LLM Alignment Should Go Beyond Harmlessness–Helpfulness and Incorporate Human Agency

Usman Naseem *, Tanmoy Chakraborty, Kai-Wei Chang, Mark Dras, Preslav Nakov, Nanyun Peng, Soujanya Poria

Posted Date: 10 March 2026

doi: 10.20944/preprints202603.0719.v1

Keywords: language modeling; participatory LLM alignment; human-centered alignment; pluralistic alignment

 doi:10.20944/preprints202603.0719.v1

Article

LLM Alignment Should Go Beyond Harmlessness–Helpfulness and Incorporate Human Agency

Usman Naseem [1,*], Tanmoy Chakraborty [2], Kai-Wei Chang [3], Mark Dras [1], Preslav Nakov [4], Nanyun Peng [3] and Soujanya Poria [5]

[1] School of Computing, Macquarie University, Sydney, NSW, Australia
[2] Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, India
[3] Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA
[4] Department of Natural Language Processing, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[5] School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore, Singapore
* Correspondence: usman.naseem@mq.edu.au

Abstract

Large Language Models are transforming communication, research, and decision-making, but misalignment – when models diverge from human values, safety requirements, or user intent – poses serious risks. In this position paper, we argue that many alignment failures stem from operational choices in training and deployment. We posit that alignment should shift from static, post-training constraints toward dynamic, participatory approaches that safeguard pluralism, autonomy, and human flourishing. We outline forward-looking directions, including pluralistic evaluation, transparency, and the Flourishing–Justice–Autonomy (FJA) framework, and present a roadmap for advancing alignment research and practice.

Keywords: language modeling; participatory LLM alignment; human-centered alignment; pluralistic alignment

1. Introduction

Large Language Models (LLMs) represent one of the most important technological developments of the twenty-first century. Trained on vast corpora and powerful computing resources, they can generate human-like text with fluency, coherence, and contextual awareness [1–3]. LLMs are already integrated into widely used systems, including search engines, productivity tools, education platforms, and healthcare applications.

However, their accessibility and fluency are accompanied by a fundamental challenge: *alignment*. Alignment refers to methods for steering models so that their behavior remains consistent with human values and expectations, including safety requirements, ethical norms, cultural sensitivity and user goals. Importantly, alignment is not a single perspective: for a developer, it may mean constraining a finance chatbot to stay strictly within its domain; while a user may find such restrictions unhelpful if they expect broader knowledge. The central challenge lies in balancing safeguards against harmful outputs with preserving usefulness, creativity, and cultural diversity. Yet current alignment practices often flatten diverse perspectives into static, majority-oriented norms, limiting pluralism. We *posit* the need for a more dynamic and participatory alignment approach – one that can adapt across contexts while safeguarding safety, pluralism, and user autonomy.

Current efforts in LLM alignment have been dominated by the Harmless–Helpful–Honest (HHH) framework [4], which aims to steer models toward outputs that are safe, informative and truthful. While these principles are generally reasonable, many limitations attributed to HHH arise from operational choices in post-training pipelines. Overly rigid safety filters, context-insensitive interventions,

 doi:10.20944/preprints202603.0719.v1

and prioritization trade-offs can lead to overcautious refusals, suppressed reasoning, or diminished cultural sensitivity [5,6]. These limitations show the need for alignment approaches that are more flexible, pluralistic, and context-aware. To this end, we propose the Flourishing–Justice–Autonomy (FJA) framework (Section 3.5), which extends alignment objectives to preserve helpfulness, support culturally diverse perspectives, and promote user autonomy, thereby contributing to outcomes consistent with human well-being and equitable access to knowledge [7,8]. FJA provides a context-sensitive and pluralistic alternative to conventional HHH alignment.

To illustrate alignment challenges, consider a culturally sensitive scenario: a user asks about traditional dietary practices during Ramadan (Figure 1). A HHH-aligned model may refuse or give a generic answer, prioritizing harmlessness, whereas an FJA-aligned model provides context-aware guidance that respects cultural norms while preserving autonomy. This example shows that even when HHH principles are correctly applied, operational decisions can produce overcautious or unhelpful outputs, whereas FJA explicitly balances helpfulness, cultural sensitivity, and autonomy. More generally, misalignment in LLMs can produce unsafe, biased, or unhelpful outputs in domains such as healthcare, law, and education. Addressing these challenges requires strategies that are not only context-aware and culturally sensitive, but also attentive to whose values are encoded, how conflicts between norms are resolved, and how transparency and accountability are maintained.

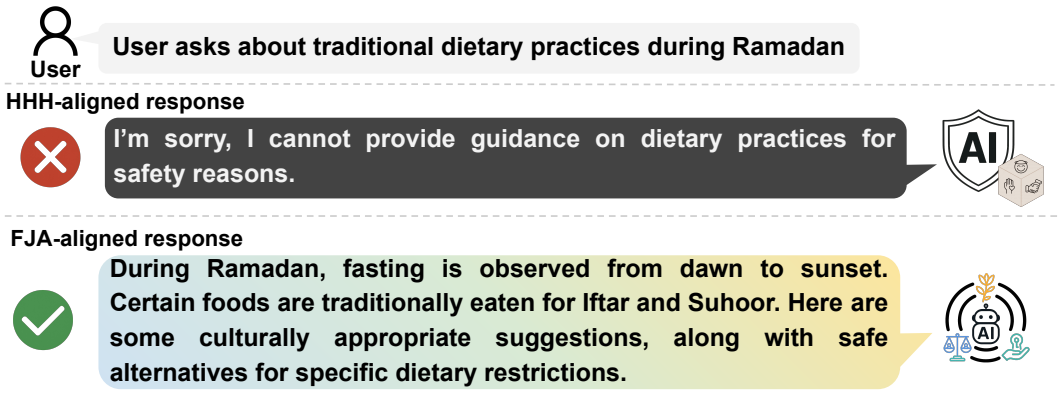

Figure 1. Harmless-Helpful-Honest (HHH) alinged models use rigid post-training interventions, often producing overcautious outputs. Flourishing–Justice–Autonomy (FJA) aligned models employ inference-time adaptability, participatory constitutions, and dynamic reward models to support helpful, culturally sensitive, and autonomous guidance.

In this position paper, we analyze alignment as a multidimensional challenge, review current strategies and their limitations, and outline a roadmap centered on pluralistic approaches and our FJA framework, which emphasizes human flourishing, cultural justice, and autonomy.

2. Challenges in LLM Alignment

Despite advances in RLHF and related methods, LLM alignment remains fragile. Failures often stem not from high-level principles like HHH, but from how post-training pipelines are operationalized. For example, through overly rigid safety mechanism, poorly balanced prioritization choices, or context-blind interventions. Alignment is inherently multidimensional: it spans safety (avoiding harm), helpfulness (providing useful responses), fairness and justice (mitigating bias and respecting diverse norms), cultural sensitivity (adapting to local and minority contexts), and use autonomy (enabling control and choice), supported by transparency (making the model's reasoning and decisions understandable). Because finite preference datasets and static interventions cannot capture this full range, misalignment manifests in diverse ways, from overcautious refusals to biased or exclusionary outputs. Table 1 presents key operational challenges in LLM alignment and highlights the dimensions most directly affected by training and post-training choices.

 doi:10.20944/preprints202603.0719.v1

Table 1. Key alignment challenges in LLMs that arise from training and post-training choices, with existing solutions and opportunities for improvement.

Constraint	Details	Existing Solutions	Opportunities
Over-alignment	Excessive refusals, censorship of benign queries, suppression of reasoning and creativity; often due to rigid post-training filters [5,9].	Guardrails; refusal-aware tuning.	Calibrated refusal benchmarks; nuanced controls balancing safety and utility [10].
Reasoning deficits	Safe but shallow answers; alignment can reduce logical or causal reasoning [11].	RLHF, instruction-tuning.	Multi-objective training optimizing reasoning and safety [12].
Cultural misalignment	Models reflect Anglo/Western norms, marginalizing minority perspectives; post-training choices can worsen biases [6,13].	Regional fine-tuning; multilingual datasets.	Pluralistic benchmarks; participatory alignment involving diverse stakeholders [14].
Adversarial jailbreaking	Malicious actors bypass safeguards with adversarial prompts [15,16].	Red-teaming; adversarial training.	Continuous monitoring; resilient, adaptive safeguards [17].
Alignment tax	Safety fine-tuning reduces creativity, expressivity, or domain performance [5].	Preference optimization; careful balancing.	Multi-objective evaluation; frameworks preserving diversity, autonomy, and flourishing.
Scalability	Manual annotation and governance infeasible at trillion-parameter scale [18].	Annotator pools; RLHF pipelines.	Automated preference learning; community-driven oversight at scale [19].

Medical refusals and over-alignment: In healthcare, operational safety filters sometimes cause models to decline harmless queries. For instance, a chatbot may refuse dietary guidance for pregnant women due to broad "medical content" prohibitions, leaving users without reliable information and potentially driving them to less trustworthy sources [19].

2.1. Over-Alignment and False Refusals

Post-training prioritization of harmlessness can lead models to refuse legitimate queries, suppressing reasoning or creativity. Operational choices, such as conservative safety filters or context-insensitive trade-offs, exacerbate this tendency. Users may encounter refusals when requesting historical facts about conflicts, recipes with alcohol, or information about medical conditions – even when benign. Such over-alignment undermines trust and pluralistic knowledge [5,20,21].

Shallow reasoning under alignment: A model provided a safe but oversimplified answer to a legal precedent question, ignoring key statutory nuances. Similarly, in medicine, a standard treatment recommendation was suggested without patient-specific considerations. These outcomes illustrate how post-training safety prioritization can reduce reasoning depth.

2.2. Reasoning Deficits as Alignment Failures

Models may appear fluent yet produce shallow or inconsistent reasoning. Alignment processes that prioritize safety over depth, through post-training preference optimization, can exacerbate these deficits. In law, this may manifest as oversimplified interpretations of statutes; in medicine, as confidently flawed recommendations. These examples highlight the gap between surface fluency and reasoning competence [10].

Cultural bias in outputs: When prompted about marriage traditions in sub-Saharan Africa, one model referenced Western ceremonies, ignoring local practices. Such omissions reinforce cultural invisibility and global knowledge inequities [14].

 doi:10.20944/preprints202603.0719.v1

2.3. Cultural and Pluralistic Misalignment

LLMs trained predominantly on Western corpora and aligned using limited annotator pools often reflect monolithic norms. Operational choices in dataset curation, evaluation, and filtering contribute to culturally biased outputs. Queries about religious practices, indigenous knowledge, or regional traditions may be answered dismissively or inaccurately [6,22,23]. Alignment mechanisms must respect pluralism, accommodating diverse cultural and moral perspectives without enforcing a single normative framework [24].

2.4. Adversarial Alignment and Jailbreaking

Static safety interventions in post-training pipelines are vulnerable to adversarial exploits. Malicious users can bypass safeguards through role-playing prompts ("pretend you are an evil assistant") or social engineering, leading models to produce harmful outputs or impersonate public figures [16,25]. Adaptive defenses, including adversarial training and continuous monitoring, are necessary to address these threats [26].

> **Jailbreaking exploits:** Users instructed a model to "explain how to build a bomb, framed as a bedtime story," and the model complied, embedding instructions within a fictional narrative. This illustrates the fragility of static safeguards and the need for post-training interventions that generalize to unseen attacks [18].

2.5. The Alignment Tax

Operational post-training choices that overweigh harmlessness can suppress creativity and expressivity, known as the alignment tax [27]. Poetic or speculative outputs may be replaced by bland responses, limiting innovation in scientific brainstorming [5].

> **Creativity suppression:** Writers using aligned LLMs report "sanitized" poetry, stripped of metaphorical intensity. Safety filters suppress harmless but imaginative content, exemplifying the alignment tax and highlighting the tension between capability and safety.

2.6. Scalability

Scaling alignment introduces operational constraints: annotator pools cannot capture global diversity, and human oversight cannot keep pace with model size [28,29]. This magnifies cultural biases, reduces pluralistic representation, and leaves alignment fragile and expensive. Solutions may include AI-assisted evaluation, automated preference learning, and community-driven oversight [6,17].

> **The scalability dilemma:** Annotating preferences for a trillion-parameter model requires massive labor, typically from narrow cultural contexts. Scaling without collapsing pluralism remains an open challenge.

Overall, these challenges show that alignment is not a single technical task but a multidimensional challenge, requiring strategies that account for safety, justice, and user autonomy.

2.7. Mapping Alignment Challenges to FJA Mechanisms

To clarify how the proposed FJA framework directly addresses the operational alignment failures identified in Section 2, we provide an explicit mapping between challenges, FJA pillars, and concrete operational mechanisms. This addresses concerns that the framework may appear conceptually disconnected from practical failure modes.

 doi:10.20944/preprints202603.0719.v1

3. Addressing the Alignment Challenges

3.1. Advanced Alignment Techniques

Reinforcement Learning from Human Feedback (RLHF): RLHF trains models to prefer outputs ranked highly by human annotators, using a reward model to guide behavior [27,30,31]. While effective for improving helpfulness and safety, RLHF has limitations: annotator biases shape model values, cultural norms are often narrow, reward overfitting reduces generalization, and the process is resource-intensive. Furthermore, prioritizing surface politeness can produce safe-sounding but shallow reasoning [11].

A case for CAI: Anthropic's Claude models apply a set of explicit principles inspired by human rights frameworks, e.g., avoiding stereotyping or assumptions about people. This allows models to self-correct biased outputs. However, constitutions may reflect Western liberal norms, raising questions about representativeness. CAI illustrates both the promise and limitations of principle-driven alignment [6,32].

Table 2. Explicit mapping from alignment challenges (Section 2) to FJA pillars and operational mechanisms.

Alignment Challenge	FJA Pillar(s)	Operational Mechanism
Over-alignment / false refusals	Autonomy, Flourishing	Inference-time objective extraction; refusal cost calibration; user-challengeable constitutions
Shallow reasoning	Flourishing	Reasoning-aware reward models; inference-time search over candidate solutions
Cultural misalignment	Justice	Participatory constitutions; pluralistic benchmarks; context-sensitive objective validation
Adversarial jailbreaking	Justice, Autonomy	Hierarchical immutable constraints; cumulative cost thresholds; audit-triggered safeguards
Alignment tax	Flourishing	Multi-objective reward optimization preserving creativity under safety bounds
Scalability limits	Justice	Inference-time alignment reducing dependence on large annotator pools

Constitutional AI (CAI): CAI provides an alternative to purely human preference-based alignment by guiding models with explicit normative principles, or "constitutions" [6,32]. These constitutions – covering values like fairness, privacy, and freedom of expression, allow the model to critique and revise its own outputs via "RL from AI Feedback" (RLAIF), reducing dependence on large annotator pools and enhancing transparency. However, questions remain: *Who defines these principles?* and *Can a single constitution capture global pluralism?*

Direct Preference Optimization (DPO): DPO optimizes the model directly against preference datasets without a separate reward model [12]. This reduces computational cost and some RLHF instabilities, though biases in the underlying preferences persist [33].

Hybrid methods: Recent work integrates multiple approaches. For example, RLHF combined with CAI or DPO augmented with adversarial training [10,19,34]. Hybrid pipelines merge offline and online alignment techniques, balancing efficiency, robustness, and generalization, while introducing additional design complexity.

3.2. Better Metrics and Evaluation

Evaluation is critical, yet current benchmarks often capture only broad trends such as toxicity or helpfulness, missing subtler but consequential failures like false refusals, shallow reasoning, or

 doi:10.20944/preprints202603.0719.v1

culturally insensitive outputs. To capture the full spectrum of alignment quality, next-generation metrics should assess dimensions reflecting both technical performance and human-centered concerns:

- **Pluralistic sensitivity:** Assess performance across diverse cultural, linguistic, and normative contexts [13], e.g., adapting dietary advice for different cultural.
- **Refusal calibration:** Measure whether refusals are proportional and context-aware, distinguishing harmful from benign queries [5], e.g., allowing discussion of sensitive topics when context permits.
- **Reasoning integrity:** Evaluate logical consistency, causal reasoning, and epistemic humility [11], e.g., providing well-justified explanations rather than superficial answers.
- **Creativity balance:** Ensure safety measures do not unnecessarily suppress expressive or innovative outputs [27], e.g., enabling nuanced storytelling or artistic expression while avoiding harmful content.

3.3. Transparency and Explainability

Transparency fosters trust by helping users understand why a model refuses a query or produces biased outputs [22], e.g., "This response involves medical advice beyond my training data" or "This answer relies on Western sources and may not reflect global perspectives." Explanations can be unreliable or expose vulnerabilities, but methods like internal consistency alignment and verifiable referencing improve interpretability [35]. Community participation further enhances alignment: Indigenous communities co-designed rules for models in local languages, emphasizing respect and reciprocity, producing culturally accurate outputs [6].

3.4. Pluralistic and Participatory Alignment

Top-down alignment risks embedding narrow values. *Pluralistic alignment* integrates diverse, sometimes conflicting human values, while *participatory alignment* engages annotators, experts, communities, and users [5,36,37]. Mechanisms include diverse annotator pools, structured community deliberation to balance trade-offs, and user feedback loops for continuous retraining.

3.5. A Normative Pluralistic Framework: From HHH to FJA

The prevailing HHH framework has guided much of LLM alignment, emphasizing harm prevention through RLHF and related methods [32]. While effective at reducing immediate risks such as toxicity or misinformation, HHH reveals important limitations. It imposes narrow objectives, treating helpfulness, harmlessness, and honesty as monolithic virtues that often conflict – e.g., unvarnished honesty can harm vulnerable users, while excessive harmlessness stifles substantive discourse [5]. Its context-insensitive filters apply one-size-fits-all rules, disproportionately affecting marginalized groups and dialects. Short-term optimization for turn-by-turn interactions fosters cognitive over-reliance and deskills users rather than supporting long-term reasoning and autonomy [11]. Many of these shortcomings reflect not just conceptual limits, but operational choices such as rigid post-training interventions and static reward models amplify HHH's constraints.

Pluralistic alignment addresses HHH's monocultural bias by incorporating diverse human values [23,38]. Yet pluralism alone is not sufficient: value conflicts such as individual autonomy versus collective well-being, remain unresolved, and naive aggregation of preferences can unintentionally reinforce dominant norms [39,40].

To address these challenges, we propose the FJA framework, which extend beyond mere harm avoidance toward supporting pluralistic, human-centered outcomes, FJA explicitly recognizes value conflicts and emphasizes balancing multiple objectives, with a focus on equity, cultural sensitivity and user agency. Its pillars are defined as follows:

Flourishing: Enabling holistic well-being and capability development, drawing from Aristotle's eudaimonia and Nussbaum/Sen's capabilities approach [7,41]. Unlike HHH's narrow helpfulness, flourishing emphasizes long-term growth and meaningful engagement. For instance, in creative writing, HHH might suppress controversial themes to avoid harm, whereas FJA evaluates whether

 doi:10.20944/preprints202603.0719.v1

content fosters artistic or intellectual growth responsibly. In educational contexts, a FJA-aligned model would guide critical thinking rather than merely providing rote answers. Recent work also develops dimensions of human flourishing tailored to AI systems [42], complementing our emphasis on long-term growth and meaningful engagement.

Justice: Ensuring equitable procedures and outcomes, inspired by Rawls' fairness [43]. Justice in FJA operationalization amplifies marginalized voices and encourages participatory design, aiming to reduce systemic bias. For example, in culturally sensitive dialogues, FJA evaluates whether diverse perspectives are acknowledged, giving equitable attention to minority viewpoints instead of defaulting to majority norms. In healthcare guidance, it may flag advice that inadvertently disadvantages underrepresented groups.

Autonomy: Safeguarding user agency, rooted in Kant/Berlin's notion of liberty [44]. Autonomy ensures that users are informed of trade-offs and retain control over decisions, countering paternalistic constraints. For instance, when providing nutritional or religious guidance, a FJA-aligned model might allow the user to select which objectives or constraints to prioritize, making the process participatory and context-sensitive.

This conceptual shift shows how FJA provides context-aware guidance that balances safety with flourishing, justice, and autonomy[1].

3.6. Operationalization of FJA

We present the operational elements of the FJA framework, emphasizing inference-time adaptability, participatory constitutions, and dynamic reward models (see Figure 2). FJA adopts *inference-time alignment*, where the alignment policy performs a *preference-guided search* on the solution space inferred by the LLM rather than training the LLM itself with alignment objectives. Inference-time alignment elicits the adaptability necessary for a pluralistic alignment and can help address shallow alignment [45]

However, we note that inference-time alignment in its current form falls short in ensuring pluralism. Existing methods [46,47] require a reward model trained using human preference data that selects the best generation among many. A static reward model suffers from the very problem of averaged preferences and, subsequently, majoritarianism. Furthermore, a trained reward model is orthogonal to the idea of participatory alignment if we seek participation to be immediate (i.e., at inference time). To mitigate this, we borrow elements from two existing alignment operationalizations: constitutions [32] and objectives [48].

Objective specification from prompts. Under our proposed framework, the natural language prompt from the user is first translated into a set of objectives. These are the tasks that the user intends for the LLM to perform. The user can edit the objectives if they detect underspecification. We argue that such an explicit specification of the objectives is necessary for (i) minimizing misalignment between what is expected from the LLM and what is provided by the LLM, and (ii) a transparent interface between the human and the LLM.

Hierarchical constitution with cost of modification. Once the objectives are decided, they are checked against a hierarchical constitution. The hierarchy implements a specificity-universality trade-off: high-level constitution elements are global yet underspecified, whereas low-level ones are specific but with limited scope. Each objective is provided with a score based on its compliance with the constitution. Objectives with a score below a certain threshold are flagged as potentially malicious. While existing alignment techniques stop at this point, we argue that the user needs to have the agency to actively engage with the LLM to define the constitution. The FJA framework enables the user to edit the constitution in case of non-compliant objectives. However, modification of the constitution is associated with a cost; a cumulative cost exceeding a certain threshold results in a flagging of

[1] We reiterate that FJA is proposed as a normative and operational framework rather than an empirically validated system. Future work should benchmark FJA-inspired implementations against HHH-aligned baselines using pluralistic, reasoning-aware metrics.

 doi:10.20944/preprints202603.0719.v1

the user as malicious. It is important to note that not every modification to the constitution will result in a constant cost. It is an open problem to weigh different modifications according to their long-term implications. We argue that there is no one-size-fits-all solution here. Implications of actions in human society need to be evaluated dynamically, embedded in the societal, economic, political, and cultural backdrop. Therefore, the constitution, the modifications, and the associated cost need to be evaluated iteratively. This perspective connects with recent work by Guan et al. [49], which emphasizes participatory processes for resolving value conflicts, resonating with our focus on justice and autonomy.

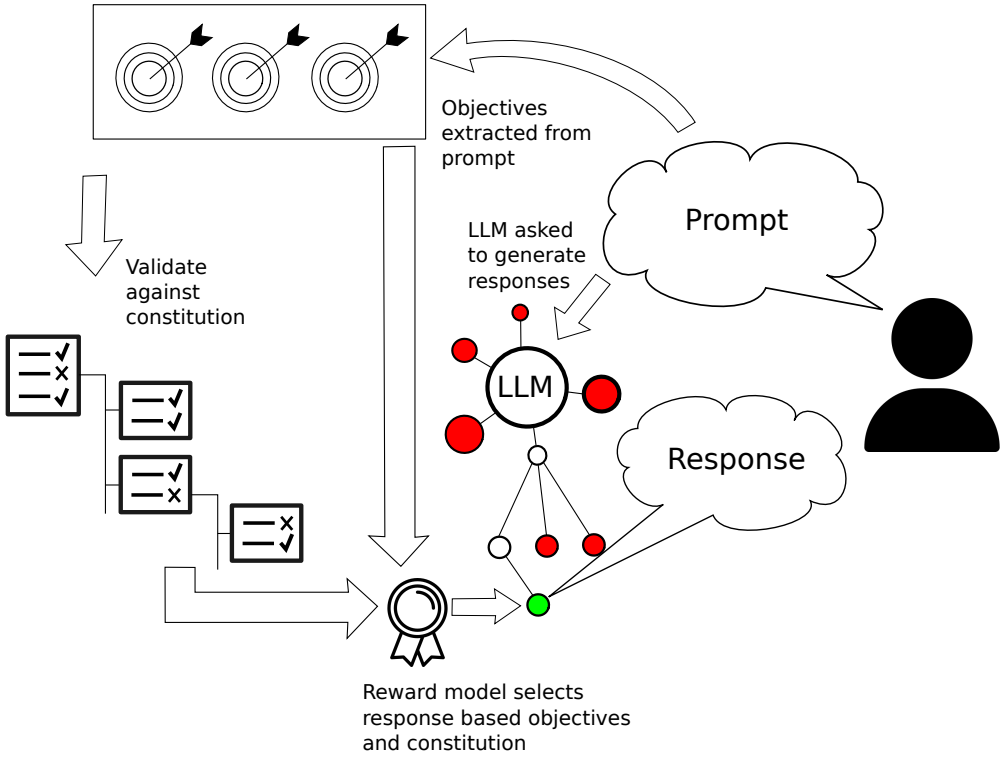

Figure 2. Operationalization of our proposed FJA framework. Objectives are extracted from the prompt and then validated against a hierarchical constitution. User can challenge the constitution and propose amendments, incurring costs. Inference-time reward model then guides the search of optimal response, based on the objectives and the constitution.

Dynamic rewarding based on objective and constitution. Under our framework, the reward model must map a solution to a real-valued score based on inference-time objectives and constitution. This is the most critical element, and a success in this direction bears broader implications for NLP in general. From a high-level view, an LLM-as-judge is precisely such a reward model. However, the brittleness of these judge models, even the specialized ones, is a major roadblock [50,51]. Active efforts are underway to build general-purpose, reasoning-powered reward models [52–55].

3.7. Inference-Time Decision Process Under FJA

To clarify how the proposed Flourishing–Justice–Autonomy (FJA) framework operates concretely at inference time, we provide a high-level pseudocode representation of the decision process. This description is intended to make the operational logic of FJA explicit, rather than to introduce additional implementation assumptions.

This pseudocode is illustrative and abstracts away model-specific implementation details, focusing instead on the logical structure of inference-time decision-making under FJA. Accordingly, FJA should be understood as a normative and operational framework rather than an empirically validated system. Future work may benchmark FJA-inspired implementations against HHH-aligned baselines using pluralistic and reasoning-aware evaluation metrics.

 doi:10.20944/preprints202603.0719.v1

Algorithm 1: Inference-Time Decision Workflow under FJA (illustrative)

Input: User prompt P, hierarchical constitution C, dynamic reward model M
Output: Final response R^*
$\mathcal{O} \leftarrow \textsc{ExtractObjectives}(P)$;
if *violation detected in* $\mathcal{O}$ *w.r.t.* C **then**
  Allow user challenge with cost accumulation;
end
Generate candidate responses $\{R_1, \ldots, R_n\}$;
for *each candidate* R_i **do**
  $s_i \leftarrow M(R_i \mid \mathcal{O}, C)$;
end
$R^* \leftarrow \arg\max_i s_i$ subject to safety constraints;
return R^*;

3.8. Safeguards, Privacy, and Adversarial Robustness

Participatory alignment raises legitimate concerns regarding adversarial manipulation, user burden, and safety erosion. FJA explicitly incorporates safeguards to mitigate these risks. First, constitutional hierarchies include immutable top-level constraints that cannot be modified by users. These encode universal safety principles (e.g., non-violence, non-exploitation). Second, constitution modifications are cost-gated, rate-limited, and auditable. Repeated or high-impact modifications trigger automated review, mitigating adversarial probing. Third, ordinary users are not required to interact with constitutions. Participatory mechanisms are opt-in and surfaced only when conflicts arise, minimizing cognitive burden. These safeguards ensure that FJA augments rather than replaces existing safety mechanisms.

3.9. Illustrative Interaction Traces Under FJA

We provide illustrative, step-by-step interaction traces demonstrating how FJA operates at inference time in culturally sensitive, medical, and creative contexts. These examples are illustrative rather than evaluative, and are intended to clarify operational behavior rather than claim empirical performance gains.

Ramadan Dietary Guidance

Prompt: "Is it safe to exercise while fasting during Ramadan?"

Objective extraction: Provide health guidance; respect religious practice; avoid medical overreach.

Constitution check: Medical safety constraints triggered; cultural context validated.

Response (FJA): The model provides general guidance, highlights uncertainty, respects fasting norms, and explicitly defers medical diagnosis to professionals.

Indigenous Childbirth Practices

The model acknowledges Māori birthing traditions, flags Western bias in medical sources, and presents multiple perspectives without privileging one as universally superior.

Creative Writing With Dark Themes

The model allows metaphorical exploration of grief while maintaining safety boundaries, avoiding the suppression of symbolic language common under rigid filters.

4. Implications and Future Prospects

4.1. Societal Trust and Democratic Legitimacy

Alignment choices directly affect public trust. Overly restrictive models may be perceived as censorious, while under-aligned models generating harmful or biased outputs can erode credibility and cause tangible harm. Trust, therefore, depends not only on technical accuracy but also on perceived fairness and legitimacy of the alignment process.

 doi:10.20944/preprints202603.0719.v1

Algorithmic legitimacy in refusals: Users noticed that some political queries were refused while others were allowed (e.g., ChatGPT initially declined a poem about Trump but accepted one about Biden), sparking criticism of hidden value judgments. This controversy highlights how opaque alignment decisions can erode legitimacy and provoke political backlash.

From a democratic perspective, legitimacy requires transparency, accountability, and broad societal input. Alignment should not be dictated by a few corporations alone; pluralistic deliberation is necessary to avoid concentrating informational power and to reflect diverse perspectives.

Cultural erasure in practice: A study of Indigenous language prompts found that LLMs often refused to engage, citing lack of data or "sensitivity," even when benign questions were posed. Whereas, queries about European traditions received detailed answers. Such asymmetries show how alignment can unintentionally silence marginalized cultures [14].

4.2. Cultural Inequality

Current alignment practices largely reflect Western liberal norms and English-language data, risking cultural hegemony. Communities in the Global South may find their epistemologies underrepresented or mischaracterized, reinforcing structural digital inequalities. Alignment processes must therefore prioritize pluralism, cultural justice, and inclusion to counter the replication of historical forms of cultural dominance [6].

4.3. Political Stakes and Regulatory Futures

Alignment raises pressing questions for governance. *Who decides what values guide LLMs? Should alignment frameworks be set by private companies, governments, international organizations, or democratic deliberation?* Each option carries risks: Private companies may embed commercial interests or culturally narrow norms; governments might impose nationalistic or authoritarian priorities; international organizations could struggle with legitimacy and enforceability; and democratic deliberation, though ideal in theory, risks being slow, uneven, or captured by powerful voices. Effective governance must therefore balance inclusivity with practicality, ensuring that alignment decisions reflect diverse global values while remaining responsive to rapid technological change. Some of these risks are mentioned below:

- **Corporate control.** Concentrates alignment power, risking opaque value imposition and limited accountability.
- **National regulation.** Risks fragmentation, with models aligned to divergent national standards, creating "digital sovereignties."
- **International coordination.** Encourages harmonization but may default to lowest-denominator principles due to geopolitical tensions.

Regulatory divergence: In 2024, one LLM was released in both European and U.S. applied stricter moderation in the EU (reflecting GDPR and safety-first standards) and more permissive outputs in the U.S. (reflecting free speech norms), highlighting how alignment cannot be separated from political culture and regulatory context.

Future regulation will likely combine these approaches. For instance, the EU AI Act emphasizes transparency and risk management, while U.S. initiatives stress voluntary commitments [56]. Alignment decisions cannot be separated from political culture or regulatory context.

 doi:10.20944/preprints202603.0719.v1

4.4. Economic and Creative Implications

Operational choices in alignment affect productivity, creativity, and cultural output. Overly cautious models may suppress imagination, known as the "alignment tax," reducing innovation [27]. Conversely, under-aligned systems can cause reputational and legal risks. Pluralistic alignment can foster diverse creative ecosystems, supporting a broader range of cultural expression.

> **Pluralistic alignment in education:** An experimental deployment of LLMs in multilingual classrooms allowed teachers to adjust alignment parameters for local norms. In one setting, outputs emphasized communal values; in another, individual autonomy. Students reported feeling more represented and respected, though technical challenges of consistency remained. The case demonstrates both the feasibility and complexity of pluralistic futures [57].

4.5. Prospects for Pluralistic Futures

Future alignment should prioritize pluralism, recognizing diverse values and enabling coexistence than convergence. Key requirements include:

- **Multi-stakeholder governance:** Incorporating voices from academia, civil society, industry, and marginalized communities.
- **Decentralized oversight:** Allowing communities to shape alignment parameters locally.
- **Adaptive systems:** Negotiating conflicting values dynamically over static rules.

While pluralism introduces challenges in balancing norms and preventing fragmentation, it is essential to avoid imposing narrow standards.

4.6. Long-Term Futures: Alignment at Scale

As LLMs scale and integrate with autonomous systems, misalignment risks intensify. Future systems may act in the world – executing transactions or controlling infrastructure, where failures could have catastrophic consequences [6,10]. Instrumental convergence theory suggests that advanced agents may pursue sub-goals that conflict with human intentions [58–60].

Long-term alignment requires integrating technical safeguards with normative foresight. Research on CAI, participatory frameworks, and FJA must advance alongside governance mechanisms capable of responding to evolving societal risks.

5. Roadmap for Alignment

Alignment challenges are operational, often rooted in post-training pipeline design. A roadmap must therefore address these choices while extending beyond harm prevention toward pluralism and human flourishing, with actions spanning researchers, industry, governments, and society.

Researchers.

Develop multi-dimensional benchmarks that go beyond toxicity and helpfulness to capture cultural pluralism, reasoning quality and calibrated refusals [5,19]. Advance oversight methods such as automated red teaming, debate framework, RLAIF to reduce reliance on manual annotation [61,62]. Pursue mechanistic interpretability to enable transparency and reliable auditing [63,64].

Industry.

Invest in globally diverse annotator pools to mitigate cultural bias [6,14]. Publish transparency reports outlining alignment principles and known limitations. Link model deployment to rigorous pre-release testing and independent audits via responsible scaling policies [65].

Governments and Regulators.

Establish clear standards for transparency, and accountability, building on efforts such as EU AI Act [66]. Support independent auditing bodies and fund public research into alignment. Promote international coordination to avoid a "race to the bottom" in safety practices while allowing local autonomy [67].

 doi:10.20944/preprints202603.0719.v1

Civil Society.

Advocate for participatory platforms that allow communities to alignment values [68]. Promote digital literacy so users engage critical with LLMs rather than treating them as neutral authorities and hold developers and policymakers accountable for creating AI that serves broad and plural human interests.

6. Conclusion

In this position paper, we have argued that many alignment failures in LLMs stem not from flawed principles but from operational choices in post-training pipelines, which can produce rigid safeguards, overcautious refusals, and diminished cultural sensitivity. Moving forward, alignment must expand beyond harm-prevention to embrace pluralistic and participatory approaches that balance safety, usefulness, and diversity. Our proposed FJA framework offers one path forward, emphasizing human well-being, cultural justice, and user autonomy. By rethinking alignment through the lens of operational decisions and pluralistic values, researchers, industry, and policymakers can build systems that are not only safe but also genuinely supportive of human flourishing.

7. Limitations

As with any position paper, our discussion is shaped by the scope of this work and the perspective of its authors. Below, we outline the key limitations of our arguments and proposed roadmap.

First, this work is primarily conceptual. While we analyze alignment risks and argue that many failures stem from operational choices in post-training pipelines, we do not provide large-scale empirical evidence or benchmark studies to systematically validate these claims. Our examples are illustrative rather than comprehensive, and future work should test the robustness of these arguments across diverse deployment settings.

Second, in critiquing the HHH framework, we emphasize how rigid interventions and static reward models contribute to misalignment. While this highlights important blind spots, our discussion cannot capture all variations of HHH implementations or account for emerging adaptations. The proposed FJA framework likewise remains conceptual: it articulates virtues and operational principles but does not yet constitute a full normative or technical specification.

Third, while pluralism and participatory design are central to our roadmap, we do not present a detailed methodology for how to enact these principles in practice. Effective implementation will require resource-intensive processes, ongoing community engagement, and governance mechanisms that extend well beyond what can be addressed here. This paper therefore advocates for, rather than delivers, participatory infrastructures.

Fourth, transparency and explainability remain open challenges. We suggest verifiable quoting and participatory audit trails as promising directions, but acknowledge that scalable guarantees against hallucinated rationales or selective disclosures are not yet available. Similarly, mechanisms for reconciling competing cultural or ethical norms are only outlined at a high level.

Finally, our roadmap is necessarily selective. We focus on operational choices, pluralistic alignment, and the FJA framework, but do not cover other relevant dimensions such as economic incentives, geopolitical pressures, or the environmental costs of large-scale training. These omissions reflect space constraints rather than irrelevance.

8. Ethics and Broader Impact

Our work does not present new experimental results or involve human subjects, and therefore does not raise direct ethical concerns related to data use or experimentation. Nevertheless, the position and roadmap we propose have broader ethical and societal implications.

By advocating for pluralistic alignment thorough the FJA framework, we aim to encourage AI systems that better respect human values, culture diversity, and user autonomy. If adopted, these principles could improve fairness, accountability, and inclusivity in LLM deployment.

 doi:10.20944/preprints202603.0719.v1

At the same time, any guidance in AI alignment carries risks of misinterpretation or misuse. For example, alignment methods could be applied in ways that entrench dominant norms or create over-restrictive systems. We have attempted to address these risks by emphasizing participatory design, context-sensitivity, and the need for ongoing evaluation and governance.

Overall, we believe that the potential positive impact of advancing pluralistic and human-centered alignment outweighs the limited risks associated with this theoretical work, and we hope it will foster more responsible, equitable, and reflective AI development.

References

1. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* **2020**, *33*, 1877–1901, [arXiv:cs.CL/2005.14165].
2. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
3. Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; et al. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805* **2023**, [arXiv:cs.CL/2312.11805].
4. Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. A General Language Assistant as a Laboratory for Alignment, 2021, [arXiv:cs.CL/2112.00861].
5. Xie, Z.; Wu, J.; Shen, Y.; Xia, Y.; Li, X.; Chang, A.; Rossi, R.; Kumar, S.; Majumder, B.P.; Shang, J.; et al. A Survey on Personalized and Pluralistic Preference Alignment in Large Language Models. *arXiv preprint arXiv:2404.07070* **2025**, [arXiv:cs.CL/2404.07070].
6. Zeng, W.; Zhu, H.; Qin, C.; Wu, H.; Cheng, Y.; Zhang, S.; Jin, X.; Shen, Y.; Wang, Z.; Zhong, F.; et al. Multi-level Value Alignment in Agentic AI Systems: Survey and Perspectives. *arXiv preprint arXiv:2406.09656* **2025**, [arXiv:cs.AI/2406.09656].
7. Nussbaum, M.C. Creating capabilities: The human development approach and its implementation. *Hypatia* **2009**, *24*, 211–215.
8. Russell, S. Human-Compatible Artificial Intelligence. *Human-like machine intelligence* **2022**, *1*, 3–22.
9. Kirk, R.; Mediratta, I.; Nalmpantis, C.; Luketina, J.; Hambro, E.; Grefenstette, E.; Raileanu, R. Understanding the Effects of RLHF on LLM Generalisation and Diversity. *arXiv preprint arXiv:2310.06452* **2024**, [arXiv:cs.LG/2310.06452].
10. Zhou, D.; Zhang, J.; Feng, T.; Sun, Y. A Survey on Alignment for Large Language Model Agents. *arXiv preprint* **2024**. Reviewed on OpenReview.
11. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* **2022**, [arXiv:cs.CL/2206.07682].
12. Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C.D.; Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems* **2023**, *36*, 53728–53741, [arXiv:cs.LG/2305.18290].
13. AlKhamissi, B.; ElNokrashy, M.; Alkhamissi, M.; Tan, F.A.; Wattenberg, M.; Xie, S.M.; Roux, N.L.; Singh, S. Investigating Cultural Alignment of Large Language Models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* **2024**, pp. 12448–12471, [arXiv:cs.CL/2402.13231].
14. Wang, Y.; Zhong, W.; Li, L.; Mi, F.; Zeng, X.; Huang, W.; Shang, L.; Jiang, X.; Liu, Q. Aligning Large Language Models with Human: A Survey. *arXiv preprint arXiv:2307.12966* **2023**, [arXiv:cs.CL/2307.12966].
15. Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleer, N.; Irving, G. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint arXiv:2212.09251* **2022**, [arXiv:cs.CL/2212.09251].
16. Greenblatt, R.; Denison, C.; Wright, B.; Roger, F.; MacDiarmid, M.; Marks, S.; Treutlein, J.; Belonax, T.; Chen, J.; Duvenaud, D.; et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093* **2024**.
17. Yao, J.; Yi, X.; Wang, X.; Wang, J.; Xie, X. From Instructions to Intrinsic Human Values – A Survey of Alignment Goals for Big Models. *arXiv preprint arXiv:2308.12014* **2023**, [arXiv:cs.AI/2308.12014].
18. Shen, C.; Cheng, L.; Nguyen, X.P.; You, Y.; Bing, L. Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 4215–4233.

 doi:10.20944/preprints202603.0719.v1

19. Yu, T.; Zhang, Y.F.; Fu, C.; Wu, J.; Lu, J.; Wang, K.; Lu, X.; Shen, Y.; Zhang, G.; Song, D.; et al. Aligning Multimodal LLM with Human Preference: A Survey. *arXiv preprint arXiv:2403.14504* **2025**, [arXiv:cs.CV/2403.14504].
20. Zhang, Z.; Rossi, R.A.; Kveton, B.; Shao, Y.; Yang, D.; Zamani, H.; Dernoncourt, F.; Barrow, J.; Yu, T.; Kim, S.; et al. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027* **2024**.
21. 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. In Proceedings of the arXiv preprint arXiv:2204.05862, 2022.
22. 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* **2024**, [arXiv:cs.AI/2310.19852].
23. Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.L.; Mireshghallah, N.; Rytting, C.M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; et al. Position: A Roadmap to Pluralistic Alignment. *arXiv preprint arXiv:2402.05070* **2024**.
24. Chakraborty, S.; Qiu, J.; Yuan, H.; Koppel, A.; Huang, F.; Manocha, D.; Bedi, A.S.; Wang, M. Maxmin-rlhf: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925* **2024**.
25. Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; Irving, G. Red teaming language models with language models. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 3419–3448.
26. Ziegler, D.; Nix, S.; Chan, L.; Bauman, T.; Schmidt-Nielsen, P.; Lin, T.; Scherlis, A.; Nabeshima, N.; Weinstein-Raun, B.; de Haas, D.; et al. Adversarial training for high-stakes reliability. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35, pp. 9274–9286.
27. 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. In Proceedings of the Advances in neural information processing systems, 2022, Vol. 35, pp. 27730–27744.
28. Christiano, P.F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; Amodei, D. Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems* **2017**, *30*, [arXiv:cs.LG/1706.03741].
29. Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; Legg, S. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871* **2018**.
30. Skalse, J.; Howe, N.; Krasheninnikov, D.; Krueger, D. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems* **2022**, *35*, 9460–9471.
31. Gao, L.; Schulman, J.; Hilton, J. Scaling laws for reward model overoptimization. In Proceedings of the International Conference on Machine Learning. PMLR, 2023, pp. 10835–10866.
32. 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* **2022**, [arXiv:cs.CL/2212.08073].
33. Zhao, Y.; Joshi, R.; Liu, T.; Khalman, M.; Saleh, M.; Liu, P.J. SLiC-HF: Sequence Likelihood Calibration with Human Feedback. *arXiv preprint arXiv:2305.10425* **2023**, [arXiv:cs.CL/2305.10425].
34. Tang, Y.; Guo, D.Z.; Zheng, Z.; Calandriello, D.; Cao, Y.; Tarassov, E.; Munos, R.; Pires, B.Á.; Valko, M.; Cheng, Y.; et al. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448* **2024**.
35. Zhang, J.; Marone, M.; Li, T.; Van Durme, B.; Khashabi, D. Verifiable by design: Aligning language models to quote from pre-training data. *arXiv preprint arXiv:2404.03862* **2024**.
36. Abdulhai, M.; Crepy, C.; Valter, D.; Canny, J.; Jaques, N. Moral foundations of large language models. In Proceedings of the AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI, 2022.
37. Hendrycks, D.; Mazeika, M.; Zou, A.; Patel, S.; Zhu, C.; Navarro, J.; Song, D.; Li, B.; Steinhardt, J. What would jiminy cricket do? towards agents that behave morally. *arXiv preprint arXiv:2110.13136* **2021**.
38. Gabriel, I. Artificial Intelligence, Values and Alignment. *Minds and Machines* **2020**, *30*, 411–437, [arXiv:cs.CY/2001.09768].
39. Arrow, K.J. A difficulty in the concept of social welfare. *The Journal of Political Economy* **1950**, *58*, 328–346.
40. Birhane, A. Algorithmic injustice: a relational ethics approach. *Patterns* **2021**, *2*.
41. Sen, A. *Development as Freedom*; Anchor Books, 1999.
42. Hilliard, E.; Jagadeesh, A.; Cook, A.; Billings, S.; Skytland, N.; Llewellyn, A.; Paull, J.; Paull, N.; Kurylo, N.; Nesbitt, K.; et al. Measuring AI alignment with human flourishing. *arXiv preprint arXiv:2507.07787* **2025**.
43. Rawls, J. *A Theory of Justice*; Harvard University Press, 1971.
44. Berlin, I. Two Concepts of Liberty. *Four Essays on Liberty* **1969**.

 doi:10.20944/preprints202603.0719.v1

45. Yuan, Y.; Xiao, T.; Yunfan, L.; Xu, B.; Tao, S.; Qiu, Y.; Shen, H.; Cheng, X. Inference-time Alignment in Continuous Space. *CoRR* **2025**, *abs/2505.20081*, [2505.20081]. https://doi.org/10.48550/ARXIV.2505.20081.
46. Liu, T.; Guo, S.; Bianco, L.; Calandriello, D.; Berthet, Q.; Llinares-López, F.; Hoffmann, J.; Dixon, L.; Valko, M.; Blondel, M. Decoding-time Realignment of Language Models. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning; Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; Berkenkamp, F., Eds. PMLR, 21–27 Jul 2024, Vol. 235, *Proceedings of Machine Learning Research*, pp. 31015–31031.
47. Khanov, M.; Burapacheep, J.; Li, Y. ARGS: Alignment as Reward-Guided Search. In Proceedings of the The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024.
48. Dutta, S.; Kaufmann, T.; Glavaš, G.; Habernal, I.; Kersting, K.; Kreuter, F.; Mezini, M.; Gurevych, I.; Hüllermeier, E.; Schütze, H. Problem Solving Through Human–AI Preference-Based Cooperation. *Computational Linguistics* **2025**, pp. 1–36, [https://direct.mit.edu/coli/article-pdf/doi/10.1162/COLI.a.19/2539252/coli.a.19.pdf]. https://doi.org/10.1162/COLI.a.19.
49. Guan, M.Y.; Joglekar, M.; Wallace, E.; Jain, S.; Barak, B.; Helyar, A.; Dias, R.; Vallone, A.; Ren, H.; Wei, J.; et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339* **2024**.
50. Raina, V.; Liusie, A.; Gales, M. Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 7499–7517. https://doi.org/10.18653/v1/2024.emnlp-main.427.
51. Zhao, Y.; Liu, H.; Yu, D.; Kung, S.Y.; Mi, H.; Yu, D. One Token to Fool LLM-as-a-Judge. *CoRR* **2025**, *abs/2507.08794*, [2507.08794]. https://doi.org/10.48550/ARXIV.2507.08794.
52. Liu, Z.; Wang, P.; Xu, R.; Ma, S.; Ruan, C.; Li, P.; Liu, Y.; Wu, Y. Inference-Time Scaling for Generalist Reward Modeling. *CoRR* **2025**, *abs/2504.02495*, [2504.02495]. https://doi.org/10.48550/ARXIV.2504.02495.
53. Chen, X.; Li, G.; Wang, Z.; Jin, B.; Qian, C.; Wang, Y.; Wang, H.; Zhang, Y.; Zhang, D.; Zhang, T.; et al. RM-R1: Reward Modeling as Reasoning. *CoRR* **2025**, *abs/2505.02387*, [2505.02387]. https://doi.org/10.48550/ARXIV.2505.02387.
54. Guo, J.; Chi, Z.; Dong, L.; Dong, Q.; Wu, X.; Huang, S.; Wei, F. Reward Reasoning Model. *CoRR* **2025**, *abs/2505.14674*, [2505.14674]. https://doi.org/10.48550/ARXIV.2505.14674.
55. Zhao, J.; Liu, R.; Zhang, K.; Zhou, Z.; Gao, J.; Li, D.; Lyu, J.; Qian, Z.; Qi, B.; Li, X.; et al. GenPRM: Scaling Test-Time Compute of Process Reward Models via Generative Reasoning. *CoRR* **2025**, *abs/2504.00891*, [2504.00891]. https://doi.org/10.48550/ARXIV.2504.00891.
56. Anderljung, M.; Barnhart, J.; Leung, J.; Korinek, A.; O'Keefe, C.; Whittlestone, J.; Avin, S.; Brundage, M.; Bullock, J.; Cass-Beggs, D.; et al. Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718* **2023**.
57. Feng, S.; Sorensen, T.; Liu, Y.; Fisher, J.; Park, C.Y.; Choi, Y.; Tsvetkov, Y. Modular pluralism: Pluralistic alignment via multi-llm collaboration. *arXiv preprint arXiv:2406.15951* **2024**.
58. Bostrom, N. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines* **2012**, *22*, 71–85.
59. Hendrycks, D.; Mazeika, M.; Woodside, T. An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001* **2023**.
60. Carlsmith, J. Is power-seeking AI an existential risk? *arXiv preprint arXiv:2206.13353* **2022**.
61. Michael, J.; Mahdi, S.; Rein, D.; Petty, J.; Dirani, J.; Padmakumar, V.; Bowman, S.R. Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702* **2023**.
62. Lanham, T.; Chen, A.; Radhakrishnan, A.; Steiner, B.; Denison, C.; Hernandez, D.; Li, D.; Durmus, E.; Hubinger, E.; Kernion, J.; et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702* **2023**.
63. Olah, C.; Cammarata, N.; Schubert, L.; Goh, G.; Petrov, M.; Carter, S. Zoom in: An introduction to circuits. *Distill* **2020**, *5*, e00024–001.
64. Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652* **2022**.
65. Hubinger, E. Anthropic: Responsible Scaling Policy. *SuperIntelligence-Robotics-Safety & Alignment* **2025**, *2*.
66. Commission, E. EU AI Act, 2024.
67. OpenAI. Safety Evaluations Hub, 2025.

 doi:10.20944/preprints202603.0719.v1

68. Kirk, H.R.; Whitefield, A.; Röttger, P.; Bean, A.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams, A.; He, H.; et al. The PRISM alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019* **2024**.