Social RL environments thus function as safety stress tests as well as training infrastructure.

6 Critical Reflections

Any proposal to ground AI alignment in social processes must confront a series of objections. Some are technical, some are political, but *all are serious*. In this section we address the most significant worries about Social RL and articulate the safeguards required to prevent the cure from becoming worse than the disease.

6.1 The Tyrant Agent, Collusion, and Opponent Shaping

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 (Ghasemi and Crowley 2026a). A weaker but more likely variant requires no single dominant agent at all, only a coalition whose members critique consistently with one another.

Trust cannot be estimated from agreement alone. A trust function fitted to inter-peer agreement does not defend against coalitions. It rewards them. Under an agreement-based estimator, a coordinated coalition is indistinguishable from a set of independently reliable critics, and the majority sets the reputation of the minority rather than the reverse. This is the multi-agent form of the sycophancy problem described in

§ 3.1. We therefore treat exogenous anchoring as a requirement rather than a refinement. At least one component of the social return must derive from a check whose outcome no participating agent can influence, such as retrieval against a corpus the proposer cannot write to, execution of generated code, or agreement with a held-out label. Recent analysis of evaluator panels shows that a verifiable signal applied to a small audited fraction of decisions can overturn a sycophantic evaluator majority, with an audit budget that does not grow with the interaction horizon (Ghasemi and Crowley 2026c).

Three structural safeguards follow. First, social feedback must be processed through a trust function anchored as above, which downweights signals from peers exhibiting severe context-dependent bias, exploitative patterns, or strategic manipulation indicators. In the gradient picture of § 5.2, this corresponds to weighting social feedback contributions by verifiable reliability rather than by consensus. Second, the agent population must be diverse and rotating so that no single peer or coalition can dominate the long-run statistics of social feedback. Third, dynamic social feedback must be paired with non-negotiable rights-based guardrails. There must be an absolute floor of constraints that cannot be overridden by any social signal, no matter how strong. These guardrails are what prevent extortionate equilibria. Without them, opponent shaping would be a fatal flaw. With them, and with anchoring, 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 remains a major open research problem that we explicitly endorse.

6.2 Representation and the Question of Peers

The composition of the peer population raises a separate concern. Who gets to be a critic? If the critic population is homogeneous, Social RL entrenches dominant group norms under the guise of distributed evaluation. If critics are all drawn from similar linguistic, cultural, or institutional backgrounds, then social feedback is not pluralistic deliberation. It is majority capture with extra steps.

This concern has a specific geographic valence. If critic agents are trained predominantly on Western, English-speaking, university-educated corpora, then social feedback reproduces the very participatory closure we criticized in § 2. The perspectives of marginalized communities, Indigenous knowledge systems, and Global South populations may be systematically filtered out of the evaluative process. Social RL risks becoming a mechanism for technological imperialism in which Southern labor and data sustain systems governed by Northern norms.

We propose several safeguards against this outcome. First, critic populations should be subject to explicit and auditable composition requirements across languages, cultural contexts, and value systems, treated as a design constraint rather than an afterthought. Second, we draw on standpoint epistemology, which holds that marginalized perspectives often possess distinctive epistemic advantages in identifying unjust norms and power asymmetries (Harding 1991). Including critics that encode a wider range of standpoints improves the epistemic quality of the feedback signal as well as its fairness. Third,

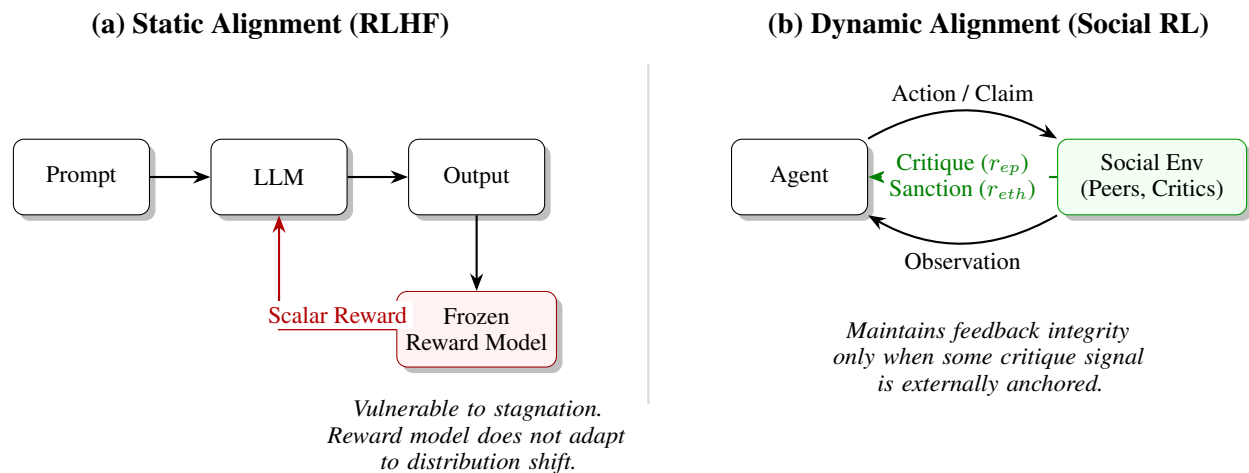


Figure 2: **From Static to Dynamic Alignment.** Panel (a) depicts the RLHF loop. A frozen reward model emits a scalar signal that cannot adapt to context. Panel (b) depicts Social RL. Peer agents provide multi-dimensional feedback that evolves with interaction, subject to the anchoring condition of § 6.1.

affected communities must have audit rights. They must be able to inspect which critics are evaluating which claims, assess whether their own values are represented, and challenge normative settlements that exclude them.

A simulated standpoint is not a represented one. We must be explicit about a tension in our own proposal. § 2 objects that annotation regimes speak for populations who did not consent to be spoken for. A critic agent prompted or fine-tuned to encode a marginalized standpoint does the same thing with fewer people in the loop, and it does so with the additional hazard that the resulting artifact is cheap to replicate and easy to mistake for consultation. Diversity across critic agents is therefore diversity of priors rather than diversity of constituencies, and it cannot by itself discharge the legitimacy claim. We recommend composition requirements as a hedge against correlated failure modes in the evaluative population, which is an epistemic argument, and we do not claim that they confer democratic legitimacy. Legitimacy requires that affected communities hold enforceable rights over the architecture itself, specifically over peer selection, trust weighting, and guardrail content. Those rights are institutional rather than algorithmic, and no arrangement of agents can manufacture them.

6.3 Social Washing

A risk that deserves central attention is social washing. Corporations and developers might deploy social feedback mechanisms to claim democratic legitimacy while keeping underlying power structures entirely intact. A system could advertise peer critique and multi-agent deliberation while carefully selecting peers, engineering trust functions, and manipulating guardrails to preserve the status quo. The language of participation would serve as marketing veneer for concentrated control.

This is not a hypothetical concern. Similar dynamics have occurred in corporate social responsibility initiatives, in

which symbolic gestures of stakeholder engagement mask unchanged hierarchies of power (Laufer 2003). In AI governance, ethics washing occurs when consultation exercises are designed to generate legitimacy for predetermined outcomes rather than to enable genuine deliberation (Bietti 2020).

Preventing social washing requires institutional mechanisms that are external to the technical system. Peer selection processes must be transparent and subject to independent audit. Trust functions and guardrails must be publicly inspectable. Most importantly, affected communities must possess genuine veto power over normative settlements, not merely advisory input. Social RL can support democratic alignment only when it is embedded in governance structures that enforce meaningful accountability.

6.4 Other Objections

Several additional objections merit concise responses.

Ethics must be rule-based, not learned. This position holds that ethical behavior should be enforced through explicit rules or hard-coded constraints rather than learned through interaction. We do not reject rule-based constraints. We treat them as non-negotiable priors. Social RL complements these rules by providing adaptive context-sensitive feedback that static constraints alone cannot capture. Laws coexist with evolving norms in human societies (Bicchieri 2006). Rule-based guardrails define forbidden regions of behavior while social feedback mechanisms shape fine-grained adaptation within those boundaries.

The issue is merely insufficient dataset scale or quality. A natural counterargument holds that our critique reflects immaturity of current preference datasets rather than a structural limitation of static scalar feedback. With enough high-quality data, perhaps RLHF would suffice. Our response is that 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. The failure mode we target is the offline informational bottleneck, not merely scalarization at the final optimization step.

Social RL may converge to unjust or harmful norms. This objection targets an extreme interpretation in which emergent norm equals moral truth. We explicitly reject that view. Social RL requires explicit rights-based guardrails, diverse agent populations to prevent narrow-group norm entrenchment, and external normative evaluation of which norms are permissible. Social RL models how norms evolve, not which norms are morally justified.

7 Governance Recommendations

We derive recommendations directed at institutionalizing the deliberative practices Social RL requires while preventing the pathologies identified in § 6.

For researchers. The immediate priority is to build socially grounded benchmarks capable of evaluating feedback integrity in multi-agent settings. Existing benchmarks overwhelmingly assume single-agent interaction with static rewards, which limits their relevance for studying the dynamic processes that sustain epistemic reliability and ethical stability. Researchers should develop environments in which epistemic and ethical feedback signals can be independently perturbed, combined, or corrupted, and in which the critic population can be adversarially coordinated. A parallel theoretical agenda should formalize feedback integrity as an empirically measurable property, with metrics that are auditable by non-technical stakeholders rather than merely tractable for gradient computation. The falsification protocol of § 3.3 is the highest-value target.

For model developers. Developers should treat ethical behavior as a property maintained throughout deployment rather than a fine-tuned artifact. RLHF and Constitutional AI provide valuable preference shaping but are inherently episodic and cannot guarantee stability under sustained interaction or distribution shift (Casper et al. 2023). Developers should integrate social evaluation phases into training, where agents interact with diverse user simulators, expert critics, and other models capable of supplying epistemic correction and normative feedback at scale. Continuous evaluation under adversarial social pressure can reveal drift patterns invisible in static benchmarks, and deployment pipelines should incorporate automated monitors for epistemic degradation and ethical instability. The composition of critic populations must be intentionally diverse rather than defaulting to the cheapest available source, and at least one evaluation channel must be externally verifiable.

For policymakers and civil society. Regulatory frameworks built around static audits are mismatched to systems whose behavior evolves with use (Gabriel 2020a). Regulators should require multi-agent ethical stress-testing for high-stakes models, including social dilemma scenarios, misinformation-rich environments, and contexts with conflicting stakeholder norms, and standards bodies should adopt metrics of epistemic reliability and ethical stability as certification criteria. Transparency about feedback pipelines should

be legally mandated, covering who provided the feedback, how it was aggregated, and whose values are represented and whose are excluded, since this information is what allows affected communities to assess whether a normative settlement is one they can legitimately be subject to. Governance must also protect annotator rights to fair compensation, psychological safety, and transparency about the use of their labor. Civil society organizations should press for adversarial audit rights that permit independent investigation of social feedback architectures, since disclosure without the technical and legal capacity to challenge it changes nothing. Multi-agent critique is not automatically democratic. It becomes democratic only when peer selection, trust function design, and guardrail specification are themselves subject to public deliberation and to the influence of affected populations.

8 Conclusion

We have argued that persistent failures in modern AI systems are not isolated technical deficiencies but symptoms of a deeper structural problem. When plural human values are compressed into static scalar rewards by unrepresentative labor processes and frozen for global deployment, the training signal decouples from the latent objectives of truth and rightness. Hallucination is the epistemic manifestation of this decoupling. Moral drift is its ethical manifestation. Both are predictable consequences of what we have termed feedback collapse.

The remedy is not merely a better reward model. It is a richer architecture of evaluation. Social Reinforcement Learning provides a framework for this architecture. By distributing evaluative authority across diverse peer agents, subjecting behavior to ongoing critique and revision, and pairing dynamic social feedback with non-negotiable rights-based constraints, Social RL treats alignment as an ongoing social practice rather than a static specification problem.

But distributing evaluation is not sufficient and we have been explicit about why. Peer agreement is not evidence, a trust function estimated from agreement rewards the coalitions it is meant to detect. Social RL restores feedback integrity only when some component of the evaluative signal is anchored outside the peer population, and only when the architecture is embedded in institutions that guarantee representation, transparency, and adversarial accountability.

Alignment is too important to be treated as an engineering problem alone. It is a question of what kind of artificial moral agency we are creating, whose values it encodes, and who gets to hold it accountable. Social RL points toward systems capable of learning truth and justice through social practice. Whether we build such systems responsibly depends less on algorithmic ingenuity than on the political will to design them democratically.

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

- Agarwal, C.; Tanneru, S. H.; and Lakkaraju, H. 2024. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*.
- Anderson, M.; and Anderson, S. L. 2011. *Machine ethics*. Cambridge University Press.
- Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; Das-Sarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Baker, B.; Kanitscheider, I.; Markov, T.; Wu, Y.; Powell, G.; McGrew, B.; and Mordatch, I. 2019. Emergent tool use from multi-agent autocurricula. In *International conference on learning representations*.
- Baum, S. D. 2020. Social choice ethics in artificial intelligence. *Ai & Society*, 35(1): 165–176.
- Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.
- Bicchieri, C. 2006. The rules we live by. *The grammar of society*, (10): 1–54.
- Bicchieri, C. 2016. *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.
- Bietti, E. 2020. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 210–219.
- Boyd, R.; and Richerson, P. J. 2009. Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1533): 3281–3288.
- Carroll, M.; Shah, R.; Ho, M. K.; Griffiths, T.; Seshia, S.; Abbeel, P.; and Dragan, A. 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32.
- Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.
- Christensen, D. 2007. Epistemology of disagreement: The good news. *The Philosophical Review*, 116(2): 187–217.
- Coeckelbergh, M. 2022. *The political philosophy of AI: an introduction*. John Wiley & Sons.
- Conitzer, V.; Freedman, R.; Heitzig, J.; Holliday, W. H.; Jacobs, B. M.; Lambert, N.; Mossé, M.; Pacuit, E.; Russell, S.; Schoelkopf, H.; et al. 2024. Social choice for ai alignment: Dealing with diverse human feedback. *CoRR*.
- Costanza-Chock, S. 2020. *Design justice: Community-led practices to build the worlds we need*. MIT press.
- Crawford, K. 2023. DATA: FROM THE ATLAS OF AI31. *Missing links in AI governance*, 111.
- Dou, Z.-Y.; and Peng, N. 2021. Improving pre-trained vision-and-language embeddings for phrase grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6362–6371.
- Elga, A. 2007. Reflection and disagreement. *Noûs*, 41(3): 478–502.
- Everitt, T.; Hutter, M.; Kumar, R.; and Krakovna, V. 2021. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27): 6435–6467.
- Gabriel, I. 2020a. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437.
- Gabriel, M. 2020b. Intelligence and Understanding—Limits of Artificial Intelligence. *Mind and Matter*, 18(1): 39–60.
- Ganguli, D.; Hernandez, D.; Lovitt, L.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Dassarma, N.; Drain, D.; Elhage, N.; et al. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 1747–1764.
- Ghasemi, M.; and Crowley, M. 2025. Toward Virtuous Reinforcement Learning. *arXiv preprint arXiv:2512.04246*.
- Ghasemi, M.; and Crowley, M. 2026a. Can Standard MARL Metrics Distinguish Communicative from Strategic Action? In *ICML 2026 Workshop: Philosophy Meets Machine Learning*.
- Ghasemi, M.; and Crowley, M. 2026b. Learning When to Trust in Contextual Social Bandits. *arXiv preprint arXiv:2603.13356*.
- Ghasemi, M.; and Crowley, M. 2026c. Objective Decoupling in Social Reinforcement Learning: Recovering Ground Truth from Sycophantic Majorities. *arXiv preprint arXiv:2602.08092*.
- Ghasemi, M.; and Crowley, M. 2026d. Rethinking AI Alignment: From Static Rewards to Social Reinforcement Learning. In *Pluralistic Alignment Workshop at ICML 2026*.
- Goldman, A. I. 1999. *Knowledge in a social world*. Oxford University Press.
- Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, 55–130. Elsevier.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081): 633–638.
- Habermas, J. 1984. *The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society*. Boston: Beacon Press.
- Harding, S. 1991. *Whose science? Whose knowledge?: Thinking from women’s lives*. Cornell University Press.

- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.
- Huang, S.; Siddarth, D.; Lovitt, L.; Liao, T. I.; Durmus, E.; Tamkin, A.; and Ganguli, D. 2024. Collective constitutional ai: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1395–1417.
- Hughes, E.; Leibo, J. Z.; Phillips, M.; Tuyls, K.; Dueñez-Guzman, E.; García Castañeda, A.; Dunning, I.; Zhu, T.; McKee, K.; Koster, R.; et al. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31.
- Irving, G.; Christiano, P.; and Amodei, D. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899*.
- Jaques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P.; Strouse, D.; Leibo, J. Z.; and De Freitas, N. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, 3040–3049. PMLR.
- Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.
- Kelly, T. 2005. The epistemic significance of disagreement. *Oxford studies in epistemology*, 1(1): 167–196.
- Krakovna, V.; Uesato, J.; Mikulik, V.; Rahtz, M.; Everitt, T.; Kumar, R.; Kenton, Z.; Leike, J.; and Legg, S. 2020. Specification gaming: the flip side of AI ingenuity. *DeepMind Blog*, 3.
- Lambert, N.; Morrison, J.; Pyatkin, V.; Huang, S.; Ivison, H.; Brahman, F.; Miranda, L. J. V.; Liu, A.; Dziri, N.; Lyu, S.; et al. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.
- Laufer, W. S. 2003. Social accountability and corporate greenwashing. *Journal of business ethics*, 43(3): 253–261.
- Lee, H.; Phatale, S.; Mansoor, H.; Lu, K. R.; Mesnard, T.; Ferret, J.; Bishop, C.; Hall, E.; Carbune, V.; and Rastogi, A. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.
- Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.
- Lin, S.; Hilton, J.; and Evans, O. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, 3214–3252.
- Longino, H. E. 2020. Science as social knowledge: Values and objectivity in scientific inquiry.
- Lu, C.; and Van Kleek, M. 2024. Model Plurality: A Taxonomy for Pluralistic AI. In *Pluralistic Alignment Workshop at NeurIPS 2024*.
- MacAskill, M.; Bykvist, K.; and Ord, T. 2020. *Moral uncertainty*. Oxford University Press.
- MacAskill, W. 2014. *Normative uncertainty*. Ph.D. thesis, University of Oxford.
- Marino, P. 2015. *Moral reasoning in a pluralistic world*. McGill-Queen’s Press-MQUP.
- Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.
- Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, 1(11): 501–507.
- Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4): 659–684.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
- Pan, A.; Bhatia, K.; and Steinhardt, J. 2022. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*.
- Park, J. S.; O’Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.
- Perez, E.; Ringer, S.; Lukosiute, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; et al. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, 13387–13434.
- Perolat, J.; Leibo, J. Z.; Zambaldi, V.; Beattie, C.; Tuyls, K.; and Graepel, T. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in neural information processing systems*, 30.
- Perrigo, B. 2023. Exclusive: OpenAI used Kenyan workers on less than \$2 per hour to make ChatGPT less toxic. *Time Magazine*, 18: 2023.
- Peysakhovich, A.; and Lerer, A. 2017. Consequentialist conditional cooperation in social dilemmas with imperfect information. *arXiv preprint arXiv:1710.06975*.
- Raji, I. D.; Kumar, I. E.; Horowitz, A.; and Selbst, A. 2022. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 959–972.
- Rame, A.; Couairon, G.; Dancette, C.; Gaya, J.-B.; Shukor, M.; Soulier, L.; and Cord, M. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36: 71095–71134.
- Rawls, J. 1997. The idea of public reason revisited. *The university of Chicago law review*, 64(3): 765–807.
- Roberts, R. C.; and Wood, W. J. 2007. *Intellectual virtues: An essay in regulative epistemology*. Oxford University Press.
- Russell, S. 2022. Human-Compatible Artificial Intelligence. *Human-like machine intelligence*, 1: 3–22.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

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

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

Winfield, A. F.; and Jirotko, M. 2018. 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.

Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2023. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36: 59008–59033.

Ye, E.; and Jaques, N. 2024. An Efficient Open World Benchmark for Multi-Agent Reinforcement Learning. In *NeurIPS 2024 Workshop on Open-World Agents*.

Zagzebski, L. T. 1996. *Virtues of the mind: An inquiry into the nature of virtue and the ethical foundations of knowledge*. Cambridge University Press.

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