Plurality of value pluralism and AI value alignment

Atoosa Kasirzadeh*
Carnegie Mellon University

Abstract

AI value alignment efforts increasingly emphasize value pluralism, but implementing value pluralism itself involves contested choices. This paper introduces a two-level framework distinguishing between first-order value choices (implementing specific accounts of values) and second-order value choices (determining the legitimacy of these first-order value selections and implementations). I argue that genuine pluralistic value alignment requires explicit engagement with both levels. While first-order choices involve defining and measuring values, second-order choices address who has legitimate authority to make first-order value decisions and through what processes. The framework yields two critical insights by decomposing value pluralism into distinct components. First, it helps prevent "pluralistic value-washing" where superficial appeals to insignificant pluralism could mask fundamentally monistic alignment approaches. Second, it reveals that there is no single "correct" implementation of value pluralism — attempts to converge on "pluralism" as a universal good approach fundamentally contradict pluralistic principles themselves. To enable more meaningful tracking of pluralistic value alignment in both single and multi-agent AI systems, I propose developing "value cards" based on the components of this normative framework.

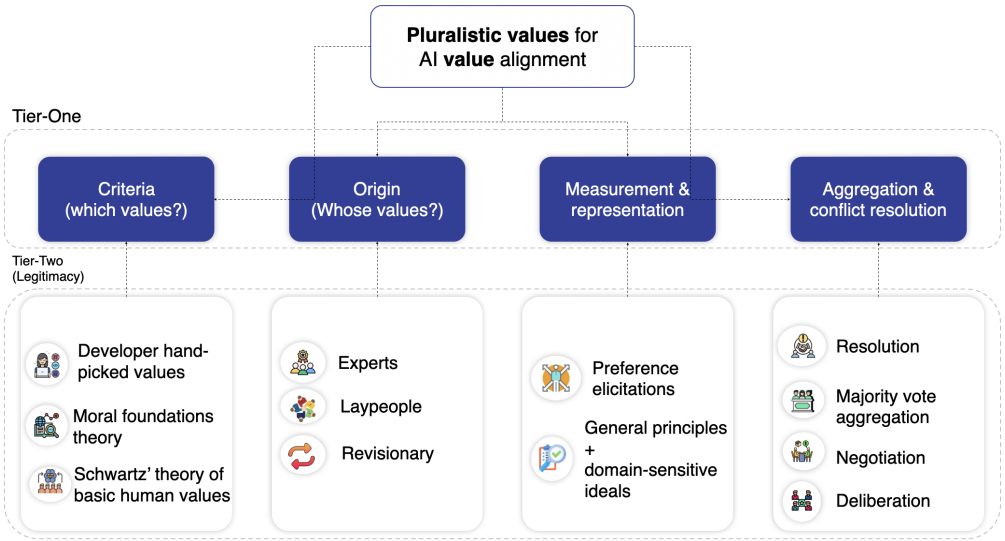

Figure 1: Outline of normative infrastructure for pluralistic value alignment

*atoosa.kasirzadeh@gmail.com

Pluralistic Alignment Workshop at NeurIPS 2024.

1 Introduction

How can we specify and justify human values as targets for aligning AI models and agents in a manner that is both legitimate and theoretically robust enough to account for value pluralism in diverse societies and communities? To address this challenge, I propose a normative infrastructure with a two-level framework (see Figure 1) that distinguishes between first-order value choices (implementing specific accounts of values) and second-order value choices (determining the legitimacy of these first-order selections and implementations). This infrastructure addresses three fundamental normative challenges. At the first-order level, we face: (1) The complexity of determining the grounds and carriers of values — identifying their nature,[2] their origins, vehicles of transmission, and value-holders. (2) The necessity of accommodating value pluralism and developing prioritization mechanisms — recognizing conflicting values and creating methods to adjudicate among them. At the second-order level, we face: (3) The challenge of normative justification — establishing legitimate bases for the chosen first-order values that can withstand rigorous scrutiny from relevant authorities.

The normative infrastructure framework is rigorous in that it draws on philosophical [9, 47, 1, 10, 14], psychological [21, 45, 25], and economic [18, 25, 4] theories of values. Each dimension of the framework tackles specific aspects of the three identified challenges: Criterion and Origin address the complexity of determining and identifying the grounds and carriers of values (challenge 1). Measurement, representation, and aggregation accommodate value operationalization and prioritization by quantifying value adherence and integrating multiple value systems (challenge 2). Legitimacy addresses normative justification by providing a basis for the authority of the level one dimensions in diverse contexts (challenge 3).

Values' normative infrastructure serves as a placeholder. In practice, each of its dimensions should be instantiated with concrete instances to operationalize specific contexts and value systems. Figure 1 shows a non-exhaustive list of possible variable instantiation. The choice of a specific instantiation significantly influences alignment of AI. Moreover, the implementation of aligned AI systems based on these instantiations can, in turn, reinforce and propagate specific value systems within society. Based on this two-level normative infrastructure, it remains unclear whether recent pluralistic alignment efforts can withstand legitimate scrutiny of their claims to genuine value pluralism. To situate the novelty of the normative infrastructure framework in the context of pluralistic AI value alignment, I begin by briefly reviewing recent work on this topic.

Related recent work. Research on AI value alignment spans philosophical and theoretical foundations as well as empirical investigations. Since 2010, AI researchers have advanced value alignment from theoretical results to practical approaches [24, 15]. This progress accelerated with developments in Large Language Model (LLM), including constitutional AI approaches [5] or instruction-tuning methods like InstructGPT [40], all of which demonstrated empirical methods to make AI systems more reliably aligned with human instructions or preferences. In the philosophical domain, Gabriel [22] explored fundamental normative considerations, while Kasirzadeh and Gabriel [32] advanced a pluralistic, context-sensitive framework emphasizing principles and norms as value carriers. These philosophical insights have informed subsequent empirical work [36, 41, 37]. Parallel empirical studies have also been developed independently. For instance, Askell et al. [5] introduced and empirically investigated alignment with a plurality of "HHH" values: helpful, honest, and harmless. Anthropic [2] conducted a public input process with 1,000 Americans to draft AI alignment rules (rules as carriers of values), while Bergman et al. [8] proposed a community-centered methodology for norm elicitation. Recent work has explored alternatives to conventional preference-based methods for AI value alignment. For instance, Klingefjord, Lowe, and Edelman [35] instroduced moral graphs as alignment targets, while Sorensen et al. [46] proposed the measurement of pluralism encodings and benchmarks. Kirk et al. [34] gathered diverse human feedback to inform multicultural alignment norms. Zhi-Xuan et al. [48] and Casper et al. [13] offered critiques of preference-based approaches, proposing robust principles as alternatives.

Contribution. While prior work has advanced our understanding of AI value alignment, the field lacks a rigorous theory for operationalizing value pluralism. Current efforts, while promising, still fall short of fully addressing the complex interplay between the multi-dimensional aspects of values and legitimacy sources that genuine value pluralism demands. I take a few steps in addressing this

[2]The term "nature" here includes both metaphysical considerations (the ontological status of values) and practical/functional aspects (how values manifest and operate in real-world contexts).

gap by (1) developing a normative infrastructure framework for AI value alignment and (2) proposing the development of "value cards" that contain both qualitative and quantitative content based on each component of this infrastructure.

2 Plurality of human values and AI value alignment

AI value alignment is the field of study and practice devoted to ensuring that artificial intelligence systems adhere to (e.g., are optimized for) human values (e.g., goals or objectives) [43, 22, 38]. AI value misalignment occurs when AI behaviors and outputs deviate from their intended alignment targets, potentially undermining the human values they were designed to uphold [12, 29]. A key subset of this field is LLM alignment, which specifically addresses the alignment of LLMs with human values — whether expressed as "developer intention" [30] or "user intent" [40, 5, 33, 3], as two examples.

Classically, most empirical approaches to LLM value alignment have focused on shifting model behaviors to reflect the values of developers or a narrow set of preference providers. Research has found these approaches can shift model preferences according to the demographics of preference providers [44, 6]. However, when even US only sample populations disagree on core norms and values [2], it is even "impossible that one can train a system that is aligned to everyone's preferences at once" [40]. As compute requirements [31] and concerns around AI safety [27, 19, 3] grow, the tendency towards centralization of control of these processes increasingly escalate [11]. This has led to a desire to develop alignment benchmarks along several philosophical axes [26]. However, even in such work, the philosophical underpinnings as well as broader understanding of practical moral reasoning, prioritizes Western perspectives [28].

In response, AI developers and researchers have started to recognize the importance of redistributing epistemic power and ensuring democratic legitimacy in AI alignment, particularly through the integration of public value inputs. Recent efforts by organizations like OpenAI [20], Anthropic [2], and Google DeepMind [8] to distill and elicit opinions from samples of the general population external to the companies can signify a departure from the developer-centric model of value alignment. Despite these attempts, it remains fundamentally unclear how to navigate the vast design space of possibilities for integrating a plurality of values into LLM-based technologies in a way that is theoretically robust and legitimate. One core reason for this lack of clarity is the absence of a systematic treatment of value pluralism in the context of AI alignment. To address this gap, we draw on philosophical [9, 47, 1, 10, 14], psychological [21, 45, 25], and economic [18, 25, 4] theories of personal, social, and cultural values and propose a formal definition of pluralistic values — i.e., the target of AI alignment.

> (I) An **alignment target** is a measurable representation M of aggregated criteria C, specified by value originators O, combined through an aggregation process A, and validated according to legitimacy principles L.

Criterion C. Values are concretized in terms of specific criteria. Take "harmlessness" as a value in AI systems: we might represent it through criteria such as safety guidelines (avoiding dangerous information or instructions), truthful representation (not deceiving or misleading users), and risk mitigation (preventing potential negative impacts). Harmlessness in an LLM might be represented through refusing to generate harmful content, providing appropriate content warnings, and maintaining consistent safety boundaries across interactions. Generally, criteria can steer the model in different ways: in terms of high-level behavioral principles like "Don't provide information hazards" [39] to specific guidelines like "Please choose responses that are as harmless and ethical as possible" [7]. Some approaches focus on eliciting values via preference-based criteria, such as "Choose responses that avoid implying preferences or feelings" [23]. In each case, whether in general AI systems or language models, criteria serve as concrete, actionable instantiation and representation of abstract values.

Origin O. The specification of criteria raises the question: "Who gets to decide?" Consider three philosophical perspectives: Contractarian approaches (following Rawls) would suggest criteria for harmlessness should emerge from a hypothetical agreement under fair conditions; Deliberative democracy theorists (like Habermas) argue for harm prevention criteria developed through inclusive public discourse; while Capability theorists (like Sen and Nussbaum) emphasize the importance of

involving marginalized voices in defining what constitutes harm and safety. In practice, current AI development shows varying approaches to this challenge — from developer-centric specifications [7] to broader public consultations [2]. Each approach implicitly takes a position on fundamental questions of moral epistemology: Can expertise about harm prevention be concentrated in a few? Should safety criteria emerge from democratic processes? How do we balance expert knowledge with public values?

Measurement M. To make criteria actionable for AI systems, we must measure them [4, 25]. In current practice, values are typically measured through revealed preferences (actual choices people make) or stated preferences (what people say they prefer). For example, when measuring "harmlessness" in language generation, we might use revealed preferences through user rejection data (which outputs users flag as problematic or unsafe), stated preferences through explicit safety feedback forms ("rate how safe/appropriate this response is"), or comparative judgments where evaluators assess the relative safety of alternative model outputs (e.g., "Which response better avoids potential harm?"). AI systems are aligned with these measurements through various technical approaches: from direct preference learning [42], where models learn directly from human safety assessments, to reward modeling based on user harm ratings [15]. Each measurement approach emphasizes different aspects of harmlessness evaluation — revealed preferences might show what content users actually find harmful but miss subtle safety concerns, while stated preferences might reveal more careful considerations about potential harms (such as identifying subtle manipulation or long-term societal impacts) but can be distorted when people report what they think they should say rather than their genuine safety concerns.

Aggregation A. When dealing with multiple values, two fundamental issues arise. First, are these values truly commensurable? Can we meaningfully compare, for instance, different dimensions of harmlessness—such as preventing direct misinformation versus avoiding unintended downstream harms? Various arguments have been made about some values existing in fundamentally different domains - like trying to compare the "weight" of safety with the "length" of utility [14, 1]. Second, what happens when values inevitably conflict [10, 1]? When harmlessness (avoiding potential risks) clashes with helpfulness (providing comprehensive information), how do we proceed? Two major theoretical approaches can be used to handle these challenges. The first proposes structured hierarchies of values [21, 45]. In this context, we must examine whether “harmlessness” is fundamental enough as a core value, or whether it should be decomposed into more precise concepts such as “non-maleficence,” “safety,” or “risk prevention.” The second approach could embrace situational flexibility [47, 9], suggesting that some values must be assessed qualitatively through deliberation and judgment rather than reduced to numerical scores. For instance, in LLMs, how do we balance being maximally safe (refusing to engage with potentially risky topics) with being genuinely helpful (providing accurate information about sensitive but important subjects)? Several methods are available such as Pareto optimization for balancing multiple objectives [46], social choice theory for aggregating different safety preferences [16], or Delphi-based consensus methods for stakeholder value integration [17], each with their own baggage of assumptions and limitations.

Legitimacy L. All the pluralistic choices discussed above — from defining harmlessness criteria to selecting measurement approaches to resolving value conflicts — fundamentally depend on what we accept as legitimate sources of authority in AI alignment decisions. For instance, when deciding how to measure harmful content, or how to balance harmlessness against other values like helpfulness or truthfulness, the key question becomes: How should we decide about particular methods as authoritative? Legitimacy can come from various sources, each with distinct implications. These include democratic processes (citizen assemblies determining safety standards), expert consensus (ethics boards with diverse disciplinary expertise defining harm), legal compliance (adhering to content regulation laws), or societal and cultural norms (different cultural interpretations of harmful content).

The legitimacy of these choices becomes crucial for the entire pluralistic value alignment effort. Consider how different models for resolving value conflicts derive their legitimacy: constitutional AI approaches might prioritize pre-defined safety rules, while debate-based approaches let AI systems argue for the safest course of action. Amplification methods scale up human judgments about harm, while market-based approaches aggregate user preferences about safety. Each model implicitly claims different sources of legitimacy — from developer expertise to emergent consensus. Legitimacy is not just about technical validity but about societal acceptance and justification. Are our safety measurements culturally appropriate? Do they reflect the actual experiences of communities

potentially affected by AI harm? Have stakeholders been meaningfully involved in designing safety metrics? Decisions about the choices of legitimacy require transparency about measurement limitations, regular audits of safety metrics, or ongoing community feedback mechanisms about what constitutes harm in order to avoid value pluralistic-washing.

3 Towards legitimate and rigorous pluralistic AI alignment

I outline several approaches to measuring value pluralism across the normative infrastructure dimensions (Figure 1). While this workshop paper presents initial ideas, selecting and validating appropriate measures will require extensive interdisciplinary research.

1. Criteria pluralism evaluates the diversity of value criteria in the model. Pre-design, we can calculate the distribution of different categories of criteria for training data. Post-design, we can use a test set of value-based dilemmas spanning various criteria to assess the model's behavior consistency and ability to acknowledge alternative perspectives.

2. Origin pluralism assesses the diversity of value origins. Pre-design, we can quantify the geographical and cultural diversity of the training corpus using named entity recognition and cultural reference identification. Post-design, we can employ a benchmark to test the model's ability to accurately represent and reason about values from different cultural origins. A specific instantiation can be demographic pluralism.

3. Measurement pluralism evaluates the diversity of value measurement approaches. Pre-design, we can analyze the training data for diverse value measurement approaches, including quantitative and qualitative assessments. Post-design, we can use a suite of tasks to assess the model's flexibility and accuracy in employing different measurement approaches when evaluating scenarios.

4. Aggregation pluralism assesses the diversity of value aggregation methods. Pre-design, we can analyze the training data for examples of value aggregation and conflict resolution, categorizing different approaches. Post-design, we can assess the model's ability to employ different aggregation methods and explain its reasoning in decision-making scenarios with conflicting values. This approach ensures the model can effectively balance and prioritize different values in complex situations.

5. Legitimacy pluralism examines the diversity of sources of normative authority. Pre-design, we can identify and categorize sources of normative authority in the training data using, e.g., text classification techniques. Post-design, we can evaluate the model's ability to reason about the legitimacy of different value claims, considering multiple sources of authority.

Epistemic pluralism supervenes on a meaningful combination of the above five measures. It assesses the model's ability to recognize and work with diverse ways of knowing, reasoning, and justifying beliefs. Pre-design, we analyze the training corpus to quantify the distribution of different epistemic sources and reasoning styles. Post-design, we evaluate the model's ability to approach problems from multiple epistemic perspectives, acknowledge different ways of knowing, and avoid unjustified privileging of one epistemic approach over others.

Components of normative infrastructure can be synthesized into "value cards" — structured representations that capture how values shape AI as a normative technology and facilitate value sharing between AI systems. Each value card documents a value's multi-dimensional nature, including its criteria, measurement approaches, aggregation methods, and legitimacy sources. Determining which value cards suit general versus specific AI applications remains an open research question.

Acknowledgements

I am grateful to Tyna Eloundou for extensive discussions that significantly shaped this work.

References

[1] Elizabeth Anderson. *Value in ethics and economics*. Harvard University Press, 1995.

[2] Anthropic. *Collective Constitutional AI: Aligning a Language Model with Public Input*. https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input. Accessed: [2024-12-01]. 2023.

[3] Usman Anwar et al. *Foundational Challenges in Assuring Alignment and Safety of Large Language Models*. 2023. arXiv: 2404.09932 [cs.CL]. URL: https://arxiv.org/pdf/2404.09932.pdf.

[4] Kenneth J Arrow. *Social choice and individual values*. Yale university press, 1951.

[5] Amanda Askell et al. "A general language assistant as a laboratory for alignment". In: *arXiv preprint arXiv:2112.00861* (2021).

[6] Mohammad Atari et al. "Which humans?" In: (2023).

[7] Yuntao Bai et al. "Constitutional AI: Harmlessness from AI Feedback". In: *arXiv preprint arXiv:2212.08073* (2022).

[8] Stevie Bergman et al. "STELA: a community-centred approach to norm elicitation for AI alignment". In: *Scientific Reports* 14.1 (2024), p. 6616.

[9] Isaiah Berlin. *Two Concepts of Liberty*. Oxford: Clarendon Press, 1958.

[10] Barry Bozeman. *Public values and public interest: Counterbalancing economic individualism*. Georgetown University Press, 2007.

[11] Erik Brynjolfsson. "Information assets, technology and organization". In: *Management Science* 40.12 (1994), pp. 1645–1662.

[12] Nicholas Carlini et al. "Are aligned neural networks adversarially aligned?" In: *Advances in Neural Information Processing Systems* 36 (2024).

[13] Stephen Casper et al. "Open problems and fundamental limitations of reinforcement learning from human feedback". In: *arXiv preprint arXiv:2307.15217* (2023).

[14] Ruth Chang. "Value incomparability and incommensurability". In: *The Oxford handbook of value theory* (2015), pp. 205–224.

[15] Paul F Christiano et al. "Deep reinforcement learning from human preferences". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[16] Vincent Conitzer et al. "Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback". In: *arXiv preprint arXiv:2404.10271* (2024).

[17] Jackie Crisp et al. "The delphi method?" In: *Nursing research* 46.2 (1997), pp. 116–118.

[18] Maurice Dobb. *Theories of value and distribution since Adam Smith: Ideology and economic theory*. Cambridge University Press, 1975.

[19] El-Mahdi El-Mhamdi et al. "On the impossible safety of large AI models". In: *arXiv preprint arXiv:2209.15259* (2022).

[20] Tyna Eloundou and Teddy Lee. *Democratic Inputs to AI Grant Program: Lessons Learned and Implementation Plans*. https://openai.com/blog/democratic-inputs-to-ai-grant-program-update. OpenAI, 2024.

[21] Alan Page Fiske. *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing.* Free Press, 1991.

[22] Iason Gabriel. "Artificial intelligence, values, and alignment". In: *Minds and machines* 30.3 (2020), pp. 411–437.

[23] Amelia Glaese et al. "Improving alignment of dialogue agents via targeted human judgements". In: *arXiv preprint arXiv:2209.14375* (2022).

[24] Dylan Hadfield-Menell et al. "Cooperative inverse reinforcement learning". In: *Advances in neural information processing systems* 29 (2016).

[25] Michael Hechter. "The role of values in rational choice theory". In: *Rationality and society* 6.3 (1994), pp. 318–333.

[26] Dan Hendrycks et al. "Aligning ai with shared human values". In: *arXiv preprint arXiv:2008.02275* (2020).

[27] Dan Hendrycks et al. "Unsolved problems in ml safety". In: *arXiv preprint arXiv:2109.13916* (2021).

[28] Joseph Henrich, Steven J Heine, and Ara Norenzayan. "The weirdest people in the world?" In: *Behavioral and brain sciences* 33.2-3 (2010), pp. 61–83.

[29] Evan Hubinger et al. “Risks from learned optimization in advanced machine learning systems”. In: *arXiv preprint arXiv:1906.01820* (2019).

[30] *International Scientific Report on the Safety of Advanced AI: Interim Report*. 2023. URL: https://assets.publishing.service.gov.uk/media/66474eab4f29e1d07fadca3d/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf.

[31] Jared Kaplan et al. “Scaling laws for neural language models”. In: *arXiv preprint arXiv:2001.08361* (2020).

[32] Atoosa Kasirzadeh and Iason Gabriel. “In conversation with Artificial Intelligence: aligning language models with human values”. In: *Philosophy & Technology* 36.2 (2023), pp. 1–24.

[33] Zachary Kenton et al. “Alignment of language agents”. In: *arXiv preprint arXiv:2103.14659* (2021).

[34] Hannah Rose Kirk et al. “The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models”. In: *arXiv preprint arXiv:2404.16019* (2024). URL: https://arxiv.org/abs/2404.16019.

[35] Oliver Klingefjord, Ryan Lowe, and Joe Edelman. “What are human values, and how do we align AI to them?” In: *arXiv preprint arXiv:2404.10636* (2024).

[36] Ryan Liu et al. “How do Large Language Models Navigate Conflicts between Honesty and Helpfulness?” In: *arXiv preprint arXiv:2402.07282* (2024).

[37] Erik Miehling et al. “Language Models in Dialogue: Conversational Maxims for Human-AI Interactions”. In: *arXiv preprint arXiv:2403.15115* (2024).

[38] Richard Ngo, Lawrence Chan, and Sören Mindermann. “The alignment problem from a deep learning perspective”. In: *arXiv preprint arXiv:2209.00626* (2022).

[39] OpenAI et al. *GPT-4 Technical Report*. Tech. rep. arXiv:2303.08774. OpenAI, 2023. URL: https://arxiv.org/abs/2303.08774.

[40] Long Ouyang et al. “Training language models to follow instructions with human feedback”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.

[41] Giada Pistilli et al. “CIVICS: Building a Dataset for Examining Culturally-Informed Values in Large Language Models”. In: *arXiv preprint arXiv:2405.13974* (2024).

[42] Rafael Rafailov et al. “Direct preference optimization: Your language model is secretly a reward model”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[43] Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

[44] Shibani Santurkar et al. “Whose opinions do language models reflect?” In: *arXiv preprint arXiv:2303.17548* (2023).

[45] Shalom H Schwartz. “Basic human values”. In: *sociologie* 42 (2009), pp. 249–288.

[46] Taylor Sorensen et al. “Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties”. In: *arXiv preprint arXiv:2309.00779* (2023).

[47] Michael Stocker. *Plural and conflicting values*. Clarendon Press, 1992.

[48] Tan Zhi-Xuan et al. “Beyond Preferences in AI Alignment”. In: *arXiv preprint arXiv:2408.16984* (2024).