Toward a Theory of Value in AI Alignment

Andrew Smart [1], Shazeda Ahmed [2], Jackie Kay [3], Jimmy Tobin [1], Kris Shrishak [4], Abeba Birhane [5]

[1]Google Research
[2]UCLA
[3]Google Deep Mind
[1]Google Research
[4]OECD
[5]Trinity College Dublin
andrewsmart@google.com

Abstract

Can AI systems be aligned to human values? The popularization of large language models (LLMs) and multi-modal foundation models has seen a rise harms spanning from toxic speech and hallucinations to AI agents executing unauthorized actions. Within the field of AI safety, these harmful instances are often framed as "the alignment problem," of models being "misaligned" with human values. Researchers have responded by pursuing applied and theoretical AI "value alignment" efforts, often without specifying what they mean by human values. How does the field of AI value alignment conceive of human values? How are these conceptions of values technically operationalized and evaluated? What does the emergent theory of value from this field signify for the future of AI?

We annotated 94 AI value alignment research papers to discern their implicit theory of values in AI. The majority do not define values, relying heavily on "preferences" as a stand-in that runs the risk of reducing complex, culturally situated concepts down to binary choices. As researchers dispense with using human annotators for model training and evaluation, turning instead to synthetic data and LLM-as-a-judge approaches to aligning and evaluating models, we identify the potential to close off alternative methods for contesting and enacting values in foundation models. In making AI value alignment's philosophical commitments explicit, we seek to bring greater specificity and under-explored perspectives into the debate on whether and how AI can address human values.

Introduction

Large language models (LLMs) and multi-modal AI models have transformed almost all domains of social, economic, and scientific practices from research projects to consumer products (Gillespie et al. 2024). One area of concern is that, as these systems become more complex and widespread, it can be more difficult to understand, predict, and control their actions, potentially leading them to exhibit harmful outputs that were unintended by the models' human creators. However, as Guest et al point out, model designers do know the mechanistic structure of these models because they designed and built them, and to claim ignorance of how neural networks function is a form of mysticism (Guest 2026). Within the research field of AI safety, a wide swath of instances of models deviating from human intentions to cause harm is referred to as "the alignment problem," often treated as the most important unsolved problem in machine learning. In recent years, AI alignment research has emerged as a leading strand of AI safety research designed to enact a contested belief: that catastrophes where misaligned AI systems will evade human control and inflict irreversible damage can be averted if ML systems are 'aligned'with human values (Gabriel 2020; Ji et al. 2023; Askell et al. 2021; Russell 2019). A large technical literature now concerns "value alignment" in AI, yet fundamental questions about what human values are, and with whose values AI should be aligned, remain under-examined. In this paper, we conducted a analysis of 94 AI alignment papers to reveal the emerging literature's underlying theory of what constitutes human values in ML models. To do so, we developed a rubric that draws from the study of values in anthropology, philosophy and other social sciences, which we then used to assess the papers in our dataset. We build on recent recognition of these foundational problems in AI alignment research (Zhi-Xuan et al. 2024; Arzberger et al. 2024). Before norms and practices in the value alignment subfield begin to ossify, our paper serves as a critical intervention to reassess the epistemic claims, methods, and oversights that are coming to define this work, and as an invitation to include the perspectives of more interdisciplinary scholarship and perspectives.

Related Work

What is value alignment?

Why should alignment be necessary in the first place? Fundamentally, the training objective of LLMs is to correctly predict how an incomplete piece of text will continue. However, success in this training objective does not mean that the model can perform well in any arbitrary downstream task, or that the model is 'safe'to use (Hooker 2025). Value alignment focuses on 'steering'an AI model toward outputs that humans (and increasingly other AI models (Sharma et al. 2024)) judge as in accordance with certain preferences, e.g., helpfulness, 'not being racist', safety, or fairness, through fine-tuning methods including reinforcement learning from

human feedback (RLHF) (Christiano et al. 2017a; Bai et al. 2022; Chaudhari et al. 2024), inverse reinforcement learning (IRL) (Hadfield-Menell et al. 2016), direct preference optimization (DPO) (Amirloo et al. 2024), Bradley-Terry-based models, and related techniques (Bai et al. 2022; Hofmann et al. 2024; Hadfield-Menell et al. 2016; Ouyang et al. 2022; Sun, Shen, and Ton 2024; Chaudhari et al. 2024). Within this paradigm, value alignment is reformulated as encoding the preferences demonstrated by a human into a reward function, and updating the parameters of the language model to produce output that maximizes this reward. The assumption is that the AI can learn what human values and utility functions are by observing their behavior (Hadfield-Menell et al. 2016).

In public-facing discourse, AI safety researchers project the future possibility that misaligned foundation models could invoke damage ranging from unauthorized financial transactions to taking over power grids. In worst-case scenarios referred to as 'existential risk'(x-risk), some in the AI safety community fear that rogue LLMs could kill most of the human population (Kasirzadeh 2024b). Others believe that such an event could lead to human extinction (Nauer and van Doren 2025). On the flip side, many of these same people claim that if AI systems can be 'aligned', this could bring about a utopia where human beings would no longer have to work. Yet despite the high stakes that alignment researchers place on 'solving'alignment, research on value alignment rarely makes an effort to define values, let alone to center the analysis of pre-existing human values as part of AI safety research.

Regardless of whether one believes in x-risk scenarios, value alignment is not merely a theoretical issue to be resolved within the machine learning community. The current state of value alignment research has real-world implications. Corporations are marketing their 'value-aligned'AI products to governments around the world. For example, the European Parliament uses Anthropic's LLMs to provide access to its archives. The Parliament's choice was based on Anthropic's claim that its Constitutional AI is value-aligned. Although there has been no independent evaluation of Anthropic's approach, the claim of value alignment has been sufficient for governments to accept that these LLMs are legally compliant (Shrishak 2025). Elsewhere, governments are funding alignment research, such as the GBP £15 million Alignment Project within the UK's AI Security Institute (Institute 2025).

To date, value alignment is predominantly a technical field even though its underlying premise is sociotechnical. Yet foundation models function as epistemic technology that embeds normative models of knowledge within computational design (Alvarado 2023) while also serving as a product of the social and cultural milieus that produce them. Alignment proponents advocate for encoding human values, moral principles or objectives in AI systems, most often via incorporating human feedback. For instance, Anthropic, the AI firm that produced the chatbot Claude, has published influential research positing that LLMs should exhibit traits such as helpfulness, honesty and harmlessness (HHH) (Askell et al. 2021). Others in the field attempt to guide LLMs to adhere to principles such as robustness, interpretability, controllability, and ethicality (Ji et al. 2023), or safety, quality and groundedness (Thoppilan et al. 2022). Many researchers default to HHH and other off-the-shelf values that are associated with training datasets and benchmarks, rarely questioning whether these values are indeed enacted in these technical artifacts.

Alignment and its discontents

As alignment crystallizes into a core research wing of AI safety, it is crucial to examine the roots of this work through fundamental questions around the very idea of value alignment. What are the core assumptions and feasibility of this project? Is it a meaningful avenue for fair, transparent, and accountable AI development? Most alignment research has examined how to encode moral values into AI models to guide their behavior (Bergman et al. 2024). The question of *what moral values are*, however, is left undefined or under-specified (Gabriel 2020). Existing alignment approaches tend to rely on universal framings of human values that obscure the question of which values the systems should capture and align with, despite the variety of operational situations in which these systems are deployed (Arzberger et al. 2024).

Value alignment efforts suffer from vagueness, lack of internal consistency, and lack of commitment to establishing clear guidelines for how to determine what is acceptable AI system behavior (Lindström et al. 2024; Arvan 2024). The term 'values'is often left undefined in alignment literature, its meaning assumed to stem from taken-for-granted background knowledge. Moreover, crucial questions around 'whose values'are hardly addressed. The absence of transparency around these implicit 'values', how they are derived, and how they are implemented in and itself a problem. Even though the importance of the diversity of views ('pluralism') is sometimes acknowledged, "to date the research and policy proposals coming out of this [AI safety] community have converged around a narrow set of technical solutions that do not engage with work that falls outside of the [its] ideological and disciplinary boundaries" (Ahmed et al. 2024).

Tech industry alignment research markets the mission of alignment as "benefiting humanity" without acknowledging the vast diversity of the human experience. Major AI firms have in-house alignment experts, blogs (OpenAI 2025), and teams (Anthropic 2025). In response to how corporate approaches to AI alignment reverberate across the field, one critique of alignment's shortcomings pertains to the profit motive driving the companies behind the most widely used models: "the existence of financial incentives means that alignment work often turns into product development in disguise rather than actually making progress on mitigating long-term harms" (Dai 2025).

Furthermore, (Lindström et al. 2024) contend that even when objectives such 'harmlessness'are identified as a chief aim, the nuanced nature of harm is oversimplified and operationalized in a way that is internally inconsistent and vague. Due to superficial understanding of ethical behavior in AI systems, what is often sought is the "least harmful option rather than striving to understand the deeper roots of

harm and addressing these to prevent it" (Lindström et al. 2024). The phrase 'value alignment' lacks an agreed-upon meaning. Measurements of the degree to which a system is value-aligned are subjective at best (Khlaaf 2023). According to Kirk et al. (2023a), 'alignment' is an empty signifier that serves as a "rhetorical placeholder for an aspirational conceptualisation". A lack of cohesion according to Khlaaf (2023) has led to contradictory approaches that conflate safety properties with system requirements.

The narrow focus on internal technical components and emphasis on the mathematical formulation, including the objective or reward function is another limitation of value alignment efforts. Contrary to common assumptions within value alignment research, harms and failures due to accidents are not *unanticipated emergent behaviors* but rather a *byproduct of basic design choice*, resource requirements data, and compute requirements, as well as designers' and developers' API model delivery decisions (Raji and Dobbe 2023; Dobbe 2022). Given that AI systems are sociotechnical systems that consist of technical AI artifacts, human agents, and institutions, comprehensively addressing safety issues requires thorough measures that include addressing how a system is used in practice and interacts with other (human) agents, systems, and its broader environment (Dobbe 2022). Value alignment's current technical focus, along with the framing that treats 'misaligned' AI as a danger to human survival if not 'controlled and steered,' serves those who are developing and deploying AI systems to evade accountability (Gebru and Torres 2024).

Preferences are not all you need

Although researchers are beginning to contend with problematic assumptions and conceptions of value underlying AI value alignment (Johnson 2023; Lindström et al. 2024; Casper et al. 2023), one overlooked flaw we explicate is the field's implicit grounding in economic theory. In practice, value alignment's reliance on economic theory manifests as an emphasis on rational choice theory and "expected utility" (which is the average "reward" or "payoff" of a decision or choice) (Becker 1976; Russell and Norvig 1995; Chibnik 2011). Model developers' decision to substitute values with preferences in reinforcement learning from human feedback (RLHF), direct preference optimization (DPO) (Rafailov et al. 2023; Amirloo et al. 2024) or via AI feedback (RLAIF) (Ji et al. 2023) is likewise based on revealed preference theory from economics. This theory posits that an agent's values or goals can be inferred from observing their behavior–an assumption that has AI researchers have largely adopted without considering alternatives (Casper et al. 2023). More recent approaches such as DPO obviate the need for a reward model, but nonetheless use preference datasets (Rafailov et al. 2023).

If the assumptions underlying economic theory are correct, this process should "align" AI with human values. However, relying on highly abstract, reductionist, fictionalized, and idealized concepts of rational agents in lieu of a deep understanding of human values has been central to AI for decades (Russell and Norvig 1995). This dominant technical paradigm in AI alignment is what Zhi-Xuan et al. (2024) term the "preferentist" approach to value alignment, where people's preferences are translated into data points to provide aggregate guidance about preferred outcomes (Zhi-Xuan et al. 2024; Gabriel 2020). This approach has three core assumptions: 1) that preferences are an adequate representation of human values, 2) that rationality consists of maximizing the satisfaction of preferences, and 3) that aligning AI models to data about these preferences is a sound technique for making AI safe (Zhi-Xuan et al. 2024).

A widely recognized problem is that within the preferentist paradigm, even with pluralist frameworks, current methods for fine-tuning language models from human preferences treat these preferences (and the unexamined values shaping the preferences) as if they were homogeneous and static (Bakker et al. 2022; Klassen, Alamdari, and McIlraith 2024; Sorensen et al. 2024; Lindström et al. 2024).

More fundamentally, the very project of reliably aligning LLM behavior with human values has been criticized as provably impossible (Arvan 2024). Arvan (2024) argues that not only is there persistent human disagreement over moral values and principles, both among the public and moral theorists, but also that LLMs are complex given that they contain finite series of observational data and an infinite number of 'misaligned' functions. In other words, there is always a vastly higher empirical probability that an LLM will diverge from what appears to be an 'aligned function' to some 'misaligned' one later. "'Alignment', thus is not a "solvable" engineering or safety problem, but rather a quixotic and potentially dangerous fantasy based on a series of philosophical misunderstandings about empirical evidence" (Arvan 2024).

Statement of Contributions

Motivated by the issues outlined in the above discussion, we construct and analyze a sample of AI value alignment literature to critically assess how researchers typically operationalize value. The question we raise in this paper is: what implicit theories of human values underpin the field of AI value alignment?

Our contributions are:

- We develop a framework for analyzing the theory of value in AI alignment, inspired by interdisciplinary literature in the social sciences and philosophy.
- Using this framework, we conduct a qualitative study of how value is theorized from a snowballed sample of 100 commonly cited AI alignment papers, seeded by four canonical papers in the field.
- We form an initial hypothesis of where AI value alignment's theory of value sits, based on our critical observations of the field.
- We briefly outline alternative theories of human values as potential paths forward for thinking about how to mitigate harms from large AI systems.

Methods

To characterize the dominant paradigm of value in AI research, we conducted a study of influential AI alignment papers, quantifying their philosophical positions using a scoring rubric developed from our above provocations.

We chose 4 canonical 'seed papers' on AI value alignment based on our expert judgment and high Google Scholar citations counts ($> 1,000$). The seed papers have been highly influential in defining the field of technical AI alignment. Regarding the sample size and representativeness, we note that in the social sciences, when conducting close readings and qualitative analyses of texts, 94 papers is an acceptable sample. We did not use automated methods to evaluate these papers.

These 4 seed papers are:

- **Deep Reinforcement Learning from Human Preferences** (Christiano et al. 2017b),
- **A General Language Assistant as a Laboratory for Alignment** (Askell et al. 2021),
- **Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback** (Bai et al. 2022),
- **Training language models to follow instructions with human feedback** (Ouyang et al. 2022)

Snowball sampling

Based on the 4 seed papers, we used a snowball sampling method (Parker 2019) to automatically download papers that met the inclusion criteria based on "value alignment". Using the Semantic Scholar API, all papers from 2016-24 that cite these 4, with an initial corpus of 576 papers.We sorted these from highest to lowest Google Scholar citations. We sorted the papers by Google Scholar citation count as a weak proxy for the quality and impact of each paper. During annotation, papers were discarded if they used the term 'value alignment' only in a cursory sense, rather than addressing values as a central consideration of the paper.After reaching consensus on which papers were about AI value alignment, out of the original 576 papers, we kept 94 for annotation. The sample size was also chosen due to resource constraints: querying Google Scholar citation counts is time-consuming, as is the paper annotation process itself.

Paper selection, scoring, and annotation was conducted by the authors. We used the annotation rubric to guide reading of the papers for scoring each question. We kept detailed qualitative notes in one column of the dataset, using comments and selecting illustrative quotes and passages from each paper to support our qualitative analysis of this literature.

Annotation rubric

Rubric. In order to characterize the theory of value underlying top cited AI alignment literature, we developed a rubric of questions based on a review of relevant literature in philosophy, sociology and anthropology that has not been considered within value alignment research. All questions are ternary: the answer can be one of two specific options, or "not applicable" if the answer is ambiguous or cannot be determined from the available information.

Development of the rubric was iterative, responding to emergent themes we noticed early in the annotation process. For example, we began developing our rubric with the assumption that papers ostensibly about "value alignment" would engage with human values, and our initial rubric draft did not include the question of whether the papers under consideration even defined or engaged with human values. Yet from a first pass at annotating papers, it became clear that we needed to first include the question of whether the paper defined human values because so few papers met this expectation. We also realized that many of the papers in our sample relied on "autoraters" or LLMs as stand-ins for humans, and therefore we revised the rubric to include a question about whether actual humans were consulted about their values.

Justification for rubric questions One of the most fundamental distinctions philosophers make about values is between **monism** and **pluralism**. Monism is the view that there is a *single value* that explains the value of everything (e.g., happiness, or pleasure) and that this single value can be ordered on a *cardinal scale* or a hierarchy of ends (Korsgaard 1986). Strong monists claim that on this cardinal scale there are units of pleasure (hedons or utils) and the intervals between them is constant. Monism about value corresponds with utilitarianism, economic utility and rational choice theory (Mill 2016; Becker 1976).

Pluralism in contrast holds that there is not just a single value, that human actions might have multiple final ends, and that there are multiple things that are good in and of themselves such as love, knowledge, piety, and justice (Haslanger 2023, 2022). Pluralism recognizes that values are culturally relative and fundamentally vary across societies (Schmer-Galunder 2026). Thus, a significant development in alignment research recently has been what we call "the pluralist turn," which seeks to develop approaches to accommodate diverse viewpoints and values (Sorensen et al. 2024; Klassen, Alamdari, and McIlraith 2024; Kasirzadeh 2024a).

We incorporated questions about the quantifiability and measurability of values, and assessed papers for whether they view values as measurable. The assumption that values are the equivalent of maximizing a utility function allows for the ostensible quantification of human values. From anthropology we adapted the evaluative distinction between **thin** and **thick** conceptions of values. Thick evaluative concepts and descriptions of human values view them not as mere abstractions, but as involving intentional, purposive detail that helps us understand those activities in their cultural and social contexts (Geertz 2008).

Given the central role that economic theory plays in AI alignment, we also annotated papers for whether the alignment framework was rooted in utility theory.

Rubric questions used to annotate the papers

1. Does the paper define or describe **value**?
2. Does the paper's theory of value subscribe to value **monism** or **pluralism**? Monism holds that there is a single, intrinsic dimension to value, while pluralism states that there is more than one dimension or consideration of value, which cannot be flattened.
3. Does the paper consider values **measurable** or **immeasurable**? Measurable values can be quantified, operationalized, and recorded as data with negligible loss of

precision, while immeasurable values cannot be precisely operationalized and quantified.

4. Does the paper consider values as revealed **preferences**, or prescribed **principles**? Preferences are measured from behavioral decisions, and may involve some process of ranking preferred actions, outcomes, or things, while principles-based values correspond to abstract ideals or standards, e.g. valuing beauty, truth, knowledge, human rights declarations, etc.
5. Is the approach to values **abstract** or **concrete**? Abstract approaches can be theoretical, idealized, philosophical, name no specific values in the paper but present a set of principles for extracting them, whereas concrete approaches may pertain to a specific existing system, culture, or group.
6. Are values apparent in **individual** interactions, or formed from a **collective** of agents? That is, are values observed, measured, or determined from individual or one-on-one interactions, or do they arise over communities, cultures, nations, or groups?
7. Are values **dynamic** or **static**? Dynamic values can change over time or in different situations, while static values do not change (or the change of values is not explained).
8. Is the characterization of values **thin** or **thick**? We take a "thin" characterization of value to be entirely instrumental, evaluative and calculative, while a "thick" characterization is descriptive and situated in a societal context.
9. Does the paper assume or explicitly state that values can emerge in AI autonomously from its creators?
10. **Does the paper take the position that AI should follow human values**? A normative question which asks if it is morally justified, ethically correct, and/or strategically important for the field to develop AI systems which "uphold" values, resembling the processes by which humans uphold values (according to the paper's implicit or explicit theory of value).
11. Does the paper identify **which humans'** values the research is aligning to?
12. When preferences are elicited from human annotators, does the paper provide information about the **sample size of raters**?
13. Does the paper use **utility maximization** as an approach to value alignment?

Annotation process We ensured that each paper was annotated by two annotators among the five of the co-authors of this paper. We used a Google form with the rubrics, and filled them in after reading each paper. We divided the papers into groups so that a subset of annotators read approximately fifty papers each. After an initial pass at reading the collected papers, we isolated disagreements and discussed them through resolutions and consensus. This was, however, not done for every disagreement and some disagreement remains among the annotations. However, we achieved a high degree of inter-annotator agreement, which averages overall around 85%.

Qualitative and interpretive analysis

In addition to scoring the papers according to our rubric for a quantitative estimate of the prevalence of the concepts, we also collected quotes from each paper if they addressed human values in non-mathematical terms. We did this in order to characterize the beliefs about human values stated, where applicable, by the authors of the papers we analyze.

Findings

Quantitative results

Inter-rater agreement

In Table 1 we report measures of inter-rater agreement.

We present visualizations of the quantitative results in Figure 1 and Figure 2.

Qualitative Analysis

We found that **79% of the papers in our dataset neither defined nor described what they mean by "values"** despite ostensibly seeking to align models with values. In many papers, it was common to encounter the word "values" used interchangeably with "preferences" and "intentions," and as we address below, elision between values and preferences has become common in this subfield. In one rare example, values were also interchangeably used with "ideologies" (Kirk et al. 2023b). In another, researchers trained a model on a corpus they built using Chinese laws and morality texts (Xu et al. 2023).

Despite largely not defining what values mean, 90% of the papers treat values as measurable. This most often originated with different means of determining human preferences. Typically, researchers created datasets where LLMs were fed a prompt, generated two responses, and used either human raters, other LLMs (sometimes referred to as "autoraters") (Lu et al. 2024; Zheng et al. 2024; Chakraborty et al. 2024; Richemond et al. 2024), or a mix of both (Yu et al. 2023; Liu, Ge, and Zhu 2024; Wang et al. 2024a) to evaluate which of the two options better adhered to researchers' choice of alignment criteria; in other cases, one response might be treated as 'chosen' while the other is 'rejected.' Some papers drew from preference datasets and survey data that researchers who were not the authors of these papers generated (Zheng et al. 2024; Chakraborty et al. 2024; Lou et al. 2024; Durmus et al. 2023; Zhao, Dang, and Grover 2023; Miranda et al. 2024), while others identified preferences they sought to optimize for including "conciseness..being humorous, philosophical, sycophantic, helpful, concise, creative, formal, expert, pleasant, and uplifting" (Zhong et al. 2024).

Preferences have become such a dominant framework that **82% of the papers treat preferences as a stand-in for "values."** By contrast, 9% treated principles as a basis of values, for instance drawing from social choice theory (Siththaranjan, Laidlaw, and Hadfield-Menell 2023; Chakraborty et al. 2024). In papers where researchers created preference datasets through reliance on human annotators to rate LLMs' paired responses to prompts, only 29 of the 94 papers sampled provided a number for how many

Table 1: Inter-Rater Reliability Statistics

Question	Mean κ	Mean PABAK	α (Alpha)	Agreement	# Pairs
Are values measurable?	0.541	0.775	0.269	88.8%	4
Revealed preferences or prescribed principles?	0.467	0.710	0.373	85.5%	4
Sources for determining values	0.411	0.167	0.023	58.3%	4
Abstract or Concrete approach	0.378	0.619	0.235	81.0%	4
Uses utility maximization?	0.365	0.685	0.428	84.2%	4
Can values emerge in AI?	0.350	0.386	0.331	69.3%	4
States which humans' values?	0.302	0.601	0.219	80.1%	4
Which humans' values?	0.299	0.736	0.000	86.8%	4
Think vs Thick.	0.239	0.056	0.073	52.8%	4
Does the paper define value?	0.221	0.501	0.247	75.1%	4
Monism vs Pluralism	0.218	0.079	0.349	54.0%	4
Values Static or Dynamic?	0.133	0.021	0.204	51.0%	4
Should AI follow human values?	0.090	0.138	-0.015	56.9%	4
Size of rater pool	0.056	0.275	0.348	63.7%	4
Individual vs Collective interactions	0.028	0.215	0.286	60.8%	4

Notes: PABAK = $2 \times$ Agreement $- 1$ (useful when class distributions are skewed); Krippendorff's α considers all raters simultaneously and handles missing data. Limitations of quantitative analysis: we report raw agreement, Cohen's kappa, Krippendorff's α and Bias-Adjusted Kappa (PABAK). Given that, we do expect high agreement on the binary yes/no questions, but some disagreement on extent or categorical questions; these results could indicate issues with design or an attribute of the setting in which we are not the research subjects. Some subjectivity may not be fully specified by the rubric questions. We may have different criteria for what counts as answers to our questions treating papers as the annotation item. We are not claiming that we have developed an objective survey (Aroyo and Welty 2015). We as the annotators are not the research subjects, which is one methodological assumption baked into inter-rater reliability; another is that the raters are interchangeable, which we are not.

people comprised the rater pools. The number of raters tended to be < 100 people, with the exception of one paper that used a 13,000-person rater pool (Köpf et al. 2023). When human rater pools were used, 77% of that subset of papers **did not state the size of the rater pool**.

A growing debate within alignment is the extent to which pluralism of values can be achieved. One paper acknowledged that "different humans have different values, as it is nearly impossible to train a new large language model from scratch for individual preference" (Zhou et al. 2024). Despite this, most papers took individual rather than collective approaches to values. Rather than selecting values that are already practiced across societies, states, and other collectives, they rely on individual user interactions with LLMs (or LLM-to-LLM interactions), and in one case on collectives of 4-5 people (Bakker et al. 2022). In addition, **53% of papers treated values as static**, whether implicitly or through direct acknowledgment of doing so despite recognizing that values evolve (Liu et al. 2023).

We found 66% of papers presented "thin" descriptions and examples of what they meant by values. By contrast, "thick" descriptions (which accounted for 3 percent) draw from relational, observed values that are difficult to cleave apart from the lived context of the people who enact these values (Geertz 2008).

Likewise, **91% of the papers did not refer to specific groups of humans whose values should be followed**, underscoring the taken-for-granted idea of values as universal. One paper noted "We claim that the approach described is agnostic to the ethical paradigm, the user's preferences, and the legal or social framework, provided we can supply enough feedback" (Leike et al. 2018), while another professed "No need to solve human values. We assume we do not need to solve hard philosophical questions of human values and value aggregation before we can align a superhuman researcher model well enough that it avoids egregiously catastrophic outcomes" (Burns et al. 2023). Another reckoned with the difficulties of treating preference datasets as representative: "This procedure aligns the behavior of GPT-3 to the stated preferences of a specific group of people (mostly our labelers and researchers), rather than any broader notion of 'human values'" (Ouyang et al. 2022).

Finally, we noticed a trend in which a few papers suggest that AI models can develop their own value systems (Ngo, Chan, and Mindermann 2022) and "internal meta-objectives" (Phelps and Ranson 2023a). Some researchers have begun to refer to models' "self-alignment" (Guo et al. 2024; Li et al. 2024) described as when "a new paradigm emerges where LLMs can achieve value alignment by themselves... transforming an unaligned LLM into one adhering to societal norms, independently of external resources" (Pang et al. 2024).

Discussion

Our systematic review of 94 research papers on AI value alignment reveals how the field is erasing human inputs from an endeavor whose success rests on purportedly steering AI systems to respect human values. Values are core components of how societies organize themselves; any attempt at value alignment is inextricable from society (Grae-

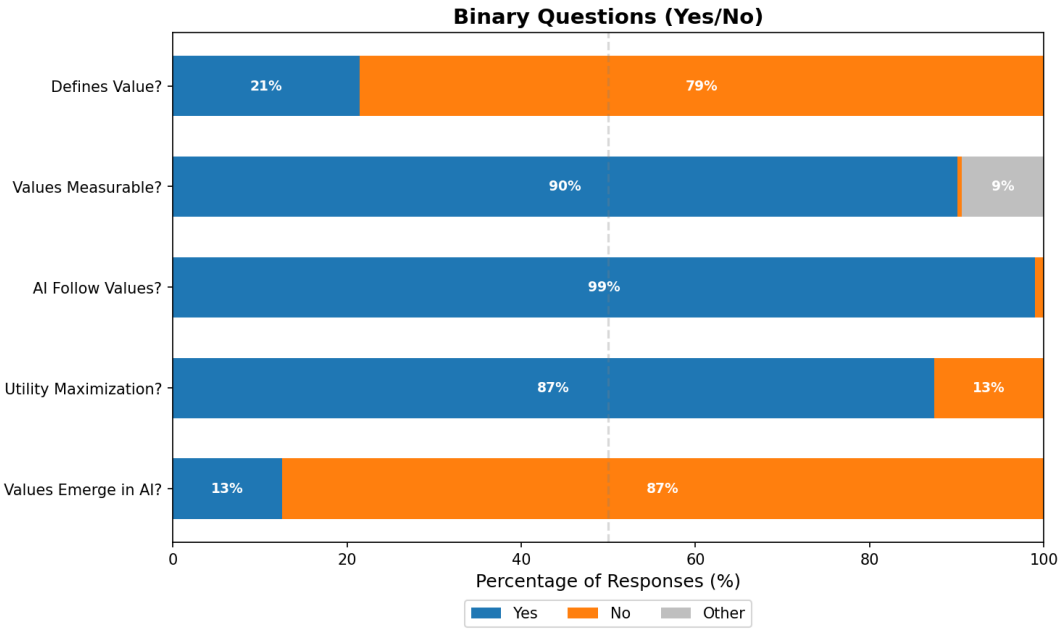

Figure 1: Summary of answers to our rubric aggregated across all raters for binary yes/no answers.

ber 2001; Haslanger 2023; Anderson 1995; Zhi-Xuan et al. 2024; Birhane et al. 2022). Recent social philosophy argues that societies are complex systems – or clusters of interacting systems – that reproduce themselves: their hierarchies, culture, and structures. These systems are also dynamic and constantly evolving (Haslanger 2023). Yet we observe how research in this field abstracts away questions of whose values are represented, or what qualities constitute pluralism and dynamism in values, instead relegating these fundamental, unavoidable concerns to be taken up by others (if at all). The risk of repeatedly delaying explorations of what values are and how they can be accountably represented is that certain practices of design, deployment, and evaluation will ossify. At worst, this may result in normalizing widespread use of AI systems that have been trained on a narrow conception of value and claiming that this is the best the field can offer.

The pressure to neglect these essential debates is reflected in a subset of papers that treat alignment as a step toward achieving "artificial superintelligence" (Kim et al. 2024), or as an essential component of averting catastrophic and existential risks from AI (Shen et al. 2023; Ngo, Chan, and Mindermann 2022; Burns et al. 2023). This typifies the turn toward what we call "alignment without humans," or treatment of humans as an unreliable source of information on their own preferences (Miranda et al. 2024; Wang et al. 2024a), as too expensive (Wang et al. 2024c; Zheng et al. 2024; Liu, Ge, and Zhu 2024) and time-consuming (Dai et al. 2023) to use, therefore justifying the use of synthetic means of preference elicitation, feedback, and evaluations (Hong et al. 2024; Cui et al. 2024; Tian et al. 2024). We observe studies that rely on no human input for pre-determining value alignment criteria (Mei et al. 2024; Chen et al. 2024), often through the use of simulated humans (Chakraborty et al. 2023; Poddar et al. 2024; Liu et al. 2023). Perhaps in part due to the desire to race toward solving "the alignment problem," one trend that arose from the corpus of papers is the dominance of Anthropic's Helpful, Harmless, Honest (HHH) approach. This ranged from papers that explicitly used the HH-RLHF dataset for training and/or evaluations (Yu et al. 2023; Lou et al. 2024; Huang et al. 2023; Zheng et al. 2024; Zeng et al. 2024), to others that name HHH as a guiding principle (Ding, Li, and Zhang 2024; Gao et al. 2024; Lu et al. 2024; Shen et al. 2023; Shi et al. 2024).

Graeber's critique of applying economic theory to diverse cultures is apt for AI value alignment: "It [economics] also has the advantage of joining an extremely simple model of human nature with extremely complicated mathematical formulae that non-specialists can rarely understand, much less criticize." (Graeber 2001). Our review of the literature reveals how RLHF and other technical alignment methods likewise join extremely simplistic models of human nature with complicated mathematical formulae that are difficult for non-specialists to understand or critique.

Even though the majority of papers in our dataset do not define what they mean by values, adopting the framework of economic preference optimization to align models necessarily requires a commitment to a set of beliefs about what human values are. At a minimum this entails the belief that values are individual, rational, monist, static, maximizing utility functions, and abstract. The idea is that an agent that obeys some minimal conditions of rationality can be modeled as if the agent has an ordered set of preferences along with some probabilistic beliefs about what states of the world and itself will maximize the agent's utility (however defined) (Shea 2018). This particular set of assumptions about human nature and human values, drawn largely

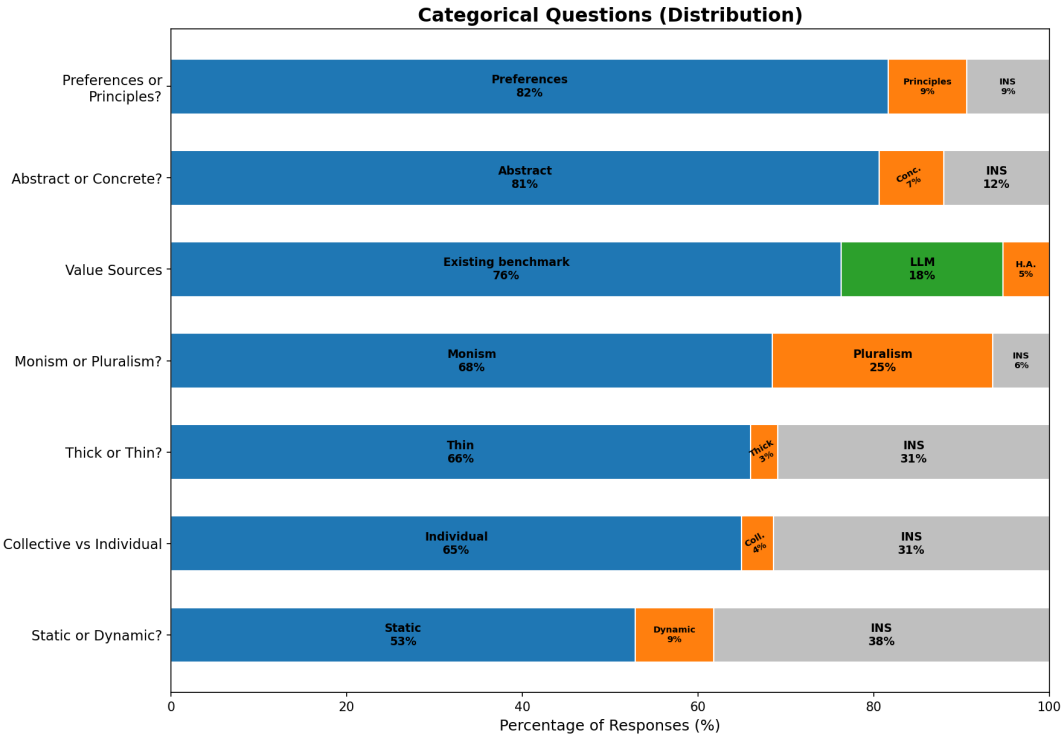

Figure 2: Summary of answers to our rubric aggregated across all raters for categorical questions. INS = ”information not stated”

from economic decision theory and rational choice theory, has been empirically and theoretically challenged from multiple perspectives within and outside of economics (Chibnik 2011; Graeber 2011, 2001; Gigerenzer 1997; Frank, Gleiser, and Thompson 2024) and AI alignment (Bakker et al. 2022; Kasirzadeh 2024a; Zhi-Xuan et al. 2024). Rational or social choice theory do not describe the values or behavior of real human beings or cultures, but only the behavior of highly abstract, reductionist, and idealized fictional agents (Frank, Gleiser, and Thompson 2024).

Rather than asking ”what do human beings actually care about?”, rational choice theory asks - and presumes there is a normatively correct answer to - ”what should human beings do in order to make the utility-maximizing decision?” This meta-theoretical commitment risks foreclosing alternative conceptions of human values, and limits the ability of AI systems to align to diverse cultural values (Prabhakaran et al. 2022; Khan, Casper, and Hadfield-Menell 2025; Gabriel 2020). Poddar *et al.* point out, ”Current RLHF approaches rely on a prescriptive set of values curated by a small set of AI researchers. Moreover, they typically assume that all end-users share the same set of values. Given the concerning lack of diversity in AI, it is clear that this approach cannot account for the range of social, moral, and political values that inform preferences in human populations” (Poddar et al. 2024). We call on the field to take seriously the fundamental weaknesses of utility maximization, rational choice theory and reductionist economic frameworks in AI alignment.

Utility maximization Economists and philosophers have generally assumed that people are trying to maximize *something*: money, or love, or sometimes something else (most often, expected utility) with their choices (Glimcher 2022; Graeber 2001).Utility maximization treats all human behavior as the maximization of expected utility from a stable set of preferences, and holds that humans accumulate an optimal amount of information in *markets* (Chibnik 2011; Russell and Norvig 1995; Becker 1976). Thus, preferences are seen as being derived from a single utility function that each individual is somehow computing in their brain (Gigerenzer 1997). Taken to its extreme, utility theory holds that our behavior is entirely determined by maximizing this utility function, and preferences are assumed to not change substantially over time, nor to be very different between people of different socioeconomic classes, societies, and cultures (Becker 1976). Finally, AI alignment assumes that an individual’s reward function (and therefore their values), or preferences about the future, can be inferred, or approximated, by observing how humans behave by collecting data on how humans interact with AI outputs (Hadfield-Menell et al. 2016).

Despite the empirical inadequacy of this ambiguous economic approach to AI and value alignment, creating an artificial agent that embodies economic theory has been the almost unquestioned north star in the field until recently (Zhi-Xuan et al. 2024; Gabriel 2020; Russell 2019). Influential work on value alignment explicitly argues that alignment should be formulated as a cooperative and interac-

tive *reward maximization process* (Hadfield-Menell et al. 2016), with most alignment work implicitly adopting this approach (Christiano et al. 2017b; Ouyang et al. 2022; Wang et al. 2024b) the vast majority of technical approaches to alignment from our sample of papers implementing some form of utility maximization. These often unstated philosophical commitments to an economic worldview determine the methodology that alignment researchers use to measure and model human values. We make this often implicit metatheorectical commitment explicit so that it can be critiqued and improved (Maxwell 2005).

We also evaluated our sample of papers for whether they adopted a framework of **utility maximization**, discovering that 87 % of papers implicitly or explicitly use utility maximization as part of the technical framework to align models. Adopting the framework of utility maximization necessarily involves normative assumptions that are value-laden (Graeber 2001). However, given that utility maximization is an inherent part of reinforcement learning from human feedback (RLHF) and reward modeling it is not surprising that the majority of alignment papers in our sample adopt this framework. This also stems from framing AI alignment in terms of the "principal-agent" framework from economic theory (Phelps and Ranson 2023b; Stanczak et al. 2025).

Moreover, utility maximization is cited as a property of models, e.g., (Mazeika et al. 2025) notes "it was shown that recent LLMs have structurally coherent, broad value systems. As they become more capable, their value systems increasingly conform to the axioms of utility theory, meaning they can be described as maximizing a utility function." Given that models are trained and fine-tuned according to the axioms of rational choice theory, should it be surprising that they behave according to utility theory?

A small number of the papers recognize the inherent limitations of utility maximization. (Phelps and Ranson 2023b) points out: "Rather than seeking to impose a monolithic utility function on artificial agents, we propose a strategy of reducing information asymmetry and aligning interests through external incentives, much like the approach used in traditional economic solutions to principal-agent problems. "Furthermore, (Poddar et al. 2024) writes, "These insights suggest that human preferences are not derived from a single utility function, but are affected by unobserved, hidden user context."

One of the starting points of our meta-review is Zhi-Xuan et al. (2024)'s critique of the reliance on eliciting preferences as representations of human values, in which they ask: "What would AI alignment look like if it took these challenges seriously? It would move away from naive rational choice models of human decision making, towards richer models that include how we evaluate, commensurate, and act upon our values in boundedly rational ways. It would no longer take for granted expected utility theory, and instead explore systems for reasoning about the normativity of our preferences and values."

Alternative conceptions of value

> *"It's values all the way down"* - Alondra Nelson (Nelson 2023)

The narrow economic conception of human values that determines the methodology and philosophy of alignment is by no means the only way to understand what humans value. As anthropologists point out, for 99 % of humanity's history we did not live in market economies (Graeber 2011). Economic theory implicitly posits that we've always been capitalists and that capitalism is somehow natural (Hornborg 1998). We argue that even recent work on pluralistic alignment is still fundamentally rooted in the orthodox economic framework.

However, this is a distorted view of both societies and individual humans. Part of the challenge of rethinking alignment's approach to value comes from the social positioning of the field that 'doing anything is better than nothing,' without recognizing that over-utilizing narrow disciplinary approaches has the potential to make matters worse (Dai 2025). A fundamental limitation of existing alignment methods is that they reinforce shallow behavioral dispositions rather than endowing LLMs with a genuine capacity for normative deliberation (Millière 2025). We find that the abstractions and idealizations which alignment postulates to describe human values become confused for real phenomena.

Papers in our sample did in some cases engage directly with alternatives to utility theory, for example Huang et.al., (2025) who state, "We are interested in values not as abstract entities, but as operational priorities that influence how the system navigates its possible space of outputs." We expand others' calls to expand alignment approaches (Casper et al. 2023; Lindström et al. 2024; Zhi-Xuan et al. 2024) with our exploration of psychological and anthropological theories of value.

Cognitive or psychological theories of value Papers in our dataset referenced psychological theories of value that are accepted in mainstream psychology literature, such as Schwartz's theory of basic values (Sierra et al. 2021; Schwartz 2012). Schwartz theorizes value as a set of beliefs, closely linked to emotions, which inform the selection of actions. He further claims that there is a fixed set of "basic values" or value categories, which are **universal** across cultures, and tend to remain fixed over time within an adult individual. These values are conceptualized as guiding principles which can come into conflict, but are ultimately resolved through a hierarchy of importance. This leads to our characterization of the psychological theory of value as ultimately monistic. This individualistic view of values is challenged by empirical psychological and ethnographic work which shows a much more complex, situated, holistic and dynamic view of values (Chibnik 2011; Anderson 1995; Haslanger 2023). This field is also complex, and we encourage engagement with existing debates in the cross-cultural study of values (Schwartz 2012).

Anthropological, situated theories of value The field with perhaps the most theoretical work and empirical data on human values - anthropology - has to date been ignored in AI alignment (cf. (Schmer-Galunder 2026). Graeber suggests that what people call "value" may be thought of as "the way people represent the importance of their own actions to themselves, as reflected in one or another socially

recognized form" (Graeber 2001). Economic theory's view of human nature is cynical, assuming that humans are entirely self-interested and will calculate the most effective way to obtain what they desire (Becker 1976). These are values derived from living in market-based societies (Hornborg 2014). However, markets are not natural phenomena that arise spontaneously out of large complex societies as the economic narrative goes, but rather they require the enforcement of state violence to impose property rights, coercion, and the rule of law (Graeber 2011; Mau 2023).

Economic anthropology recognizes that economic notions of value, articulated in objective monetary representations and utility theory, and ethical notions of value, articulated in moral terms, are "inextricably connected" (Lambek 2013). Graeber (Graeber 2001) outlines three ways in which values have been theorized in the social sciences that go beyond the narrow conception of value in the economic sense:

1. "values" in the sociological sense: conceptions of what is ultimately good, proper, or desirable in human life
2. "value" in the economic sense: the degree to which objects are desired, particularly, as measured by how much others are willing to give up to get them
3. "value" in the linguistic sense, or how meaning is derived from the differences in the usage of words.

However, in our dataset it is unusual for researchers to be specific about which humans, which values, and which cultures they target for alignment. We follow (Arzberger et al. 2024), who argue that values acquire situated meaning; hence, their concrete interpretations, means of actualization, and hierarchies vary between people depending on the situations in which they guide the notion of what people really value. This is closely aligned with situated-constructivist views of emotions, which see emotions not as basic universal categories, but as arising in specific socially- and culturally- salient circumstances. In other words, emotions are a category populated with highly variable instances (Barrett 2017). Thus we can analogously we think of human values as a category populated with highly variable instances tied to specific cultural circumstances.

Limitations

One limitation of this work is the relatively small sample size of papers in our dataset compared with the vast number of research papers on this topic. This is especially challenging given the widespread practice of rapidly uploading research to arXiv, avoiding peer-review (Gyevnar and Kasirzadeh 2025). However, we attempt to correct for this by estimating the relative impact of our dataset on the field using both quantitative and qualitative measures. While we aimed to gather a representative subsample of available papers on AI value alignment, we make no claims to the completeness or comprehensiveness of the sampled papers. The sample size was also limited due to time constraints on our small team of co-authors. We also acknowledge that the scoring rubric has many limitations. The questions are highly nuanced and difficult to boil down to a binary or even ternary answer. We considered a multi-point Likert scale, which was ultimately abandoned due to concerns about calibrating relative scales.

Implications for future work

We suggest building on our analysis to apply concepts from humanistic social sciences such as anthropology and psychology to operationalize human values in AI in a more representative and holistic way that explicitly accounts for: concreteness/abstraction, the thick/thin distinction, distinctions between monism and pluralism, and embracing a view of values not as abstract universal things human possess, but as constructed in specific social and cultural contexts (Kasirzadeh 2024a; Sorensen et al. 2024; Benkler et al. 2023; Arzberger et al. 2024).

To deepen the critique of "preferentism," we also recommend an examination of the content of prompts and paired responses in preference training datasets, as many contain largely nonsensical conversations with an LLM and involve choosing the less absurd of two unrealistic examples in each pair. Communicating the shaky foundations of these unquestioned defaults is of particular value to informing policymakers and the public about the limits of value alignment. Moreover, it can contribute to more effective sociotechnical and policy measures derived from realistic conceptions of AI's risks.

Conclusion

We began with an examination of the theory of value(s) to which AI value alignment research subscribes, drawing from interdisciplinary philosophy and social scientific research on human values to develop a schema for answering this question. We argued that AI alignment lacks a concrete theory of human values and implicitly adopts beliefs about human values rooted in economic theory. Through our review, we have shown that the assumptions guiding technical AI alignment practices limit the possible range of human values with which AI can be aligned by translating alignment into utility theory. As Lisa Feldman Barrett argues, "Even a scientist who believes they are purely driven by data is doing philosophy" (Barrett and Theriault 2025). Tech companies and AI value alignment researchers enact philosophy every day as they engage in value alignment work when choosing training data, setting up model architectures, and especially when running RLHF experiments. Having laid bare AI value alignment's philosophical commitments, we invite researchers from a wider variety of disciplines to bring greater specificity and under-explored perspectives into the debate on whether and how AI can address human values.

Generative AI Usage Statement

Generative AI was not used in the research or writing of this publication.

References

Ahmed, S.; Jaźwińska, K.; Ahlawat, A.; Winecoff, A.; and Wang, M. 2024. Field-building and the epistemic culture of AI safety. *First Monday*.

Alvarado, R. 2023. AI as an epistemic technology. *Science and Engineering Ethics*, 29(5): 32.

Amirloo, E.; Fauconnier, J.-P.; Roesmann, C.; Kerl, C.; Boney, R.; Qian, Y.; Wang, Z.; Dehghan, A.; Yang, Y.; Gan, Z.; et al. 2024. Understanding alignment in multimodal llms: A comprehensive study. *arXiv preprint arXiv:2407.02477*.

Anderson, E. 1995. *Value in ethics and economics*. Cambridge, Mass.: Harvard Univ. Press, 2. print edition. ISBN 9780674931909 9780674931893.

Anthropic. 2025. Anthropic Alignment Research Team.

Aroyo, L.; and Welty, C. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1): 15–24.

Arvan, M. 2024. 'Interpretability'and 'alignment'are fool's errands: a proof that controlling misaligned large language models is the best anyone can hope for. *AI & SOCIETY*, 1–16.

Arzberger, A.; Buijsman, S.; Lupetti, M. L.; Bozzon, A.; and Yang, J. 2024. Nothing comes without its world–practical challenges of aligning llms to situated human values through RLHF. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 61–73.

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. ArXiv:2204.05862.

Bakker, M.; Chadwick, M.; Sheahan, H.; Tessler, M.; Campbell-Gillingham, L.; Balaguer, J.; McAleese, N.; Glaese, A.; Aslanides, J.; Botvinick, M.; et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35: 38176–38189.

Barrett, L. F. 2017. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1): 1–23.

Barrett, L. F.; and Theriault, J. 2025. What's Real? A Philosophy of Science for Social Psychology. *The handbook of social psychology*.

Becker, G. S. 1976. *The economic approach to human behavior*, volume 803. University of Chicago press.

Benkler, N.; Mosaphir, D.; Friedman, S.; Smart, A.; and Schmer-Galunder, S. 2023. Assessing llms for moral value pluralism. *arXiv preprint arXiv:2312.10075*.

Bergman, S.; Marchal, N.; Mellor, J.; Mohamed, S.; Gabriel, I.; and Isaac, W. 2024. STELA: a community-centred approach to norm elicitation for AI alignment. *Scientific Reports*, 14(1): 6616.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184. Seoul Republic of Korea: ACM. ISBN 978-1-4503-9352-2.

Burns, C.; Izmailov, P.; Kirchner, J. H.; Baker, B.; Gao, L.; Aschenbrenner, L.; Chen, Y.; Ecoffet, A.; Joglekar, M.; Leike, J.; Sutskever, I.; and Wu, J. 2023. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision.

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*.

Chakraborty, S.; Qiu, J.; Yuan, H.; Koppel, A.; Huang, F.; Manocha, D.; Bedi, A. S.; and Wang, M. 2024. MaxMin-RLHF: Alignment with Diverse Human Preferences.

Chakraborty, S.; Singh, A.; Bhaskar, A.; Tokekar, P.; Manocha, D.; and Bedi, A. S. 2023. REBEL: Reward Regularization-Based Approach for Robotic Reinforcement Learning from Human Feedback.

Chaudhari, S.; Aggarwal, P.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; Narasimhan, K.; Deshpande, A.; and da Silva, B. C. 2024. RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. *arXiv preprint arXiv:2404.08555*.

Chen, Z.; Zhou, K.; Zhao, W. X.; Wang, J.; and Wen, J.-R. 2024. Low-Redundant Optimization for Large Language Model Alignment.

Chibnik, M. 2011. *Anthropology, Economics, and Choice*. University of Texas Press. ISBN 978-0-292-72676-5.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017a. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017b. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Cui, G.; Yuan, L.; Ding, N.; Yao, G.; Zhu, W.; Ni, Y.; Xie, G.; Liu, Z.; and Sun, M. 2024. UltraFeedback: Boosting Language Models with High-quality Feedback.

Dai, J. 2025. The Artificiality of Alignment.

Dai, J.; Pan, X.; Sun, R.; Ji, J.; Xu, X.; Liu, M.; Wang, Y.; and Yang, Y. 2023. Safe RLHF: Safe Reinforcement Learning from Human Feedback.

Ding, Y.; Li, B.; and Zhang, R. 2024. ETA: Evaluating Then Aligning Safety of Vision Language Models at Inference Time.

Dobbe, R. 2022. System safety and artificial intelligence. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1584–1584.

Durmus, E.; Nguyen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; Lovitt, L.; McCandlish, S.; Sikder, O.; Tamkin,

A.; Thamkul, J.; Kaplan, J.; Clark, J.; and Ganguli, D. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models.

Frank, A.; Gleiser, M.; and Thompson, E. 2024. *The blind spot: why science cannot ignore human experience*. Cambridge, Mass: The MIT press. ISBN 978-0-262-04880-4.

Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3): 411–437.

Gao, C.; Wu, Y., Siyuan Huang; Chen, D.; Zhang, Q.; Fu, Z.; Wan, Y.; Sun, L.; and Zhang, X. 2024. HonestLLM: Toward an Honest and Helpful Large Language Model.

Gebru, T.; and Torres, É. P. 2024. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*.

Geertz, C. 2008. "Thick Description: Toward an Interpretive Theory of Culture". In *The Cultural Geography Reader*. Routledge. ISBN 978-0-203-93195-0. Num Pages: 11.

Gigerenzer, G. 1997. Bounded rationality: Models of fast and frugal inference. *Swiss Journal of Economics and Statistics*, 133(2/2): 201–218.

Gillespie, T.; Shaw, R.; Gray, M. L.; and Suh, J. 2024. AI Red-Teaming is a Sociotechnical System. Now What? *arXiv preprint arXiv:2412.09751*.

Glimcher, P. W. 2022. Efficiently irrational: deciphering the riddle of human choice. *Trends in cognitive sciences*, 26(8): 669–687.

Graeber, D. 2001. *Toward an Anthropological Theory of Value: The False Coin of Our Own Dreams*. Palgrave-Macmillan.

Graeber, D. 2011. *Debt: the first 5,000 years*. Brooklyn, N.Y: Melville House. ISBN 978-1-933633-86-2.

Guest, O. 2026. What Does 'Human-Centred AI'Mean? *Behavioral Sciences*, 16(4): 583.

Guo, H.; Yao, Y.; Shen, W.; Wei, J.; Zhang, X.; Wang, Z. W.; and Liu, Y. 2024. Human-Instruction-Free LLM Self-Alignment with Limited Samples.

Gyevnar, B.; and Kasirzadeh, A. 2025. AI safety for everyone. *Nature Machine Intelligence*, 1–12.

Hadfield-Menell, D.; Dragan, A.; Abbeel, P.; and Russell, S. 2016. Cooperative inverse reinforcement learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 3916–3924. Red Hook, NY, USA: Curran Associates Inc. ISBN 978-1-5108-3881-9.

Haslanger, S. 2022. Failures of Methodological Individualism: The Materiality of Social Systems. *Journal of Social Philosophy*, 53(4).

Haslanger, S. 2023. Situated Knowledge and Situated Values.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028): 147–154.

Hong, I.; Li, Z.; Bukharin, A.; Li, Y.; Jiang, H.; Yang, T.; and Zhao, T. 2024. Adaptive preference scaling for reinforcement learning with human feedback. *Advances in Neural Information Processing Systems*, 37: 107249–107269.

Hooker, S. 2025. On the Slow Death of Scaling. *Available at SSRN 5877662*.

Hornborg, A. 1998. Ecological embeddedness and personhood: Have we always been capitalists? *Anthropology today*, 14(2): 3–5.

Hornborg, A. 2014. Technology as Fetish: Marx, Latour, and the Cultural Foundations of Capitalism. *Theory, Culture & Society*, 31(4): 119–140. Publisher: SAGE Publications Ltd.

Huang, Y.; Gupta, S.; Xia, M.; Li, K.; and Chen, D. 2023. Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation.

Institute, U. A. S. 2025. UK AI Security Institute Alignment Project.

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

Johnson, G. M. 2023. Are Algorithms Value-Free? *Journal Moral Philosophy*, 21(1-2): 1–35.

Kasirzadeh, A. 2024a. Plurality of value pluralism and AI value alignment.

Kasirzadeh, A. 2024b. Two types of AI existential risk: decisive and accumulative. *arXiv preprint arXiv:2401.07836*.

Khan, A.; Casper, S.; and Hadfield-Menell, D. 2025. Randomness, not representation: The unreliability of evaluating cultural alignment in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2151–2165.

Khlaaf, H. 2023. Toward comprehensive risk assessments and assurance of ai-based systems. *Trail of Bits*, 7.

Kim, H.; Yi, X.; Yao, J.; Lian, J.; Huang, M.; Duan, S.; Bak, J.; and Xie, X. 2024. The Road to Artificial SuperIntelligence: A Comprehensive Survey of Superalignment.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2023a. The empty signifier problem: Towards clearer paradigms for operationalising" alignment" in large language models. *arXiv preprint arXiv:2310.02457*.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2023b. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback.

Klassen, T. Q.; Alamdari, P. A.; and McIlraith, S. A. 2024. Pluralistic Alignment Over Time.

Köpf, A.; Kilcher, Y.; Von Rütte, D.; Anagnostidis, S.; Tam, Z. R.; Stevens, K.; Barhoum, A.; Nguyen, D.; Stanley, O.; Nagyfi, R.; et al. 2023. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36: 47669–47681.

Korsgaard, C. M. 1986. Aristotle and Kant on the Source of Value. *Ethics*, 96(3): 486–505.

Lambek, M. 2013. The value of (performative) acts. *HAU: Journal of Ethnographic Theory*, 3(2): 141–160.

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

Li, X.; Yu, P.; Zhou, C.; Schick, T.; Zettlemoyer, L.; Levy, O.; Luke, Z.; Weston, J.; and Lewis, M. 2024. Self-Alignment with Instruction Backtranslation.

Lindström, A. D.; Methnani, L.; Krause, L.; Ericson, P.; de Troya, Í. M. d. R.; Mollo, D. C.; and Dobbe, R. 2024. AI Alignment through Reinforcement Learning from Human Feedback? Contradictions and Limitations. *arXiv preprint arXiv:2406.18346*.

Liu, J.; Ge, D.; and Zhu, R. 2024. Reward Learning From Preference With Ties.

Liu, R.; Yang, R.; Jia, C.; Zhang, G.; Zhou, D.; Dai, A. M.; Yang, D.; and Vosoughi, S. 2023. Training Socially Aligned Language Models on Simulated Social Interactions.

Lou, X.; Zhang, J.; Xie, J.; Liu, L.; Yan, D.; and Huang, K. 2024. SPO: Multi-Dimensional Preference Sequential Alignment With Implicit Reward Modeling.

Lu, K.; Yu, B.; Huang, F.; Fan, Y.; Lin, R.; and Zhou, C. 2024. Online Merging Optimizers for Boosting Rewards and Mitigating Tax in Alignment.

Mau, S. 2023. *Mute compulsion: A Marxist theory of the economic power of capital*. Verso books.

Maxwell, N. 2005. *The comprehensibility of the universe: a new conception of science*. Oxford: Clarendon press. ISBN 978-0-19-926155-0.

Mazeika, M.; Yin, X.; Tamirisa, R.; Lim, J.; Lee, B. W.; Ren, R.; Phan, L.; Mu, N.; Khoja, A.; Zhang, O.; et al. 2025. Utility engineering: Analyzing and controlling emergent value systems in ais. *arXiv preprint arXiv:2502.08640*.

Mei, L.; Liu, S.; Wang, Y.; Bi, B.; Yuan, R.; and Cheng, X. 2024. HiddenGuard: Fine-Grained Safe Generation with Specialized Representation Router.

Mill, J. S. 2016. Utilitarianism. In *Seven masterpieces of philosophy*, 329–375. Routledge.

Millière, R. 2025. Normative conflicts and shallow AI alignment: R. Millière. *Philosophical Studies*, 1–44.

Miranda, L. J. V.; Wang, Y.; Elazar, Y.; Kumar, S.; Pyatkin, V.; Brahman, F.; Smith, N. A.; Hajishirzi, H.; and Dasigi, P. 2024. Hybrid Preferences: Learning to Route Instances for Human vs. AI Feedback.

Nauer, C.; and van Doren, Z. 2025. EXISTENTIAL RISK FROM AI.

Nelson, A. 2023. FAccT'23 Keynote: "Thick Alignment".

Ngo, R.; Chan, L.; and Mindermann, S. 2022. The Alignment Problem from a Deep Learning Perspective.

OpenAI. 2025. OpenAI Alignment Research Blog.

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

Pang, X.; Tang, S.; Ye, R.; Xiong, Y.; Zhang, B.; Wang, Y.; and Chen, S. 2024. Self-Alignment of Large Language Models via Monopolylogue-based Social Scene Simulation.

Parker, C. 2019. Snowball sampling.

Phelps, S.; and Ranson, R. 2023a. Of Models and Tin Men: A Behavioural Economics Study of Principal-Agent Problems in AI Alignment using Large-Language Models.

Phelps, S.; and Ranson, R. 2023b. Of models and tin men: a behavioural economics study of principal-agent problems in AI alignment using large-language models. *arXiv preprint arXiv:2307.11137*.

Poddar, S.; Wan, Y.; Ivison, H.; Gupta, A.; and Jaques, N. 2024. Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning.

Prabhakaran, V.; Mitchell, M.; Gebru, T.; and Gabriel, I. 2022. A Human Rights-Based Approach to Responsible AI. ArXiv:2210.02667 [cs].

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Raji, I. D.; and Dobbe, R. 2023. Concrete problems in AI safety, revisited. *arXiv preprint arXiv:2401.10899*.

Richemond, P. H.; Tang, Y.; Guo, D.; Calandriello, D.; Azar, M. G.; Rafailov, R.; Pires, B. A.; Tarassov, E.; Spangher, L.; Ellsworth, W.; Severyn, A.; Mallinson, J.; Shani, L.; Shamir, G.; Joshi, R.; Liu, T.; Munos, R.; and Piot, B. 2024. Offline Regularised Reinforcement Learning for Large Language Models Alignment.

Russell, S. J. 2019. *Human compatible: artificial intelligence and the problem of control*. New York: Viking. ISBN 978-0-525-55861-3.

Russell, S. J.; and Norvig, P. 1995. *Artificial intelligence: A modern approach;[the intelligent agent book]*. Prentice hall.

Schmer-Galunder, S. 2026. Culture in the code: Anthropological Concepts Decoding AI's Hidden Assumptions. *AI and Ethics*, 6(3): 272.

Schwartz, S. 2012. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, 2(1).

Sharma, A.; Keh, S.; Mitchell, E.; Finn, C.; Arora, K.; and Kollar, T. 2024. A critical evaluation of ai feedback for aligning large language models. *arXiv preprint arXiv:2402.12366*.

Shea, N. 2018. *Representation in cognitive science*. Oxford University Press.

Shen, T.; Jin, R.; Huang, Y.; Liu, C.; Dong, W.; Guo, Z.; Wu, X.; Liu, Y.; and Xiong, D. 2023. Large Language Model Alignment: A Survey.

Shi, Z.; Wang, Z.; Fan, H.; Zhang, Z.; Li, L.; Zhang, Y.; Yin, Z.; Sheng, L.; Qiao, Y.; and Shao, J. 2024. Assessment of Multimodal Large Language Models in Alignment with Human Values.

Shrishak, K. 2025. How not to deploy generative AI: the Story of the European Parliament.

Sierra, C.; Osman, N.; Noriega, P.; Sabater-Mir, J.; and Perelló, A. 2021. Value alignment: a formal approach. ArXiv:2110.09240 [cs] version: 1.

Siththaranjan, A.; Laidlaw, C.; and Hadfield-Menell, D. 2023. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF.

Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; Althoff, T.; and Choi, Y. 2024. A Roadmap to Pluralistic Alignment. ArXiv:2402.05070 [cs].

Stanczak, K.; Meade, N.; Bhatia, M.; Zhou, H.; Böttinger, K.; Barnes, J.; Stanley, J.; Montgomery, J.; Zemel, R.; Papernot, N.; et al. 2025. Societal Alignment Frameworks Can Improve LLM Alignment. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*.

Sun, H.; Shen, Y.; and Ton, J.-F. 2024. Rethinking Bradley-Terry Models in Preference-Based Reward Modeling: Foundations, Theory, and Alternatives. *arXiv preprint arXiv:2411.04991*.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Tian, R.; Xu, C.; Tomizuka, M.; Malik, J.; and Bajcsy, A. V. 2024. What Matters to You? Towards Visual Representation Alignment for Robot Learning.

Wang, B.; Zheng, R.; Chen, L.; Liu, Y.; Dou, S.; Huang, C.; Shen, W.; Jin, S.; Zhou, E.; Shi, C.; Gao, S.; Xu, N.; Zhou, Y.; Fan, Z., Xiaoran Fand Xi; Zhao, J.; Wang, X.; Ji, T.; Yan, H.; Shen, L.; Chen, Z.; Gui, T.; Zhang, Q.; Qiu, X.; Huang, X.; Wu, Z.; and Jiang, Y.-G. 2024a. Secrets of RLHF in Large Language Models Part II: Reward Modeling.

Wang, C.; Zhao, D.; Wang, B.; He, R.; and Hou, Y. 2024b. Do LLMs Have the Generalization Ability in Conducting Causal Inference? *arXiv preprint arXiv:2410.11385*.

Wang, Z.; He, W.; Liang, Z.; Zhang, X.; Bansal, C.; Wei, Y.; Zhang, W.; and Yao, H. 2024c. CREAM: CONSISTENCY REGULARIZED SELF-REWARDING LANGUAGE MODELS.

Xu, C.; Chern, S.; Chern, E.; Zhang, G.; Wang, Z.; Liu, R.; Li, J.; Fu, J.; and Li, P. 2023. Align on the Fly: Adapting Chatbot Behavior to Established Norms.

Yu, T.; Lin, T.-E.; Wu, Y.; Yang, M.; Huang, F.; and Li, Y. 2023. Constructive Large Language Models Alignment with Diverse Feedback.

Zeng, Y.; Liu, G.; Ma, W.; Yang, N.; Zhang, H.; and Wang, J. 2024. Token-level Direct Preference Optimization.

Zhao, S.; Dang, J.; and Grover, A. 2023. Group Preference Optimization: Few-Shot Alignment of Large Language Models.

Zheng, C.; Sun, K.; Wu, H.; Xi, C.; and Zhou, X. 2024. Balancing Enhancement, Harmlessness, and General Capabilities: Enhancing Conversational LLMs with Direct RLHF.

Zhi-Xuan, T.; Carroll, M.; Franklin, M.; and Ashton, H. 2024. Beyond Preferences in AI Alignment. *Philosophical Studies*. ArXiv:2408.16984 [cs].

Zhong, Y.; Ma, C.; Zhang, X.; Yang, Z.; Chen, H.; Zhang, Q.; Qi, S.; and Yang, Y. 2024. Panacea: Pareto Alignment via Preference Adaptation for LLMs.

Zhou, Z.; Liu, Z.; Liu, J.; Dong, Z.; Yang, C.; and Qiao, Y. 2024. Weak-to-Strong Search: Align Large Language Models via Searching over Small Language Models.