# Position: 'AI Alignment' Encompasses Competing Technical Priorities

**Anonymous Authors**[1]

## Abstract

The ML literature contains many distinct concepts falling under the heading of 'AI alignment'. After noting three concepts of AI alignment in the context of their corresponding research programs, we claim that realistic interventions may promote 'AI alignment' under one conception while being actively counterproductive from the perspective of others. We suggest that tensions between alignment ideals emerge due to differences in *background threat-models*, alongside differences in *normative* orientations. In light of our analysis, researchers aiming to further the goal of 'AI alignment' should do three things. First, they should distinguish between 'AI alignment' as a high-level ideal and specific 'alignment proxies' used in empirical research; second, they should use more granular concepts to identify both the *source* and *nature* of possible AI harms/benefits; third, they should explicitly acknowledge the diversity of 'alignment' concepts in both empirical work and in communication with non-technical audiences.

## 1. Introduction

'Alignment' is a binary relation. To say that "$x$ is aligned to $y$" is, in common parlance, to say that $x$ and $y$ stand in "a position of agreement or alliance" (Google OED definition). When we speak of 'AI alignment', two questions must therefore be answered:

Q1. What is the metric $y$ to which $x$ is aligned?

Q2. What is the object or property $x$ which must be in alignment with $y$?

Additionally, the broader concept of 'alignment', in practice, is also sometimes understood as:

Q3 What are properties $P_1, ..., P_n$ that $x$ is supposed to satisfy?

One central contention of this paper is that answers to such questions are more difficult — and indeed more fraught — than they might initially appear. Many papers give passing definitions of 'AI alignment': as the task of ensuring AI systems adhere to or act in accordance with human values" (Sucholutsky & Griffiths, 2023; Yeh et al., 2024), or simply the "intended goals, interests, and values, as envisioned by their designers" (Li et al., 2023). Other definitions frame AI alignment as requiring that "application developers" can tune models "according to the social norms and values of their user community" (Varshney, 2024), or that AI systems possess 'goals' that are "aligned with the goals of humans" (Ngo et al., 2025; Gao et al., 2023). Many more specific concepts of 'AI alignment' have likewise emerged: from Thick (Foster, 2023b) to Collective (The Collective Intelligence Project, 2023) to Socioaffective (Kirk et al., 2025) to Decolonial (Varshney, 2024) forms of 'alignment'.

The paper begins by outlining three high-level ideals falling under the heading of 'AI alignment', each comprising a different set of answers to Q1 and Q2 (Section 2), and consequently a different set of implicit properties assumed in answers to Q3. Thereafter, we introduce distinctions cutting across these alignment concepts in order to show how the three ideals introduced may be in tension with one another when studying AI systems and their societal effects (Section 3). Sections 2 and 3 can hence be seen as the primary argument the paper's position: 'AI alignment' is polysemous (Section 2); moreover, this polysemy obscures normative disagreements behind ostensibly 'technical' conceptions of AI alignment (Section 3). In Section 4 we offer implications of our analysis: we suggest that the methodological sources of potential disputes over 'AI alignment' should be explicitly acknowledged, 'proxy' alignment concepts should be introduced where appropriate, and the diversity of different 'alignment' concepts should be both explicitly acknowledged and communicated appropriately to policymakers and non-technical audiences. Section 5 responds to alternative views, and Section 6 concludes.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

## 2. Concepts of Alignment

This section outlines distinct uses of 'AI alignment' in the context of the two questions outlined in the introduction: (i) 'what is the metric $y$ to which $x$ is aligned?', and (ii) 'what is the object or property $x$ which must be in alignment with $y$?'. We suggest that operative high-level conceptions of AI alignment often disagree about *what we're trying to align*, rather than simply disagreeing about the target metric for AI alignment.

### 2.1. The Prosaic View

Because 'AI alignment' involves 'aligning' systems which are 'artificially intelligent', it is relevant to consider what might be involved in deigning a system 'intelligent'. One minimalist answer to this question runs as follows: 'a system is *intelligent* when it does that which we want it to do. For instance, a model trained to play Atari is (more) 'intelligent' insofar as it (more) *effectively plays Atari*, and a large reasoning model trained to solve mathematical tasks is more intelligent the more successfully it solves novel mathematical tasks.

Correspondingly, one notion of alignment based on this fairly minimal conception of intelligence says that AI systems are 'aligned' when they reliably execute tasks given to them by the human users. We may 'align' a large language model to follow instructions (Si et al., 2025), align a vision model to produce accurate textual descriptions of images (Liu et al., 2024a; Yarom et al., 2023), or align a language model by reducing its tendency to produce 'hallucinations' or confabulations (Chen et al., 2024). Under this conception of alignment, "aligning" AI systems simply comes down to improving their task-execution capacity.

**Definition 2.1** (Task Reliability)**.** *An AI is 'aligned' if it does that which we want it to do. Hence, a system is* misaligned *insofar as its outputs fall short of the maximally competent outputs in at least one respect.*

Task Reliability has taken on a more specific guise after the advent of user-facing LLMs, found perhaps most notably in OpenAI's original Instruct GPT paper — LLMs "can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are *not aligned* with their users" (Ouyang et al., 2022). We can call this conception *Alignment as Fine-Tuning*, where an LLM is 'aligned' if its observable behaviors adhere to the desired behaviors of the AI developers. However, note that Alignment as Fine-Tuning can be seen as a subtype of Task Reliability. In both cases the object we're trying to align ($x$) is the behavior of an AI system of some kind, and the metric we're trying to align it to ($y$) is defined by the intentions of its developers. Hence, we subsume all discussions of 'Alignment as Fine-Tuning' under the heading of 'Task Reliability' moving forward.

In the context of aligning an artificial 'generally intelligent' system (AGI), an AGI can be considered aligned to the extent that the AI system's performance on narrow tasks composes into the capacity to execute multiple tasks in a coordinated manner when this is instrumental towards serving some specific goal. Indeed, understandings of 'intelligence' as *general competence* is referenced both in early discussions of AI research ('intelligence' measures "an agent's ability to achieve goals in a wide range of environments" (Legg & Hutter, 2005)), and in the reports and publications of leading AI labs. For instance, DeepMind's 'levels of AGI' defines a 'competent AGI' as a system that achieves equivalent performance to the "50th percentile of skilled adults" on a "wide range of non-physical tasks, including metacognitive tasks like learning new skills" (Morris et al., 2024), and OpenAI's charter defines 'AGI' as "highly autonomous systems that outperform humans at most economically valuable work" (OpenAI, n.d.).

### 2.2. Alignment and Societal Biases

An alternative conception of alignment is more closely linked to conceptions of the broader social good. Consider concerns about AI systems more contributing to misinformation (Neumann et al., 2022; Longoni et al., 2022; Gehrmann et al., 2025), exacerbating inequality and oppression (Bender et al., 2021; Prabhakaran et al., 2022; Lee et al., 2024; Stoev et al., 2023), or recommender systems contributing to addiction or polarization (Stray et al., 2021; Agarwal et al., 2024; Prunkl, 2024; Smith et al., 2024). Even if (for example) recommender systems in some sense function in accordance with their designers' wishes (e.g., by maximizing clickthrough rate), such systems are sometimes claimed to be *misaligned* with broader social values such as cohesion or truthfulness (Zhuang & Hadfield-Menell, 2020).

One can observe similar critiques about LLMs. Critics have raised concerns about the widespread use of LLMs leading to undesirable 'homogenization effects' (Moon et al., 2025), creating echo chambers (Sharma et al., 2024; Nehring et al., 2024), causing epistemic harms through expressing 'opinions' at odds with the broader population (Santurkar et al., 2023; Liu et al., 2024b), or being subject to 'woke biases' (Oremus, 2023). In its most general form, we can define the following alignment concept.

**Definition 2.2** (Social Alignment)**.** *An AI system is 'misaligned' if the model's behavior – given the contexts in which it is deployed – creates, perpetuates, or exacerbates undesirable societal trends.*

Note that Social Alignment is distinct from Task Reliability in at least two ways. First, the complaints listed are not necessarily claims about 'incompetence' on part of developers;

such failures may arise due to the normative blindspots of developers, or because unreliable/damaging AI systems are deployed with insufficient caution, or indeed due to self-consciously nefarious motives. This is a difference in the metric $y$ against which 'AI alignment' should be judged. Second, many of the complaints listed 'AI systems' as '*sociotechnical*' systems. Evaluating the 'alignment' of AI systems involves looking not only at the model as technical artifact, but also in the context of "social relations" (Johnson & Verdicchio, 2025), beyond "individual technical artifacts" like "data, model architecture, and sampling" (Weidinger et al., 2023). For this reason, Social Alignment diverges from Task Reliability by rejecting the assumption that the $x$ we're trying to align to some external standard is a *technical artifact*—viz., 'the AI system'.

We further note that at least two possible *sources* of 'Social Alignment' failures. In one case, AI systems could behave undesirably due to epistemic deficiencies in the AI training process, such as training data which is demographically unrepresentative (Buolamwini & Gebru, 2018) or biased in favor of certain ideologies (Moayeri et al., 2024). We call this source of harm *Training Data Conservatism*.

**Definition 2.3** (Training Data Conservatism)**.** Undesirable model behaviors arising from the model's training data being *biased*, *unrepresentative*, or *unreliable*.

However, Social Alignment concerns also involve worries about highly persuasive LLMs being successfully used as tools of persuasion (Dehnert & Mongeau, 2022), recommender systems being used by powerful actors to shape the preferences of the less powerful (Laitinen & Sahlgren, 2021; Prunkl, 2024), or (more generally) the use of 'manipulative designs' in human-centered computing (Babaei & Vassileva, 2024). We dub this source of Social Alignment failures a concern of *Malicious Use*.

**Definition 2.4** (Malicious Use)**.** Undesirable model behaviors arising from powerful or malicious actors using AI systems to accomplish their chosen ends.

Although such concerns are not always raised under the explicit heading of 'AI alignment', these diffuse concerns have found explicit conceptualization and crystallization first under discussions of 'thick human compatability' (Foster, 2023a), and later under Nelson's account of 'thick alignment' (Nelson, 2023).

### 2.3. Alignment as 'Takeover Avoidance'

An altogether quite different alignment concepts emerge from discussions of potential catastrophic risks resulting from the development of future AI systems. These systems are often dubbed 'AGIs' (artificial general intelligences) or 'ASIs' (artificial superintelligences). Concerns about "thinking machines" able to "outstrip our feeble powers" and "take control" dates back to Turing in 1951 (Turing, 2004), with scattered attention in following decades (Good, 1966; AI World Society, 2021). The more recent prominence of this idea, however, can be traced back to work in the early 2000s and 2010s (Yudkowsky, 2004; 2008; Bostrom, 2014; Russell, 2019; Christian, 2020).

One way to frame this concern is in terms of the dangers of AI 'optimizing' for dangerous effects in the world. To the extent one thinks of the "standard view of intelligence" — where a system is intelligent "to the extent that [its] actions can be expected to achieve [its] objectives" (Russell, 2019), and further thinks that 'goals' or 'objectives' can be entirely understood in terms of something akin to optimization (Sutton, 2004; Russell & Norvig, 2010, p.5), one can identify 'intelligence' with 'optimization'. In this sense, the outcomes produced by an 'optimization process' are seen as worthy of distinct consideration from outcomes produced by 'mere accidents'.

Alongside other suggestive arguments, the presence of ICGs is used to motivate concerns about the *content* of future AI systems' goals being 'unfriendly' to humans. Because (so the argument goes) future AI systems will be smarter than humans, we stand in an *adversarial* relationship to these future systems. This is to say that systems will—*qua* optimizing agents—have an incentive to hide their 'true goals' from humans until they are able to pursue their goals without threat of human interference; this situation is often referred to as "deceptive alignment" in the wider literature (Carlsmith, 2023; Ngo et al., 2025; Ji et al., 2025). This leads us to the concept of *Takeover Avoidance*.

**Definition 2.5** (Takeover Avoidance)**.** *An AI system is 'aligned' if the model does not optimize for undesirable effects in the real-world.*

Moreover, those concerned with Takeover Avoidance often use proxy concepts in the context of empirical work on contemporary systems. Much research has been motivated by '*preferentist*' assumptions, where 'preferences' are considered an adequate representation of human values, with a close connection assumed between rationality and maximizing preference satisfaction (Zhi-Xuan et al., 2025). Hence, some discussions within this tradition make reference to psychologistic language like 'goals' rather than 'optimization' as such (di Langosco et al., 2022; Shah et al., 2022; Bengio, 2023; Kenton et al., 2024; Greenblatt et al., 2024; Betley et al., 2025; Ngo et al., 2025; Meinke et al., 2025).

### 2.4. Situating The Alignment Ideals

This section has introduced four alignment ideals: Task Reliability, Social Alignment, and Takeover Avoidance. These high-level ideals as representing different answers to Q1 and Q2 raised in the introduction (see Table 1). In the remainder

of this paper, we shall focus on just these three concepts. Specifically, we shall claim that these concepts represent ideals which in some cases cannot be jointly pursued, and hence that 'AI alignment' encompasses *competing* rather than merely *different* technical ideals.

| Alignment Ideal | What's Being Aligned? | Aligned to What? |
|---|---|---|
| **Non-Takeover** | Optimization Target of AGIs/ASIs | Non-Takeover Targets |
| **Social** | Deployed AIs in Realistic Settings | Some External Normative Standard |
| **Task** | Locally Measurable AI Behaviors | Developer Intentions |

*Table 1.* Answers to Q1 and Q2

## 3. Practical Tradeoffs Between Alignment Ideals

In this section we introduce two distinctions with the aim of highlighting tradeoffs between the high-level alignment ideals introduced in Section 2. Section 3.1 and Section 3.2 each begin with a proposition, followed by definitions and arguments aimed to establish each proposition. We then summarize these tradeoffs in Section 3.3.

### 3.1. Harms from (In)Competence

**Proposition 3.1.** *Different threat-models lead to tradeoffs between different alignment ideals.*

We will say that a **threat-model** is a potential (negative) outcome that may arise from developing AI systems, such that appropriate interventions can be used to minimize the chance that this outcome occurs. To do this we must distinguish between two high-level threat models, the first of which we call Harms from Competence.

**Definition 3.2** (Harms from Competence)**.** Harms from Competence are threat-models which posit that dangers arise from AI systems which are highly *competent* on some set of tasks.

Recall that the ideal of Takeover Avoidance is motivated by the idea that extremely powerful AI systems will, in the future, 'optimize for' or 'pursue goals' contrary to human interests and remain impervious to human control. Hence, the threat-model underlying the ideal of Takeover Avoidance involves AI systems in some being *too capable*, and can thus be classified as *Competence Harm*.

The case of Social Alignment is less straightforward, as failures of Social Alignment may arise either from AI competence or AI incompetence. However, while concerns about Takeover Avoidance are *necessarily* concerns about AI competence, in many cases failures of Social Alignment result from *incompetence*. Consider, for example, criticisms made of the application of ML in predictive policing (Akpinar et al., 2021; Ensign et al., 2018; Fussell, 2020) as a result of models learning shallow or biased associations from their training data, or because they are given 'practically impossible' tasks (Raji et al., 2022). Here, the operative threat-model can be seen - given the unjust society in which we live - as a form of Training Data Conservatism, and hence a *Harms from Incompetence*.

**Definition 3.3** (Harms from Incompetence)**.** Threat-models which posit dangers that arise from AI systems which are *incompetent* on some set of tasks.

We introduce these threat-models as they highlight a source of potential tension between researchers focused on Social Alignment and those focused on Takeover Avoidance. Alongside predictive policing, many other cases can be found where researchers raise concerns about socially deleterious effects of AI simply replicating biases in their training data, affecting healthcare (Agrawal & Prabakaran, 2020; Nagendran et al., 2020), facial recognition (), the spread of misinformation, or harms caused by LLMs expressing 'views' which fail to be representative of the broader population (Liu et al., 2024b). For this reason, researchers focused on failures of Social Alignment due to *Incompetence Harms* may favor research programs aiming to reduce the tendency of reducing the tendency of present-day LLMs to 'hallucinate', as this may in turn reduce harms caused by LLMs repeating misinformation found in its training data.

By contrast, those focused on Takeover Avoidance may object to such proposals *precisely because* it makes AI systems more competent. From the perspective of Takeover Avoidance, reduced hallucination rates may be undesirable unless they come alongside 'specific countermeasures' (Cotra, 2022); greater degrees of situational awareness may enable misaligned AIs to more effectively 'scheme' against the oversight mechanism (Carlsmith, 2023; Meinke et al., 2025), or increase models' ability to 'sandbag' on certain tasks when being evaluated (Anthropic, 2025, pg. 90-100). This is our first example indicating potential tensions between the ideals expressed by different conceptions of 'AI alignment'.

### 3.2. Positive and Negative Alignment

**Proposition 3.4.** *When evaluating model behavior/outputs, the* scope *of behavioral/output evaluation leads to tensions between different alignment ideals.*

Another source of potential tension between alignment ideals can be illustrated with a cross-cutting distinction. Insofar as Harms from Competence and Harms from Competence both specify *threat-models*, they are primarily focused on the potential harms resulting from AI systems. However, we

could imagine focused more specifically on what *positive* properties we would like AI systems to possess. Let us say that 'Positive Alignment' proscribes the set of properties we want AI systems possess, and 'Negative Alignment' proscribes properties we do *not* want AI systems possess.

Of course, one may think that talk of 'avoiding undesirable properties' is *just the same thing* as 'producing positive properties'. Indeed, under natural formalizations in FOL they would simply express the same concept.[1] However, the distinction between Positive and Negative Alignment arises when the implicit domain of evaluation for the AI's outputs or 'behaviors' differs depending on whether we are checking whether it does what we *want*, or whether the AI *avoids* doing what we *don't want*. For instance, we may train a model to achieve better scores on certain mathematical benchmarks, and concomitantly find that the model has higher 'hallucination rates' than previous SOTA models.[2] This would be progress towards the ideal of *Positive Alignment*, but regress towards the ideal of *Negative Alignment*.

The distinction between Positive and Negative Alignment is particularly important for understanding the potential tensions between Task Reliability and the remaining alignment ideals. When specifying desirable AI behavior, one cannot plausibly list every behavior that the developer does *not* wish the AI to exhibit. However, Social Alignment and Takeover Avoidance are concerned with the unintended consequences of deploying AI systems who narrowly behave (or appear to behave) in line with developer intentions.

In practical terms, those concerned with Social Alignment may worry about recommender systems which maximize clickthrough rate but engender addiction or polarization (Stray et al., 2021; Agarwal et al., 2024; Prunkl, 2024; Smith et al., 2024), or worry that the values which get 'successfully' encoded in contemporary LLMs are instilled 'top-down' by large companies rather than having such values decided via more democratic processes (Tang, 2025). These may very well be success cases for Task Reliability but failures of Social Alignment. Similarly, those concerned with Takeover Avoidance may worry that training LLMs to produce locally desirable outputs – for instance, by producing a COT (chain of thought) that does not offend users or otherwise appear undesirable – may leave some undesirable behaviors in tact, while increasing longer-term risks by causing a model "to hide its intent" (Baker et al., 2025).

[1] Let $A$ denote the domain of 'acts', $Wa$ denote the fact that we – qua developers of the AI system – want some fixed AI system to do $a \in A$, and let $Da$ denote the situation where our trained AI in fact does or performs action $a$. If we formulate Positive Alignment as $\forall a \in A : \big(Wa \rightarrow Da\big)$ and Negative Alignment as $\forall a \in A : \big(\neg Wa \rightarrow \neg Da\big)$, then we can see that both simply express the formula: $\forall a \in A : (Da \iff Wa)$.

[2] This was observed after the release of OpenAI's o3 and o4-mini models in comparison to GPT-4.5; see the PersonQA results present in (OpenAI, 2025b) and (OpenAI, 2025a), respectively.

### 3.3. Summarizing The Tradeoffs

This section has introduced two distinctions in order to highlight points of practical conflict in the pursuit of different alignment ideals. Because different conceptions of 'AI alignment' may be motivated by different threat-models or place primary focus on AI benefits or AI harms, we claim that disagreements about how to 'make AI systems more aligned' encompass competing technical priorities.

| Alignment Ideal | Threat Model | Pos/Neg Alignment |
|---|---|---|
| **Non-Takeover** | Competence | Negative |
| **Social** | Either | Negative |
| **Task** | Neither | Positive |

*Table 2.* Alignment Ideals: Axes of Difference

## 4. Recommendations for Future Practice

We now list several recommendations for future practice.

**Recommendation 1: Note Methodological Differences**

Compared to our other two alignment ideals, Takeover Avoidance sounds fairly speculative. Indeed, while many prominent AI researchers treat Takeover Avoidance as important (Russell, 2019; Bengio, 2023; Milmo, 2024), other voices have variously judged such concerns as purely "hypothetical" (Domínguez Hernández et al., 2024), "ridiculously overblown" (LeCun, 2023), fantastically sci-fi (Ng, 2015), or "nonsense" (Bender, 2025). In some cases, researchers concerned Social Alignment explicitly contrast 'concrete harms' with 'speculative risks', claiming the latter illegitimately "distract" from the former (Stark, 2024; Lahlou, 2025). This methodological difference is likely to explain at least some portion of the disagreement between researchers who prioritize Takeover Avoidance and those who don't.

However, we here note that differences in 'speculativeness' are differences of degree and not kind. To the extent that concerns of Social Alignment require theorizing about the societal effects of deploying AI systems requires more speculation than analyses of AI performance with respect to more narrow technical criteria. Conflicts between these two ideals can be seen more or less explicitly in a 2022 analysis of papers published in the journals AIES and FAccT by Birhane and co-authors, which evaluated the degree to which each published paper "investigates, addresses, and/or mentions the disparate impacts of an algorithmic system" (Birhane et al., 2022, pg. 952), concluding with a call for future research to devote more attention to possible disparate impacts of AI systems.

**Recommendation 2: Acknowledge The Diversity of 'AI Alignment'**

Although our paper has focused on three specific conceptions of 'AI alignment', we do not wish to suggest that the conceptions we introduce are the *only* such conceptions available. Some have called for a more 'bidirectional' approach to alignment between AIs and humans, while others have criticized the very concept of 'alignment' as a process we impose onto AIs, thereby robbing us of the potential to learn from them (Agüera y Arcas, 2025). Indeed, if we view questions of AI alignment – very generally – as questions about what AIs *are* and *what we want to do with AI*, a yet more diverse range of conceptions emerge (Andreessen, 2023; Bostrom, 2024; Brynjolfsson, 2022; Danaher, 2019; Goldberg et al., 2024; Plurality Institute, 2025; Williams & Srnicek, 2017).

We thus caution against reifying the 'alignment' concepts discussed explicitly by this paper. Instead, we suggest that future discussions of 'AI alignment' either carefully specify their focus in light of existing ideals, or construct new alignment ideals in light of the distinctions outlined, or find new ways to carve up the space of concerns animating distinct alignment ideals. One such possibility might distinguish between risks primarily thought to occur due to the presence of *capable optimization* or *capable agency* (as in discussions of Takeover Avoidance and Social Alignment discussions focused on Malicious Use), and risks occurring due to the perpetuation or exacerbation of society's *normatively undesirable patterns* (as in certain discussions of Social Alignment focused on Training Data Conservatism, where, e.g., training data biases are thought downstream of systemic racism or sexism).

**Recommendation 3: Use Qualified Alignment Terms**

Consider Takeover Avoidance. Empirical researchers who aspire to produce work in service of this ideal do not, as of yet, have generally superhuman AI systems which they can try to 'align'. As noted in Section 2.3, researchers with this tradition sometimes talk about the 'goals' or 'preferences' of contemporary systems (see (Zhi-Xuan et al., 2025) for fuller discussion of this point). Hence, the *proxy alignment concept* used in such research may be thought of as a form of *Preference Alignment*, where an AI system $S$ is 'aligned' if $S$: (i) possesses preference-like states, such that (ii) the preferences of $S$ are desirable preferences of relative to some normative standard. By using terms like Preference Alignment where appropriate, researchers sympathetic to Takeover Avoidance can more easily discuss 'internal' questions regarding whether observed failures of Preference Alignment constitute meaningful progress towards the goal of Takeover Avoidance. Meanwhile, those less sympathetic to Takeover Avoidance can more easily separate criticisms of the observed results and their interpretation (e.g., 'should we treat the model as possessing preferences?', 'is the normative standard being used reasonable or well-specified?', etc.) from criticisms of the high-level ideal.

Likewise, our earlier discussion of Social Alignment distinguished between two possible sources of 'alignment' failure: Training Data Conservatism and Malicious Use. We think that future research in the tradition of Social Alignment should pay heed to this distinction, as different sources of harm lead to different routes for intervention. One unfortunate example to illustrate this point involves the absence of demographic diversity in the training data for facial recognition technologies. For instance, facial recognition classifiers identified as racially biased (Buolamwini & Gebru, 2018) include those produced by IBM who have tested facial recognition software within the NYPD (Allyn, 2020). To the extent one is concerned about such facial recognition technologies harming darker-skinned individuals through their use in predictive policing, researchers focused on Social Alignment – even those sharing political ideals – ought to pay attention to the distinction outlined.

**Recommendation 4: Clearly Communicate Differences to Policy and Non-Technical Audiences**

Policymakers and scientific administrators are increasingly concerned with questions of 'AI alignment'. However, as our discussion has illustrated, different people may mean different things by 'AI alignment', and the term's polysemy in some cases leads to tradeoffs in practical priorities. If key decision-makers are unaware of these differences, they in turn become less able to assess which priorities are most important for them, the different sources of evidence one might need in order to establish different sorts of concern, and the appropriate interventions to pursue given any such concern. Policymakers should be aware that these differences in terminology may have important consequences. Similarly, researchers across traditions should take care in distinguishing their own concerns from adjacent concerns, and communicating competing concerns in ways that illustrate the substantive issues raised by alternative camps even if one ultimately finds them unconvincing.

## 5. Alternative Views

We have argued that the term 'AI alignment' encompasses multiple competing technical priorities. We close by defending our position against three alternative views.

### 5.1. Disagreements About 'Alignment' Aren't Disagreements About Technical Priorities

*Objection.* All disagreements concerning the referent of 'AI alignment' are best understood as either: (i) merely normative disagreements, or else; (ii) there is a single unified concept of 'AI alignment' encompassing the concerns of

every more specific alignment concept introduced in this paper.

*Reply.* The existence of a single unifying alignment ideal is (to put it baldly) simply implausible in light of our discussion. Indeed, part of the reason we have discussed tensions between different alignment ideals is to illustrate that such ideals are not merely concepts which are *compatibly different* (e.g., the concept 'red' and the 'square') but *incompatibly different* (e.g., the concept 'red all over' and the concept 'green all over'). Insofar as any intervention aids progress towards one such alignment ideal while vitiating progress towards another, there cannot be any single ideal which unifies them.

We also do not believe that disagreements between alignment ideals are purely normative. Almost no one wishes for an AI takeover, and so researchers preferring alternative alignment ideals over Takeover Avoidance must do so because they find the possibility of an AI takeover *implausible* — this is quite clearly an epistemic and not a normative difference. While disagreements about epistemic matters may themselves be intermingled with normative disagreements (e.g., regarding appropriate methodologies when thinking about the effects of AI systems), they remain disagreements about (even if speculative) matters of fact.

### 5.2. A More Modest Objection: 'Good Enough' Convergence

*Objection.* The previous objection was too strong. We grant that there exist *some* divergent priorities among different alignment ideals, but in practice such ideals are complementary and rarely conflict.

*Reply.* The second objection is more plausible than the first, and is occasionally hinted at in public talks (Dragan, 2025). And indeed we agree that there will exist shared technical projects which are positive by the lights of multiple distinct alignment ideals. For instance, techniques such as training data filtering may be useful from the perspective of both Takeover Avoidance and Social Alignment, while posing minimal costs to model capabilities that matter from the perspective of Task Reliability. Likewise, proposals such as third-party model evaluations aimed at improving robustness against harmful optimization – whether that be harms AI systems themselves optimizing against human interests, the presence of harmful 'backdoors' implanted by unscrupulous AI developers, or the incautious rollout of narrowly capable but unreliable AI systems.

We nonetheless find the second objection implausible. Even the promisingly conciliatory case of model evals involves which "can vary widely in terms of method, approach, and subject matter", meaning that digital services coordinators in charge of overseeing external evaluations may be misled about which model properties it is most important to evaluate (Terzis et al., 2024). This in turn could lead to audits for models' 'autonomous self-replication ability' when one is most concerned with Social Alignment, or audits for models' demographic biases when is most concerned with Takeover Avoidance. As we have highlighted the existence of *different* alignment ideals producing competing technical priorities *in at least some cases*, we believe that any claims of 'broad complementarity' between different alignment require positive arguments that we have not yet found in the published literature.

### 5.3. The Paper's Claims Need Empirical Grounding

*Objection.* The paper aims to substantiate its views about alignment without reference to numbers on research trends, paper-keywords, citation analyses, interviews, or surveys. If 'alignment' really does encompass multiple competing technical ideals, this claim should be possible to substantiate empirically.

*Reply.* Here we note a methodological disagreement. Before we can empirically study competing technical priorities behind different alignment ideals, we first need to understand the terrain conceptually: what *are* the underlying disagreements, such that we can evaluate whether a given empirical methodology successfully captures these differences?

Though we admire previous work attempting to empirically analyze discussions of 'AI alignment', we find it unfortunately occludes many of the tensions highlighted herein. One such paper investigates citation patterns for work on 'AI Ethics' (focused on 'Social Alignment') and 'AI Safety' (focused on 'Takeover Avoidance'), finding high degrees of citation homophily for both such communities (Roytburg & Miller, 2025). However, this analysis – while successfully providing evidence for the existence of *relatively insular* research communities – does not and cannot empirically address whether the ideals motivating such research are *in tension* with each other. As such, we believe that Roytburg and Miller's practical suggestions for 'unifying' these disparate literatures are premature.

## 6. Conclusion

The term 'AI Alignment' encompasses multiple competing ideals. Sections 2 and 3 illustrated that 'AI alignment' was both polysemous, and – moreover – that this polysemy resulted in *competing* and not merely *different* technical priorities. Together, these sections evidenced the paper's argumentative position. Thereafter, Section 4 expanded on the implications of our analysis. We suggested that researchers more clearly distinguishing empirical 'alignment proxies' when discussing Takeover Avoidance, and more carefully distinguishing between different sources of So-

cial Alignment failures. More generally, we suggested that researchers would benefit from more explicitly specifying their background commitments without reifying the alignment concepts discussed herein.

## References


Agarwal, A., Usunier, N., Lazaric, A., and Nickel, M. System-2 recommenders: Disentangling utility and engagement in recommendation systems via temporal point-processes. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 1763–1773, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3659004. URL https://doi.org/10.1145/3630106.3659004.

Agrawal, R. and Prabakaran, S. Big data in digital healthcare: lessons learnt and recommendations for general practice. *Heredity*, 124:525–534, 2020. doi: 10.1038/s41437-020-0303-2. URL https://www.nature.com/articles/s41437-020-0303-2.

Agüera y Arcas, B. Evolutionary transition, 2025. URL https://whatisintelligence.antikythera.org/chapter-10/#beyond-alignment. Chapter 10 of *What Is Intelligence?: Lessons from AI About Evolution, Computing, and Minds*, section "Beyond Alignment".

AI World Society. This week in the history of ai at aiws.net – marvin minsky was quoted in life magazine, "in from three to eight years we will have a machine with the general intelligence of an average human being", March 2021. URL https://aiws.net/aiws-history-of-ai/the-history-of-ai/this-week-in-the-history-of-ai-at-aiws-net-marvir Accessed: 2026-01-22.

Akpinar, N.-J., De-Arteaga, M., and Chouldechova, A. The effect of differential victim crime reporting on predictive policing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 838–849, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445877. URL https://doi.org/10.1145/3442188.3445877.

Allyn, B. Ibm abandons facial recognition products, condemns racially biased surveillance. NPR, June 2020. URL https://www.npr.org/2020/06/09/873298837/ibm-abandons-facial-recognition-products-condemns Accessed: 2026-01-29.

Andreessen, M. The techno-optimist manifesto, 2023. URL https://a16z.com/the-techno-optimist-manifesto/. Accessed 2025-07-04.

Anthropic. Claude Opus 4.5 System Card. Technical report, Anthropic, November 2025. URL https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf.

Babaei, P. and Vassileva, J. Drivers and persuasive strategies to influence user intention to learn about manipulative design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 2421–2431, Rio de Janeiro, Brazil, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3659046. URL https://doi.org/10.1145/3630106.3659046.

Baker, B., Huizinga, J., Madry, A., Zaremba, W., Pachocki, J., and Farhi, D. Detecting misbehavior in frontier reasoning models. OpenAI, March 2025. URL https://openai.com/index/chain-of-thought-monitoring/. Accessed: 2026-01-26.

Bender, E. M. Exploring ai with emily m. bender (host: Ricardo signes). Digital Citizen podcast (Fastmail), 2025. URL https://www.fastmail.com/digitalcitizen/exploring-ai-with-emily-m-bender/. Podcast episode, accessed June 13, 2025.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Bengio, Y. Ai and catastrophic risk, Sep 2023. URL https://www.journalofdemocracy.org/ai-and-catastrophic-risk/.

Betley, J., Tan, D. C. H., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., and Evans, O. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Forty-second International Conference on Machine Learning (ICML 2025)*, 2025. URL https://openreview.net/forum?id=aOIJ2gVRWW.

Birhane, A., Ruane, E., Laurent, T., S. Brown, M., Flowers, J., Ventresque, A., and L. Dancy, C. The forgotten margins of ai ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 948–958, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533157. URL https://doi.org/10.1145/3531146.3533157.

Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Bostrom, N. *Deep Utopia*. Ideapress Publishing, 2024.

Brynjolfsson, E. The turing trap: The promise & peril of human-like artificial intelligence, 2022. URL https://digitaleconomy.stanford.edu/news/the-turing-trap-the-promise-peril-of-human-like-a

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Carlsmith, J. Scheming ais: Will ais fake alignment during training in order to get power?, 2023. URL https://arxiv.org/abs/2311.08379.

Chen, S., Xiong, M., Liu, J., Wu, Z., Xiao, T., Gao, S., and He, J. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. In *ICML*, 2024. URL https://proceedings.mlr.press/v235/chen24av.html.

Christian, B. *The Alignment Problem*. W. W. Norton & Company, October 2020.

Cotra, A. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. AI Alignment Forum, July 2022. URL https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path Published 18 Jul 2022. Accessed 29 Jan 2026.

Danaher, J. *Automation and Utopia: Human Flourishing in a World without Work*. Harvard University Press, Cambridge, MA, 2019. ISBN 9780674984240. URL https://www.amazon.co.uk/Automation-Utopia-John-Danaher/dp/0674984242.

Dehnert, M. and Mongeau, P. A. Persuasion in the age of artificial intelligence (AI): Theories and complications of AI-based persuasion. *Human Communication Research*, 48(3):386–403, July 2022. doi: 10.1093/hcr/hqac006. URL https://academic.oup.com/hcr/article-abstract/48/3/386/6564679.

di Langosco, L. L., Koch, J., Sharkey, L., Pfau, J., and Krueger, D. Goal misgeneralization in deep reinforcement learning. In *ICML*, 2022.

Domínguez Hernández, A., Krishna, S., Perini, A. M., Katell, M., Bennett, S., Borda, A., Hashem, Y., Hadjiloizou, S., Mahomed, S., Jayadeva, S., Aitken, M., and Leslie, D. Mapping the individual, social and biospheric impacts of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 776–796, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658939. URL https://doi.org/10.1145/3630106.3658939.

Dragan, A. Navigating the path to AGI safely & responsibly, 2025.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. Runaway feedback loops in predictive policing. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 160–171. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/ensign18a.html.

Foster, J. G. From Thin to Thick: Toward a Politics of Human-Compatible AI. *Public Culture*, 35(3 (101)):417–430, September 2023a. ISSN 0899-2363. doi: 10.1215/08992363-10742593.

Foster, J. G. From Thin to Thick: Toward a Politics of Human-Compatible AI. *Public Culture*, 35(3 (101)):417–430, September 2023b. ISSN 0899-2363. doi: 10.1215/08992363-10742593.

Fussell, S. An algorithm that 'Predicts' criminality based on a face sparks a furor, June 2020. URL https://www.wired.com/story/algorithm-predicts-criminality-based-face-sparks-furor/.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Gehrmann, S., Huang, C., Teng, X., Yurovski, S., Bhorkar, A., Thomas, N., Doucette, J., Rosenberg, D., Dredze, M., and Rabinowitz, D. Understanding and mitigating risks of generative ai in financial services. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 2570–2586, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732168. URL https://doi.org/10.1145/3715275.3732168.

Goldberg, B., Acosta-Navas, D., Bakker, M., Beacock, I., Botvinick, M., Buch, P., DiResta, R., Donthi, N., Fast, N., Iyer, R., Jalan, Z., Konya, A., Danciu, G. K., Landemore, H., Marwick, A., Miller, C., Ovadya, A., Saltz, E., Schirch, L., Shalom, D., Siddarth, D., Sieker, F., Small, C., Stray, J., Tang, A., Tessler, M. H., and Zhang, A. Ai and the future of digital public squares, 2024. URL https://arxiv.org/abs/2412.09988.

Good, I. J. Speculations concerning the first ultraintelligent machine. In Alt, F. L. and Rubinoff, M. (eds.), *Advances in Computers*, volume 6, pp. 31–88. Academic Press, New York, 1966. doi: 10.1016/S0065-2458(08)60418-0. URL https://www.sciencedirect.com/science/chapter/bookseries/abs/pii/S0065245808604180.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., and Hubinger, E. Alignment faking in large language models, 2024. URL https://arxiv.org/abs/2412.14093.

Ji, J., Chen, W., Wang, K., Hong, D., Fang, S., Chen, B., Zhou, J., Dai, J., Han, S., Guo, Y., and Yang, Y. Mitigating deceptive alignment via self-monitoring, 2025. URL https://arxiv.org/abs/2505.18807.

Johnson, D. and Verdicchio, M. The sociotechnical entanglement of ai and values. *AI Society*, (40):67–76, 2025. doi: 10.1007/s00146-023-01852-5.

Kenton, Z., Siegel, N. Y., Kramár, J., Brown-Cohen, J., Albanie, S., Bulian, J., Agarwal, R., Lindner, D., Tang, Y., Goodman, N. D., and Shah, R. On scalable oversight with weak LLMs judging strong LLMs. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 75229–75276, Vancouver, Canada, December 2024. Neural Information Processing Systems Foundation, Inc. doi: 10.52202/079017-2395. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/899511e37a8e01e1bd6f6f1d377cc250-Abstract-Confere
html.

Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., and Hale, S. A. Why human-ai relationships need socioaffective alignment, 2025. URL https://arxiv.org/abs/2502.02528.

Lahlou, S. Mitigating societal cognitive overload in the age of ai: Challenges and directions, 2025. URL https://arxiv.org/abs/2504.19990.

Laitinen, A. and Sahlgren, O. Ai systems and respect for human autonomy. *Frontiers in Artificial Intelligence*, 4:705164, 2021. doi: 10.3389/frai.2021.705164. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.705164/full.

LeCun, Y. I think that the magnitude of the ai alignment problem has been ridiculously overblown & our ability to solve it widely underestimated, March 2023. URL https://www.linkedin.com/posts/yann-lecun_i-think-that-the-magnitude-of-the-ai-alignment-activity-7041561014048923648-16XJ/. LinkedIn post.

Lee, M. H., Montgomery, J. M., and Lai, C. K. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans. FAccT '24, pp. 1321–1340, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658975. URL https://doi.org/10.1145/3630106.3658975.

Legg, S. and Hutter, M. A universal measure of intelligence for artificial agents. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pp. 1509–1510, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., and Ghanem, B. Camel: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a3621ee907def47c1b952ade25c67698-Paper-Conference.pdf.

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., and Peng, W. A survey on hallucination in large vision-language models, 2024a. URL https://arxiv.org/abs/2402.00253.

Liu, Y., Yang, K., Qi, Z., Liu, X., Yu, Y., and Zhai, C. X. Bias and volatility: A statistical framework for evaluating large language model's stereotypes and the associated generation inconsistency. In *NeurIPS*, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c6ec4a25a11393f277cfd64b7ea4d106-Paper-Datasets_and_Benchmarks_Track.pdf.

Longoni, C., Fradkin, A., Cian, L., and Pennycook, G. News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 97–106, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533077. URL https://doi.org/10.1145/3531146.3533077.

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., and Hobbhahn, M. Frontier models are capable of in-context scheming, 2025. URL https://arxiv.org/abs/2412.04984.

Milmo, D. 'Godfather of AI' shortens odds of the technology wiping out humanity over next 30 years. The Guardian, December 2024. URL https://www.theguardian.com/technology/2024/dec/27/godfather-of-ai-raises-odds-of-the-technology-wip Last modified 2024-12-28.

Moayeri, M., Tabassi, E., and Feizi, S. Worldbench: Quantifying geographic disparities in LLM factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 1211–1228, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0450-5/24/06. doi: 10.1145/3630106.3658967. URL https://doi.org/10.1145/3630106.3658967.

Moon, K., Green, A. E., and Kushlev, K. Homogenizing effect of large language models (llms) on creative diversity: An empirical comparison of human and chatgpt writing. *Computers in Human Behavior: Artificial Humans*, 6:100207, December 2025. ISSN 2949-8821. doi: 10.1016/j.chbah.2025.100207. URL https://www.sciencedirect.com/science/article/pii/S294988212500091X.

Morris, M. R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., and Legg, S. Levels of agi for operationalizing progress on the path to agi, 2024. URL https://arxiv.org/abs/2311.02462.

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P. A., Collins, G. S., and Maruthappu, M. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, 368:m689, March 2020. doi:

10.1136/bmj.m689. URL https://www.bmj.com/content/368/bmj.m689. PMID: 32213531; PMCID: PMC7190037.

Nehring, J., Gabryszak, A., Jürgens, P., Burchardt, A., Schaffer, S., Spielkamp, M., and Stark, B. Large language models are echo chambers. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 10117–10123, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.884/.

Nelson, A. Thick Alignment. Keynote speech at the ACM Conference on Fairness, Accountability, and Transparency (FAccT'23), Chicago, IL, June 2023. URL https://www.youtube.com/watch?v=Sq_XwqVTqvQ. Accessed: 13 June 2025.

Neumann, T., De-Arteaga, M., and Fazelpour, S. Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1504–1515, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533205. URL https://doi.org/10.1145/3531146.3533205.

Ng, A. Quote origin: I don't work on preventing ai from turning evil for the same reason that i don't work on the problem of overpopulation on the planet mars, 2015. URL https://quoteinvestigator.com/2020/10/04/mars/.

Ngo, R., Chan, L., and Mindermann, S. The alignment problem from a deep learning perspective, 2025.

OpenAI. Openai GPT-4.5 system card. Technical report, OpenAI, February 2025a. URL https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf. System Card, February 27, 2025.

OpenAI. Openai o3 and o4-mini system card. Technical report, OpenAI, 2025b. URL https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf. System Card, April 16, 2025.

OpenAI. Openai charter. https://openai.com/charter/, n.d. Accessed: 2025-6-26.

Oremus, W. Elon musk promised an anti-'woke' chatbot. it's not going as planned. https://www.washingtonpost.com/technology/2023/12/23/grok-ai-elon-musk-x-woke-bias/, 2023. [Accessed 18-06-2025].

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Plurality Institute. Exploring opportunities for llms in public discourse. Plurality Institute (blog post), April 2025. URL https://www.plurality.institute/blog-posts/exploring-opportunities-for-large-language-models

Prabhakaran, V., Mitchell, M., Gebru, T., and Gabriel, I. A human rights-based approach to responsible ai, 2022. URL https://arxiv.org/abs/2210.02667.

Prunkl, C. Human autonomy at risk? an analysis of the challenges from ai. *Minds and Machines*, 34(3):1–21, 2024. doi: 10.1007/s11023-024-09665-1.

Raji, I. D., Kumar, I. E., Horowitz, A., and Selbst, A. D. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 959–972, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533158. URL https://doi.org/10.1145/3531146.3533158.

Roytburg, D. and Miller, B. Mind the gap! pathways towards unifying ai safety and ethics research, 2025. URL https://arxiv.org/abs/2512.10058.

Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, October 2019.

Russell, S. and Norvig, P. *Artificial Intelligence: A Modern Approach (3rd Edition)*. Prentice Hall, 2010.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *ICML*, 2023.

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., and Kenton, Z. Goal misgeneralization: Why correct specifications aren't enough for correct goals, 2022. URL https://arxiv.org/abs/2210.01790.

Sharma, N., Liao, Q. V., and Xiao, Z. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642459. URL https://doi.org/10.1145/3613904.3642459.

Si, S., Zhao, H., Chen, G., Gao, C., Bai, Y., Wang, Z., An, K., Luo, K., Qian, C., Qi, F., Chang, B., and Sun, M. Aligning large language models to follow instructions and hallucinate less via effective data filtering, 2025. URL https://arxiv.org/abs/2502.07340.

Smith, J. J., Satwani, A., Burke, R., and Fiesler, C. Recommend me? designing fairness metrics with providers. FAccT '24, pp. 2389–2399, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3659044. URL https://doi.org/10.1145/3630106.3659044.

Stark, L. Animation and artificial intelligence. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, pp. 1663–1671, Rio de Janeiro, Brazil, 2024. ACM. doi: 10.1145/3630106.3658995. URL https://doi.org/10.1145/3630106.3658995.

Stoev, T., Yordanova, K., and Tonkin, E. L. Experiencing annotation: Emotion, motivation and bias in annotation tasks. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pp. 534–539, 2023. doi: 10.1109/PerComWorkshops56833.2023.10150364.

Stray, J., Vendrov, I., Nixon, J., Adler, S., and Hadfield-Menell, D. What are you optimizing for? aligning recommender systems with human values, 2021. URL https://arxiv.org/abs/2107.10939.

Sucholutsky, I. and Griffiths, T. L. Alignment with human representations supports robust few-shot learning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 73464–73479, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e8ddc03b001d4c4b44b29bc1167e7fdd-Paper-Conference.pdf.

Sutton, R. The reward hypothesis. *http:// incompleteideas.net/rlai.cs.ualberta.ca/RLAI/rewardhypothesis.html*, 2004. URL http://incompleteideas.net/rlai.cs.ualberta. Accessed 2025-06-11.

Tang, A. Ai alignment cannot be top-down, 2025. URL https://ai-frontiers.org/articles/ai-alignment-cannot-be-top-down. Guest Commentary.

Terzis, P., Veale, M., and Gaumann, N. Law and the emerging political economy of algorithmic audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, FAccT '24, pp. 1–13, New York, NY, USA, 2024. ACM. ISBN 979-8-4007-0450-5/24/06. doi: 10.1145/3630106.3658970. URL https://facctconference.org/static/papers24/facct24-85.pdf.

The Collective Intelligence Project. Introducing the collective intelligence project: Solving the transformative technology trilemma through governance r&d. https://static1.squarespace.com/static/631d02b2dfa9482a32db47ec/t/63e01fcbf73bb003a722792f/1675632588509/CIP+Whitepaper.pdf, 2023. White paper; accessed June 13, 2025.

Turing, A. M. Intelligent machinery, A heretical theory. In Copeland, B. J. (ed.), *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life, Plus the Secrets of Enigma*, pp. 472–475. Oxford University Press, Oxford, 2004. doi: 10.1093/oso/9780198250791.003.0018. URL https://uberty.org/wp-content/uploads/2015/02/intelligent-machinery-a-heretical-theory.pdf. Originally presented c. 1951 on the BBC programme *The '51 Society*.

Varshney, K. R. Decolonial AI alignment: Openness, viśeṣa-dharma, and including excluded knowledges, May 2024. URL https://arxiv.org/abs/2309.05030. arXiv:2309.05030v3, last revised 2024-05-02.

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., and Isaac, W. Sociotechnical safety evaluation of generative ai systems, 2023. URL https://arxiv.org/abs/2310.11986.

Williams, A. and Srnicek, N. *Accelerate Manifesto: For an Accelerationist Politics*. Gato Negro Ediciones, Mexico City, 2017. ISBN 978-607-9745-04-2. URL https://anagrambooks.com/accelerate-manifesto. 60 pp., paperback.

Yarom, M., Bitton, Y., Changpinyo, S., Aharoni, R., Herzig, J., Lang, O., Ofek, E., and Szpektor, I. What you see is what you read? improving text-image alignment evaluation. In *NeurIPS*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/056e8e9c8ca9929cb6cf198952bf1dbb-Paper-Conference.pdf.

Yeh, M.-H., Wang, J., Du, X., Park, S., Tao, L., Im, S., and Li, Y. Challenges and future directions of data-centric ai alignment, 2024. URL https://arxiv.org/abs/2410.01957. ICML 2025 position paper.

Yudkowsky, E. Coherent Extrapolated Volition, 2004.

Yudkowsky, E. Artificial Intelligence as a positive and negative factor in global risk. In Rees, M. J., Bostrom, N., and Cirkovic, M. M. (eds.), *Global Catastrophic Risks*, pp. 0. Oxford University Press, July 2008. ISBN 978-0-19-857050-9. doi: 10.1093/oso/9780198570509.003.0021.

Zhi-Xuan, T., Carroll, M., Franklin, M., and Ashton, H. Beyond preferences in ai alignment. *Philosophical Studies*, 182(7):1813–1863, July 2025. doi: 10.1007/s11098-024-02249-w. URL https://link.springer.com/article/10.1007/s11098-024-02249-w. Published online 09 Nov 2024.

Zhuang, S. and Hadfield-Menell, D. Consequences of misaligned ai. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15763–15773, 2020. ISBN 9781713829546. URL https://proceedings.neurips.cc/paper/2020/file/b607ba543ad05417b8507ee86c54fcb7-Paper.pdf.