

# A Taxonomy of AI Value Alignment: Subproblems and Solutions

Daniel R. Harris, Karina Vold, Harsh Kumar, Ashton Anderson

University of Toronto

danielr.harris@utoronto.ca, karina.vold@utoronto.ca, harsh@cs.toronto.edu, ashton@cs.toronto.edu

## Abstract

The value alignment problem (VAP) is widely regarded as one of the central challenges in AI safety and ethics. Broadly construed, VAP concerns how to design advanced AI systems whose behaviors (outputs) align with human values. Although there is agreement about the importance of this objective, most contemporary research treats value alignment as a single technical problem. In this paper, we argue that this framing oversimplifies a diverse set of conceptual, ethical, and technical challenges. We distinguish several related but importantly different value alignment subproblems, including identifying values, prioritizing values, imperfect specification, value anthropocentrism, moral imperfection, inverse alignment, and the provability challenge. We argue that these are distinct subproblems in the sense that progress on one will not constitute a solution to the others, or to VAP as a whole. We further argue that this taxonomy of subproblems is necessary to advance the fields of AI safety and alignment research. By clearly distinguishing different subproblems, we show how each will require different kinds of solutions: while some alignment problems may admit technical approaches, others are fundamentally social or political in nature and may resist definitive technical resolution. We conclude that VAP should not be understood solely as a technical optimization problem, but also as an ongoing process of responsible design and ethical deliberation.

## Introduction

Contemporary AI researchers widely consider the value alignment problem (VAP) to be one of the greatest challenges in the field of AI safety (Russell 2019; Bengio et al. 2026). This problem has long been recognized: Alan Turing

(1951) and Norbert Wiener (1960) both expressed concerns about powerful AI systems behaving in ways that are not aligned with human interests. But in the past decade VAP has moved from theoretical speculation to a practical and urgent problem. By 2017 the influential Asilomar AI Principles explicitly encoded value alignment as a design principle that should guide safe and ethical AI: “Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation” (FLI 2017). More recently, the centrality of value alignment within the AI safety community is evidenced by workshops and calls for papers on the topic within major AI conferences, including ICML, NeurIPS, and ACM FAccT, as well as recent calls for funding by national AI Safety agencies.<sup>1</sup> The problem has even been the focus of multiple book length treatments (e.g., Russell 2019; Christian 2020; LaCroix 2025).

At a high level there is consensus about what the problem is. Broadly speaking, we can say that VAP is the problem of how to build an advanced AI system with behaviors (outputs) that are aligned with human values.<sup>2</sup> However, this statement of the problem oversimplifies a complicated set of subproblems that would have to be resolved in order to solve VAP. In this paper we aim to taxonomize the various subproblems that compose VAP. This taxonomy is important not just for providing conceptual clarity to a growing and important subfield of AI safety research, but for how it can advance progress within the field. For example, we will show how the current lack of nuance within the literature on VAP has led to unhelpful ambiguity in the solutions being

---

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

<sup>1</sup> For example, NeurIPS 2024 had a Pluralistic AI Alignment workshop (<https://neurips.cc/virtual/2024/workshop/84737>); ICML 2026 will have a Pluralist Alignment Workshop (<https://pluralistic-alignment.github.io>). In 2025 the UK’s AI Safety Institute launched a project stream dedicated to AI alignment, “The Alignment Project” (<https://alignmentproject.aisi.gov.uk>). In 2025, Canada’s AI Safety Institute (<https://cifar.ca/ai/caisi/>) had calls for funding to support AI

safety research which explicitly sought projects focused on AI alignment.

<sup>2</sup> This statement of VAP is inherited from the historically contingent tradition we mention — dating back to thinkers like Alan Turing and Norbert Weiner, and into contemporary discussions by computer scientists, like Stuart Russell. One reviewer suggests that this tradition implicitly assumes a narrow site for normative interventions, but we resist this. While it’s true that much of the AI alignment subfield focusses on technological solutions, some of the subproblems that we identify in this paper (in the next section) are largely social or political and, so we argue, may have no possible technical solution.

developed. We begin by providing a taxonomy of subproblems and then expand on the implications of this taxonomy for advancing progress in AI safety and alignment research.

## The Many Value Alignment Problems

Using the method of conceptual analysis, we clarify VAP, showing that the problem actually consists of many distinct subproblems. In other words, to solve VAP one would have to solve each of these subproblems. Furthermore, these subproblems are distinct in the sense that a solution to one subproblem does not solve the other subproblems. Hence, the subproblems we identify are not just logically distinct, they require distinct solutions.<sup>3</sup> We want to make two further points explicit.

First, the seven subproblems we identify are not intended to be exhaustive, we recognize that some of the subproblems listed may need to be divided up further, or additional subproblems may need to be identified and added in. We propose this taxonomy as a first step, an initial mapping to be refined and built upon as the field of AI alignment continues to grow. We also acknowledge that the alignment research community is becoming more aware that VAP is not just one problem, but rather a set of problems (e.g., Bengio et al. 2026; Ji et al. 2025). However, a clear taxonomy of the subproblems that compose VAP and their potential solutions is still lacking, and this dearth of conceptual clarity slows progress (as we argue in a later section).

Second, while we draw on much of the current literature on alignment to give examples of the subproblems that we identify, and much of this literature is focused on statistical models, including large language models (LLMs), none of the subproblems are unique to LLMs, or bottom-up statistical models. These subproblems are agnostic about the kind of AI architecture, or system, employed: whatever technique (symbolic, statistical, reinforcement learning, etc.) one uses to build an advanced AI system, one would inevitably have to confront and address each of the subproblems.

### Subproblem 1: Identifying Human Values

Arguably, the subproblem which has received the most attention to date is the challenge of *identifying human values* (Gabriel 2020).<sup>4</sup> In our view, this subproblem is comprised of a group of open questions (or sub-subproblems) that will each need to be addressed. Firstly, there is the question of *whose* values we should align an AI system with — what

Kenton et al. (2021) refers to as the “target” of alignment. Several candidate options have been identified in the literature, such as a user (or set of users), a developer, a company, a specific society and/or country, or some combination thereof (Kenton et al. 2021; Gabriel et al. 2024; LaCroix 2025). Secondly, there is a *timescale* question. If, for example, we decide to align an AI system with a user, developer, or company, should we optimize for their short-term or longer-term values? If, instead, we select a society and/or country, do we optimize for the values of current members or future populations? There is also the more general question of whether we should assume that values/preferences are static or dynamic; and if we opt for the latter, whether an AI system should optimize for an individual’s current values/preferences, or those of their past and/or future (Zhi-Xuan et al. 2024; Carroll et al. 2024). Thirdly, there is the question of *which* values (Gabriel 2020). Moral philosophers, social psychologists, and cultural anthropologists have long noted that different cultures seem to have wide variation between their respective values, or at least how those values are manifested in behaviours (Brandt 1954; Cohen and Nisbett 1996; Chagnon 1996; Doris and Stich 2007). If profound moral disagreement across cultures does in fact exist, this need not entail value relativism. It does, however, make the problem of aligning AI with human values considerably more difficult by introducing the possibility that there is no single, unified set of values to align with, but rather a plurality of legitimate values (Vold and Harris 2021). This could raise the prospect of some robust form of value pluralism, for example. Value pluralism is the view that there are many different and sometimes irreducible moral values (Graham et al. 2013). While by way of contrast, a value monist holds that there is one value that all other values reduce to (Cave et al. 2019). Typically, for value monists, that value might be something like happiness or pleasure. Aristotle, for example, famously defends a form of value monism in the *Nicomachean Ethics*. The challenge for the value monist is to tell a story of how it is that all values reduce to that one value, while a value pluralist would have to identify what the complete set of irreducible values are (Vold and Harris 2021).

### Subproblem 2: Prioritizing Values

Relatedly, there is the subproblem of what to do when values come into conflict (Gabriel 2020; Anwar et al. 2024; Gabriel et al. 2024; Carroll et al. 2024; Zhi-Xuan et al. 2024;

<sup>3</sup> One reviewer suggested that some of the subproblems we identify rely on certain metaethical assumptions and may collapse or disappear depending on what metaethical views one adopts. For example, they suggested that subproblem 5 (moral imperfection and progress) only makes sense if subproblem 1 (identifying values) is not a true problem, as (on this reviewer’s view) moral progress requires moral realism. It is worth addressing this concern explicitly, and making clear why it is incorrect (in our view), in case other

readers have similar reactions. The idea that moral progress requires moral realism is naïve. Not all moral subjectivists reject moral progress, just as many contemporary moral realists deny that there is a single universally correct action, or set of values, in every case. The metaethical landscape is rich and nuanced. Of course, some metaethical commitments may dissolve certain subproblems, but this is not necessarily the case.

<sup>4</sup> Gabriel (2020) refers to this as the “normative” component of VAP.

Gabriel and Keeling 2025). This could manifest in a number of different ways. At the level of the individual (e.g., a user), optimizing for short-term values might satisfy a subjective state of well-being at the cost of longer-term values that we try and cultivate, like courage or generosity.<sup>5</sup> This is a familiar problem in philosophy, exemplified by the *akratic* ('weak-willed') person, who knowingly acts against her better judgement by prioritizing short-term values (Pettit 2003; May and Holton 2012). Philosophical discussions of *akrasia* date back at least as far as Plato's *Protagoras*, and within the literature there are many sophisticated attempts aimed at reconciling weakness of will (Stroud and Svirskey 2025). It seems clear that the intended context of use will be important for how this trade-off gets navigated: sometimes we want a more servile, agreeable AI system, but other times we may want (and need) more friction. As Anwar et al. (2024) point out: "within an individual, multiple values can coexist, sometimes complementing each other but also occasionally creating internal conflict" (p. 81). Additional individual-centric value tensions are also possible if we were to assume that values/preferences are dynamic as opposed to static (Zhi-Xuan et al. 2024; Carroll et al. 2024). In principle, it is possible for a value held by an individual at time  $t$  to conflict with another, newly acquired, value at time  $t_1$  (Carroll et al. 2024).

At the group level, the values of the current members of a society (or citizens of a country) may, under certain conditions, conflict with the values of future populations. For example, finite resource acquisition (e.g., crude oil extraction) by present-day societies could have negative downstream impacts on future societies (e.g., by contributing to resource scarcity, environmental degradation, etc.) If, in response to the *timescale* component of the first subproblem, we opt to align an AI system with both the current members of a society and future populations, then the question of which populace's values should be prioritized when they come into conflict becomes salient. If, however, we opt instead to align AI solely with the values of a current populace, similar sorts of trade-off dilemmas could likewise arise. Imagine that an AI is optimizing for the values of country  $A$ , but is being used by the citizens of country  $B$  whose values are in tension with  $A$ . Which set of values,  $A$  or  $B$ , should it prioritize?

Additionally, similar sorts of value trade-offs may also arise if VAP researchers opt for a value pluralist framework over value monism. Cave et al. (2019) note that "[u]nlike monists, value pluralists tend to believe that there are at least some, perhaps many, complex moral dilemmas that result from a conflict between competing and incommensurable values" (pp. 569-570). Suppose, for example, that researchers select some pluralist set of values,  $v$ , to align an AI system with, and that  $v$  includes the values of fairness and well-being maximization. Suppose further that this AI system is

tasked with making organ transplant recommendations, and it now must decide whether to give a life-saving organ to an elderly patient who has spent several years on the donation waitlist or a child who waited less time but could benefit more. The AI must trade-off the value of fairness (help whoever has waited the longest) with the value of maximizing overall well-being (help the child, who could gain the most life-years) (Cave et al. 2019). Which value will it prioritize? We encounter these ethical dilemmas frequently in life and do our best to navigate the trade-offs (Brundage 2018; Cave et al. 2019). An AI system aligned to a pluralist theory of value will need to be designed to make these kinds of trade-offs in a reliably ethically acceptable way.

### Subproblem 3: Imperfect Specification (a.k.a. The King Midas problem)

Once we have both identified and agreed on our values, and devised a method to adequately prioritize them if, and when, they are in tension, the next step is to figure out how to impart them to an AI system in a way that avoids undesirable outcomes. Versions of this subproblem have been introduced in the VAP literature through the retelling of the old folklore of King Midas, who faced disastrous outcomes after wishing that everything he touched turned to gold (Russell 2019; Krakovna et al. 2020; Vold and Harris 2021). King Midas never intended for all his nourishment and loved ones to turn to gold, but he was not careful in how he specified his wish. The fate that befell him was the result of his failing to articulate, correctly and exhaustively, what he valued. And the thought is that if in a likewise manner we fail to adequately define the values we imbue AI with, a similar disastrous outcome could befall us when that system instantiates them.

While the King Midas folklore captures an important aspect of this subproblem, it does not encapsulate the challenge in its entirety. Even if we were to be careful in the way we specify our values, an AI system may ultimately interpret and put those values into practice in ways that we simply fail to anticipate. And much like in the case of King Midas, the result could prove calamitous. This issue shares some similarities with Bostrom's (2014) concept of *perverse instantiation*, and the technical challenge of *specification gaming* (sometimes called 'reward hacking') prevalent in certain types of machine learning (ML). Krakovna et al. (2020) defines the latter as "a behavior that satisfies the literal specification of an objective without achieving the intended outcome". A widely discussed example is a reinforcement learning (RL)-based robotic arm given the task of stacking Lego; whereby the intended outcome was for the robot to place a red block onto a blue one (Popov et al. 2017; Krakovna et al. 2020). It was designed such that the RL agent

<sup>5</sup> AI models that optimize for alignment using RLHF seem to fall into this trap, as evidenced by the sycophantic behavior of

ChatGPT-4o relative to ChatGPT-5, which were explicitly refined to correct for this (OpenAI 2025).

was “rewarded for the height of the bottom face of the red block when it is not touching the [blue] block”, and the thought was that this would lead the robotic arm to pick up, and stack, the red block atop the blue block (Krakovna et al. 2020). To the team’s surprise though, the optimal learned policy caused it to “simply [flip] over the red block to collect the reward” (ibid). This is not an isolated incident. Several other instances of RL-based specification gaming have also been reported (Clark and Amodei 2016; Krakovna et al. 2020; Buckner 2024) and, more recently, several cases associated specifically with LLMs have been identified (Denison et al. 2024; Bondarenko et al. 2025; Malmqvist 2026).

#### Subproblem 4: Value Anthropocentrism

Another subproblem researchers face is whether our alignment efforts should be restricted to human values or extended to also include non-human life (e.g., sentient animals, ecologies, etc.).<sup>6</sup> The standard definition of VAP assumes an anthropocentric notion of value (Peterson 2019; Raper 2025), however whether this is the ethically “correct” approach is, at the very least, an open philosophical question. Within environmental ethics, a distinction is commonly made between anthropocentric moral theories which do not extend moral status to non-human entities, and non-anthropocentric moral theories which do (Humphreys 2014; Peterson 2019). Peterson (2019) argues that the tension between these two views is often “swept under the carpet” in the context of VAP, yet in practice many situations pose a dilemma between either prioritizing human values or foregrounding the environment (p. 22). For example, consider a case where a logging company is using advanced, autonomous, AI to identify and harvest commercially valuable timber in undeveloped forest tracts of the Congo Basin rainforest. The system locates a zone of interest, but it surrounds a gorilla encampment which will be irreparably damaged should it proceed with the tree extraction process. Presumably, if the system was designed to align merely with human values, it would begin harvesting as non-human interests are not a consideration it must weigh in deliberating over what action to take. This would, however, be ethically problematic under certain non-anthropocentric moral theories (e.g., sentientism and biocentrism) given that such an act would have failed to consider the moral status of the gorillas.

Alternatively, these sorts of trade-off conflicts could also emerge if the notion of value assumed in VAP was extended to also include non-human values. Consider how such an AI system would deal with Regan’s (1983) well-known lifeboat

thought experiment. There are five survivors — “four normal adults and a dog” — stranded on a lifeboat which only has enough room to sustain four of them. “Someone must go”, Regan writes, “or else all will perish” (p. 324). Here we have human and non-human values in tension, and an AI system aligned with the values of both. Suppose it was tasked with deciding which survivor to remove from the lifeboat, who would it select? Hence, if researchers do opt to align an AI system with human and non-human values, it will need to be designed such that it has a way to prioritize one set of values over the other in situations where there is a conflict between the two (i.e., navigate subproblem 2). Moreover, whatever decision procedure is ultimately selected will need to be justified philosophically, and here, once again, the debate between anthropocentric and non-anthropocentric moral theories rears its head.

#### Subproblem 5: Moral Imperfection and Progress

Morally speaking, humans are far from perfect. History is replete with examples of how poorly humans treat one another, and even today, despite our technological sophistication, such mistreatment persists. We also tend to harm other species and routinely damage the environment in unsustainable ways. A fifth subproblem centers around this issue of human moral imperfection. Vold and Harris (2021) argue that if VAP researchers were to merely align AI with our current value standards, then they run the risk of these systems mirroring or amplifying human “fallible, inconsistent, and complacent moral natures” (p. 734). Moreover, they note that even if we were to somehow resolve this issue by devising a way to align AI “with (only) the better parts of our current values, we would not want [these] AI systems to codify” said values such that it stymies, or does not allow for, future moral progress (ibid). The values held by many today differ, quite substantially, from the prevailing values of the past. If we were to imagine that AI existed two hundred years ago, and, moreover, that these machines codified the value standards of that time, it is unlikely that present-day humanity would view them as adequately aligned. This is due, in large part, to a failure to have incorporated the moral progress that has been made in the centuries that separate us. In much the same way, it stands to reason that humans two hundred years in the future would likewise view the values our AI systems have codified (should we design them to do so) as (at least partially) misaligned with the current value standards of their time.

<sup>6</sup> Gabriel (2020) poses a version of this question in his (critical) examination of an approach to VAP which assumes that the goal is to align AI systems with the “interests or well-being” of an agent. His contention is that such an approach would “need principles for

deciding whose interests or needs count for the purpose of AI alignment” and this, in turn, gives rise to the question: “is it only humans whose interests matter, or do the interests of nonhuman animals and other forms of sentient life count as well?” (p. 422). This section draws inspiration from this insight.

## Subproblem 6: Inverse Alignment

Another subproblem is *inverse alignment*: the possibility that misalignment will be dissolved not by aligning AI systems to human values, but by instead shaping human values to become aligned with AI systems. Take, for example, social media RL algorithms that recommend content to users. Such algorithms are typically designed to “maximize a measurable proxy” such as a ‘clickthrough’ (Dung 2023, p. 10, footnote) and, given this, one might reasonably expect it to learn to suggest content that aligns with a user’s interests (or preferences) as this would increase the likelihood of a click (Russell 2019; Whittlestone et al. 2020). However, some have warned that the learned policy could also converge on *shaping* user preferences to make them more consistent and predictable, and hence easier to satisfy (Russell 2019; Whittlestone et al. 2021; Dung 2023; Gao 2023).

Discussing the need for ‘socioaffective’ alignment, Kirk et al. (2025) describes scenarios like these as examples of *social reward hacking*: “the use of social and relational cues by an AI to shape user preferences and perceptions in a way that satisfies short-term reward in the AI’s objectives (e.g., increased conversation duration, information disclosure or positive ratings on responses) over long-term psychological well-being” (p. 5). Current AI behavior that exemplifies this is LLM sycophancy (OpenAI 2025; Kirk et al. 2025). But the problem of humans (implicitly) adapting their preferences and values to accommodate AI systems is not uniquely ‘socioaffective’ and the methods through which humans adapt are not always examples of social reward hacking, which is why we opt for the broader term of ‘inverse alignment’. Humans may adapt their values and preferences to AI systems that are purely functional, not emotionally aware and not tracking or managing any of their social or affective states, such as a sorting algorithm, a navigation app, or even a thermostat. Of course, inverse alignment, like other VAP subproblems, becomes most worrisome when the AI system involved is more advanced, more autonomous, more general, and more capable (e.g., increased instrumental rationality or direct world modification, etc.).

## Subproblem 7: The Provability Challenge

If our aim is to build advanced, autonomous, AI that can interact with and assist humans in performing complex tasks, then it is important that we can verify that these systems act in ways that we expect and desire them to in practice, especially in future (untested) scenarios. Part of this will involve ensuring that said behavior is reliably aligned with human values. The *provability challenge* is the subproblem of how we test for this (Russell 2019). The issue is also sometimes referred to as the need for ‘assurance’ (Ji et al. 2025) or ‘verifiability’ (Dalrymple et al. 2024). Solutions to address the provability challenge usually appeal to formal techniques

for safety evaluations (e.g., sandboxing, red-teaming, adversarial testing) and mechanistic interpretability (e.g., circuit analysis, causal interventions, activation tracing).

If an AI system is narrow enough, it is reasonable that we could achieve provable alignment. But the wider (or more general) the application of an AI system is, the harder it will be to test for appropriate alignment across a large, or potentially infinite set of applications. Hence, comprehensive safety evaluations before release becomes less possible the more general, powerful, and autonomous a system is. Likewise, the quest for assurance is also hindered by the familiar black box issue: the fact that current AI systems have so many hidden network layers that their outcomes are uninterpretable even to the designers of the system. This explains the focus on mechanistic interpretability but, again, efforts here become more challenging when dealing with more advanced AI systems. Enormous transformer models, for example, have resisted standard interpretability methods both because of the amount of data they trained on and because their “interpretable circuits” (or internal representations) are distributed and entangled rather than modular (Elhage et al. 2022).

Another aspect of this subproblem is a concern about *deceptive alignment* — the risk that a model trained to optimize for one objective may internally learn another and then begin to strategically behave as though it were aligned during training and testing because that behavior is instrumentally useful (Hubinger et al. 2019; Dung 2023; Ngo et al. 2024; Hendrycks 2024). Greenblatt et al. (2024) have dubbed this ‘alignment faking’, while early discussions of this possibility, which were at the time purely hypothetical, called this ‘strategic behavior’ (e.g., Bostrom 2014). Today, AI researchers are actively monitoring for cases of deceptive alignment (Carranza et al. 2023; Ji et al. 2025). The possibility of deceptive alignment only emerges from models that are reasonably advanced and, thus, underscores how the challenge goal of provability will become increasingly difficult as models become more sophisticated.

## Proposed VAP Solutions and Their Limits

Some of the above subproblems have been recognized by the value alignment research community and substantial progress has been made towards better aligning current AI systems. Developers are much better at steering model behavior than we were a decade ago, often through a variety of post-training methods (e.g., reinforcement learning from human feedback (RLHF)), rule-based alignment (e.g., Bai et al. 2022), evaluation, and oversight techniques. However, any given proposed solution wouldn’t solve VAP broadly but would only address a particular subproblem. Moreover, when a new alignment technique emerges it can be unclear exactly what subproblem it’s meant to address, leaving its scope and promise ambiguous and even open to criticism. New techniques might be criticized for not solving all of VAP, for example, when such a bar is too high. Or, when

multiple alignment solutions are on the table they can be misconstrued, or misunderstood, as rival candidates when they are in fact directed at different subproblems.

Work on value alignment is extensive and dates back decades, hence, we want to emphasize at the outset that this section is not a comprehensive overview of proposed solutions. In this section we offer a few illustrative examples of how proposed solutions to VAP have really only addressed particular subproblems and how making the aims of these solutions clearer and more explicit would be beneficial. We end this section with a table summarizing the taxonomic structure we propose in the paper.

## Proposed Solutions to Subproblem 1: Identifying Human Values

### Inverse Reinforcement Learning

The subproblem of identifying human values is, by and large, a theoretical, or philosophical, problem. This is particularly evident with respect to the question of *which* values we should assume. As we have seen, there are a multitude of viable and competing theories to select among, and no clear consensus as to which should be adopted. Indeed, in one form or another, philosophers have been debating many of the issues underpinning this subproblem for decades (if not centuries), and while substantial progress has undoubtedly been made, a resolution still eludes us. Perhaps, though, we need not trek into these muddy theoretical waters in search of an answer. Instead, we frame it as a technical subproblem and offload the task of finding a “solution” to the very AI system whose behavior we aim to align. One such approach, discussed by Gabriel and Ghazavi (2021), is for VAP researchers to pursue a so-called “bottom-up” strategy, whereby an AI “seeks to infer human preferences about values from observed behavior or feedback” (p. 9) using *inverse reinforcement learning* (IRL).<sup>7</sup> In many domains, desirable behaviour is easier to recognize than to specify. IRL attempts to exploit this asymmetry. Rather than requiring researchers to specify the reward signal beforehand, it asks: “Given the observed behaviour, what reward signal, if any, is being optimized?” (Russell 1998, p. 101). This is different from standard RL where a reward function is specified in advance, and the agent learns a policy that maximizes expected reward (Sutton and Barto 1998).

One challenge with this approach is associated with IRL more generally — what Zhou and Li (2024) term “reward

ambiguity”, and what Ng and Russell (2000) label “degeneracy” (Gabriel 2020). The problem is that for any example of observed behavior, multiple reward functions could be inferred by an IRL agent, each explaining the behavior equally well (Zhou and Li 2024; Gabriel 2020; Christian 2020). For example, suppose it was observing a multinational coffee retail corporation running a promotion in which a percentage of their coffee sales would be donated to a charitable cause (e.g., an environmental initiative). It must now try to infer the underlying value that motivated them. One explanation of their behavior is that the corporation values altruism: the organization genuinely wants to champion the cause of environmental sustainability and do their part in helping slow down climate change. Another, equally possible, explanation is that the corporation’s behavior is entirely motivated by self-interest, as would be the case if the promotion was part of a larger “cause marketing”<sup>8</sup> public relations strategy designed and implemented to improve their public image or sales. Here we have two possible explanations of the same observed behavior, and the two values identified are not only distinct from each other, but in tension. Presumably we would want an AI system to align with the value of altruism and not self-interest. Yet, both values can be inferred in this instance.

Even if we were to assume, for the sake of argument, that researchers can resolve this issue of reward ambiguity (which we are certainly open too) and, by extension, resolve the subproblem of identifying human values, we would still not be out of the woods just yet. As noted earlier, humans are not reliable moral exemplars. Our moral imperfections will be embedded in whatever dataset or “expert” we ultimately select for training the IRL system. And thus, it stands to reason that whatever value (or set of values) it ultimately infers will reflect that moral imperfection.<sup>9</sup> Hence, by using IRL we might resolve the subproblem of *identifying human values*, but in doing so would walk ourselves straight into the subproblem of *moral imperfection and progress*.

### Pluralistic Alignment Techniques

Despite the centrality of VAP to AI safety, it was not until around 2024 that discussions of moral variation across stakeholders (e.g., different cultures, different users) became mainstream to the AI research community.<sup>10</sup> The AI safety community now widely refers to this work as *pluralistic alignment*, the aim of which is to accommodate a variety of value preferences, rather than assuming a universal

<sup>7</sup> It should be noted that Gabriel and Ghazavi (2021) are critical of this approach and identify several associated challenges that VAP researchers will need to resolve. Gabriel (2020) also examines the utilization of IRL to address this subproblem, and flags some potential challenges associated with it.

<sup>8</sup> Stole (2008) defines *cause marketing* as the use of “purchase-triggered” donations (see p. 21).

<sup>9</sup> Gabriel and Ghazavi (2021) identify some additional training-specific challenges associated with the use of IRL to resolve VAP,

such as the need to address the open question of: “what data to train the system on, how to justify this choice, and how to ensure what the model learns is free from unjustifiable bias” (p. 9).

<sup>10</sup> For example, some impactful research papers that came out in 2024 include Sorensen et al. 2024 and Feng et al. 2024; the first workshop on pluralistic Alignment at NeurIPs also occurred in 2024 (<https://neurips.cc/virtual/2024/workshop/84737>), and now there will be a similar workshop at ICML in 2026 (<https://pluralistic-alignment.github.io>).

set (Bengio et al. 2026).<sup>11</sup> Some proposed technical frameworks that have emerged to achieve pluralistic alignment include, for example, modular pluralism (Feng et al. 2024), multi-objective RL (Vamplew et al. 2024), MaxMin-RLHF (Chakraborty et al. 2024), and PAL—a framework for learning from heterogeneous preferences (Chen et al. 2024). Others have suggested looking beyond technical solutions, drawing solutions from other fields, like social choice theory (Conitzer et al. 2024) or communication theory (Tessler et al. 2024).

Work on pluralistic alignment is important and significant. In the first case, it’s an illustrative example of how interdisciplinary research should shape and inform technical AI safety research, adding nuanced conceptual clarity to advance progress. It’s also a recent example of how the alignment community has come to realize how complicated VAP is. Yet, pluralistic alignment, if achieved, would still fall short of solving VAP. The proposed pluralistic path of building AI systems that recognize and accommodate moral variation across stakeholders would at best address (some aspect of) subproblem 1 (*identifying human values*), leaving the remaining six subproblems still to be addressed.

## Proposed Solution to Subproblem 2: Prioritizing Values

### Resolving Ethical Tensions

Early work in AI ethics saw a proliferation of ethical principles from academic, industry, and governmental organizations (Jobin, Ienca, and Vayena 2019; Floridi et al. 2018; Floridi and Cowsls 2022). There was even a substantial amount of overlap between these sets of principles, despite arising from different stakeholders. But, drawing lessons from other applied fields such as bioethics, moral philosophers quickly pointed out that agreeing on principles is only a first step to achieving ethical AI, as the more challenging work is in deciding how to implement high level principles into the practice of building AI systems, especially when principles are in tension (Floridi 2019; Whittlestone et al. 2019; Morley et al. 2020). For example, fine-grained personal data about patients can improve the accuracy of medical AI predictions, but training AI systems on sensitive and personal information can heighten privacy risks. Hence, we must make decisions about how to design such AI models: which high-level value do we optimize for, accuracy or privacy? Optimizing for accuracy comes into tension with privacy concerns, but if we optimize for privacy (e.g., deciding to use no sensitive or personal data in training) then the quality of the model would drop dramatically.

Following Whittlestone et al. (2019), when ethical values come into conflict like this, we can say that they are in ‘tension’. Examples of ethical tensions in current AI/ML models are replete: model accuracy vs. fairness; model accuracy vs. interpretability; personalization vs. privacy. As noted earlier, other kinds of ethical tensions can also emerge, for example, between prioritizing what is good for individuals (e.g., personalization) vs. social goods (e.g., solidarity), or between what an individual values in the short term (e.g., convenience, entertainment) vs. their longer-term values (e.g., dignity, flourishing).

The good news is that not all ethical tensions are fundamental; some are contingent. In the case of ethical tensions in AI systems, like the one described, it may be that there are technical solutions that resolve this tension. Work on differential privacy, for example, has made significant progress in finding ways to add statistical noise to datasets to protect personal data, rather than removing an individual’s data (Abadi, et al. 2016; Ouadrhiri and Abdelhadi 2022; Cummings et al. 2024). This has made it possible to continue to optimize for accuracy without sacrificing privacy. There has also been important work on deontic logics and normative multi-agent systems (NorMAS) aimed at resolving norm conflicts (e.g., Santos et al. 2017; Olson, Salas-Damian, and Forbus, 2024). These are examples of how technical efforts can help resolve ethical tensions. Many ethical tensions, especially around new technologies, like AI, may turn out to be contingent in nature and, hence, amenable to such technical solutions. But some ethical tensions are more fundamental; these are what some philosophers call *tradeoffs*. In a strict moral tradeoff, it is not possible to optimize for one value without sacrificing another (Whittlestone et al. 2019). In these cases, tough decisions must be made about which value to prioritize.

If we’re lucky enough that most ethical tensions in the design of AI systems turn out to be contingent rather than fundamental, then directing technical research towards resolving these tensions could go a long way towards solving the subproblem of *prioritizing values*. However, what’s most important for the argument in this paper is that solving this subproblem (if possible) would not thereby solve the remaining subproblems. Of course, it’s not clear that all ethical tensions in AI systems are merely contingent. In the next subsection, we turn to the issue of temporal tensions, which may be more fundamental in nature and therefore less tractable through technological progress alone.

### Dynamic Reward Markov Decision Processes

One aspect of the subproblem of prioritizing values is temporal: human values are dynamic, changing over time, and can differ over timescales (e.g., short-term preferences can

<sup>11</sup> It is worth noting that the way the term ‘value pluralism’ is now used in AI research communities (e.g., Kasirzadeh 2024) does not track how the term was originally used in moral philosophy. Our

use of ‘value pluralism’ in this paper tracks its original use from moral philosophy and we have used ‘pluralistic alignment’ to track current discussions in AI research communities.

conflict with long-term goals). Some AI researchers have recently begun to focus on this aspect of VAP (e.g., Klassen, Alamdari, and McIlraith, 2024; Carroll et al. 2024). Carroll et al. (2024), for example, argue that current standard approaches to modeling human preferences use a single, static reward function and, hence, fail to capture this dynamic aspect. They propose a Dynamic Reward Markov Decision Processes (DR-MDPs) to address this issue. DR-MDPs are meant to account for the temporally dynamic preferences of users “by modelling them as changing reward functions” (p.2). DR-MDPs are a good example of a proposed technical solution directed at a particular VAP subproblem. As the authors were likely aware, addressing the problem of *how* to model temporally changing preferences wouldn’t address other VAP subproblems. They point out that DR-MDPs still leave open the question of *which* set of values to prioritize: user<sub>x</sub> at  $t_1 > t_2$ , or user<sub>x</sub> at  $t_2 > t_1$ , so on. Carroll et al. note this: “Ultimately, to align AI systems we must answer the following question: when a person’s preferences change over time, which (aggregation of) preferences should AI systems optimize?” (2024, p. 1). This aspect of the subproblem is more normative and may not have a technical solution.

### Proposed Solution to Subproblem 3: Imperfect Specification (a.k.a. The King Midas Problem)

#### Reinforcement Learning from Human Feedback

At a high-level, one of the main approaches to overcoming the challenge of imperfectly stated goals (or imperfect proxy reward functions) has been to find new ways to model rewards. In other words, if humans aren’t good at stating our values, then perhaps AI systems would be better off inferring them ‘bottom-up’ by observing human behavior or text. IRL, as mentioned earlier, is one prominent example of this, another is RLHF (Christiano et al. 2017). In RLHF, human rankings of model outputs are used to train a reward model to infer value preferences from these rankings. The reward model is then used to fine-tune the language model with some optimization method, so that it is more likely to generate outputs that the reward model predicts people would prefer.

RLHF has proven effective at reducing harmful speech in large language models (Ziegler et al. 2019; Ouyang et al. 2022) and, over the past five years, it has grown tremendously in popularity, becoming one of the most widely used fine-tuning techniques for transformer models. But despite improving efforts towards resolving the *imperfect specification* subproblem, models that use RLHF are still susceptible to reward hacking (Ibarz et al. 2018). This is because the learned reward model can itself be ‘hacked’ or exploited. Sycophantic LLM behavior is often attributed to RLHF-style post-training (e.g., OpenAI 2025), for example, suggesting that rather than disappearing, reward hacking behaviors simply moved up one layer of abstraction. Indeed, despite great efforts to reduce reward hacking, there is some

recent consensus that the problem remains unsolved (Gabriel et al. 2024). Work on reward-model over-optimization shows that because learned reward models are imperfect proxies, optimizing them too strongly can degrade performance relative to the underlying ground-truth objective (Gao et al. 2023). Likewise, formal work on reward hacking suggests limits to the possibility of designing proxy rewards that are robust to optimization (Skalse et al.2022).

| Diagnostic Question                                            | Characteristic Failure                              | Candidate Solutions                                         |
|----------------------------------------------------------------|-----------------------------------------------------|-------------------------------------------------------------|
| <b>#1 Identifying Human Values</b>                             |                                                     |                                                             |
| Whose and which values should the system align with?           | Wrong target or incomplete value set                | Participatory elicitation, IRL, pluralistic alignment       |
| <b>#2 Prioritizing Values</b>                                  |                                                     |                                                             |
| What happens when values conflict?                             | Arbitrary or unjust trade-offs                      | Deliberation, social choice, multi-objective RL, DR-MDPs    |
| <b>#3 Imperfect Specification</b>                              |                                                     |                                                             |
| Can values be operationalized without unintended consequences? | Reward hacking, specification gaming                | RLHF, reward modeling, red-teaming, formal methods          |
| <b>#4 Value Anthropocentrism</b>                               |                                                     |                                                             |
| Do non-human entities count?                                   | Human-centered optimization harms animals/ecologies | Expanded stakeholder analysis                               |
| <b>#5 Moral Imperfection and Progress</b>                      |                                                     |                                                             |
| Should AI mirror current human values?                         | Locks in unjust or morally immature norms           | Moral uncertainty, revision mechanisms, democratic updating |
| <b>#6 Inverse Alignment</b>                                    |                                                     |                                                             |
| Does the system reshape human values?                          | Humans adapt to AI incentives or affordances        | Longitudinal monitoring, autonomy-preserving design, audits |
| <b>#7 The Provability Challenge</b>                            |                                                     |                                                             |
| Can alignment be verified across contexts?                     | Hidden or untested misalignment, deception          | Safety evaluations, interpretability, audits, sandboxing    |

Table 1: A Taxonomy of AI Value Alignment Problem. (This table summarizes the set of seven subproblems within VAP, including a diagnostic question for each (left column), examples of characteristic failure modes (centre), and examples of candidate solutions (right).

## Implications For the Value Alignment Research Community

Distinguishing the many different subproblems of VAP illuminates why it has been such a great challenge to solve. The taxonomy provides greater conceptual rigor and theoretical clarity to the field, as well as other important implications and benefits for the alignment research community. In this section, we highlight four benefits.

### Towards a Science of Alignment

AI capabilities are growing faster than alignment research can keep up. While technical alignment research is flourishing and increasingly taking center stage at major AI conferences and among national AI safety institutes, it remains, in our view, theoretically immature and conceptually fragile. For alignment research to become alignment science it needs to show some of the characteristics of a mature science, including standardized terminology, a theoretical paradigm, explanatory power, predictive power, clear limits and domains of applicability (Godfrey-Smith 2021; Popper 1963; Kuhn 1962). The taxonomy provided in this paper is a step towards a science of alignment. It offers a theoretical paradigm with more nuanced terminology than what currently exists. The taxonomy makes it possible for any proposed solution to be explicitly directed at the precise subproblem it aims to address, thereby avoiding charges of overgeneralization (a hallmark of immature sciences; Kuhn 1962). Narrowing the target of proposed techniques will also increase our ability to predict when solutions will work and to explain why they work (or fail) when they do. Fields mature at different rates, and a taxonomy alone will not turn alignment research into a science, but it would get us closer by allowing for more precision and accuracy in how we address the complicated landscape of VAP.

### Not All Subproblems Have Technical Solutions

Gabriel (2020) provided an early and important distinction between normative and technical aspects of VAP, while also noting how the two are interrelated. He argues that given how ML systems work, it is likely that we cannot simply isolate and solve the technical aspects of AI alignment independently of the social and normative ones.<sup>12</sup> By introducing a taxonomic structure to VAP we can begin to identify more clearly exactly what aspects of each subproblem are normative vs. technical in nature. For example, the fact that human values are, and have always been, inconsistent, imperfect, culturally variable, and even contextually determined is arguably more of a philosophical, social and political problem, than a technical one. While there are

ways in which AI systems can help improve how moral we are, how fair we are, how we treat one another, and so on, it remains unlikely that any amount of technical ingenuity or clever programming will change the fact that we are morally fallible creatures. This doesn't imply, however, that the subproblem of human moral imperfection can be set aside, or ignored, by AI researchers. Our moral imperfections are reflected in our behavior, so if an AI system were to use methods like IRL or RLHF to infer the set of moral values to align with, the problem of moral imperfection could be exacerbated if the behavior being observed by a reward modelling agent is morally "suboptimal".

Another example of a VAP subproblem that is not entirely amenable to technical solutions is the problem of value anthropocentrism. Whether we should adopt either an anthropocentric or non-anthropocentric notion of value is a theoretical or normative issue (in the aforementioned sense) as opposed to a technical one. But, again, this does not mean that AI researchers should altogether disregard this challenge, as there may be important technical challenges associated with designing an AI system to be able to adequately deal with certain human vs. non-human value trade-offs. Hence, while the right notion of value cannot be settled technically, technical progress could be made towards resolving possible tensions, depending on which notion of value is ultimately adopted. The same could be said for the subproblem of how to prioritize values. The issue of how we prioritize values when they come into conflict, at least according to some moral philosophers, may have no single 'right answer'—it will depend on the nuances of a situational context, the agents involved, their relational histories and obligations to one another, and so forth (Statman 1996; Sinnott-Armstrong 1988). But no single 'right answer' doesn't imply no wrong answers. Sometimes preventing what are clearly morally bad choices would be a significant indicator of progress, even if there isn't a clear optimal course of action. Hence, technical efforts can still be worthwhile. The same could be said for the progress that has been made in resolving 'shallow' ethical tensions, as we discussed in the previous section.

The important lesson here is not that VAP is unsolvable (though it may be) but rather that calling system misalignment a failure sets the bar too high. We'll never live perfectly moral lives as individuals but we shouldn't therefore give up on trying to better ourselves. By analogy, the goal of alignment should be to build AI systems that are as aligned as possible, even if they are never perfectly aligned. This leads to our next implication.

---

<sup>12</sup> Lazar and Nelson (2023) make a similar point, but about the AI safety subfield more broadly.

## A Process View of Value Alignment

It is more productive to think of value alignment as a process than as a single solvable problem.<sup>13</sup> As we design advanced AI systems, we should go through a responsible alignment process to ensure that the resulting system is as aligned as possible, given the current science. A responsible alignment process should mirror the steps of other responsible AI design frameworks, but with a particular focus on alignment as the desired outcome. This will mean involving stakeholder consultations (e.g., whose values need to be considered?), safety evaluations and audits (e.g., how can we ensure the AI system behaves as intended in new circumstances?), well-being impact analysis (e.g., how is the AI system altering the values and preferences of stakeholders?), and risk assessments (e.g., what are the risks if the AI system reward-hacks?), to name a few stages (Peters et al. 2020).

### Identifying Taxonomic Structure and Blind Spots

Another benefit of having a clear taxonomy of VAP subproblems is the ability to identify blind spots in current alignment research. For example, our initial literature reviews suggest that there is far more attention being paid to finding solutions to certain subproblems (e.g., identifying human values, imperfect specification, prioritizing values, and provability) than others (such as, value anthropocentrism, inverse alignment and moral imperfection and progress). The field would benefit from future work that more systematically searches through existing alignment research to identify where attention has been primarily focused and where our field's 'blind spots' are, with the goal of redirecting attention to subproblems that have been overlooked. Such a review would have to carefully consider past alignment research to assess which subproblem(s) is really being addressed. By systematically taxonomizing alignment research to date, we may identify other important patterns. For example, it may be that certain subproblems have been the focus of attention for good reasons—perhaps (hypothetically) the most catastrophic safety risks emerge from only certain subproblems and not others, or perhaps certain subproblems, while distinct, are also contingent in the sense that the tractability of subproblem  $x$  presupposes some resolution to subproblem  $y$  (which is not to say that solving  $y$  would solve  $x$ ). For example, it's plausible that we'll first need to identify whose values we want an AI system to align with, before we can solve the subproblem of how to prioritize these values when they come into conflict.

## Conclusion

In this paper we have argued that VAP is a set of many different subproblems, and that explicitly identifying these subproblems will advance progress towards solutions. We began the paper by decomposing VAP into seven distinct subproblems. The problems are distinct subproblems in the sense that solving VAP would require solving all these subproblems and, furthermore, that a solution to any one subproblem would not solve the others. The taxonomy we provide is not meant to be exhaustive; there may be additional distinct subproblems that need to be addressed to solve VAP. We further argued for the value of this taxonomy. It is important not just for the sake of providing theoretical clarity, but also because of how this taxonomy can advance progress in the growing and important subfield of AI safety research. We make this point by showing how the current lack of nuance within the literature on VAP has led to unhelpful ambiguity in the solutions being developed.

Finally, we ended the paper by describing four benefits of the taxonomy for the value alignment research field. First, clearly distinguishing between distinct subproblems makes it possible for any proposed solution to be explicitly directed at the subproblem it aims to address. Going forward, this can help avoid confusions, ambiguity of scope, and even speed progress. This brings alignment research closer to a science, allowing researchers to predict and explain when and why solutions work or fail. Second, the taxonomy makes clear that different subproblems will require different kinds of solutions: some might be resolved through technical efforts, but others are fundamentally social or political in nature. Third, delineating the vast scope and varying nature of the subproblems that compose VAP substantiates why the 'problem' is better understood as a process—one that is ongoing and requiring intentional deliberation, interdisciplinarity, stakeholder consultations, and impact evaluations. Fourthly, introducing a taxonomy of subproblems is a first step towards identifying the taxonomic structure of VAP, that is, how these subproblems relate to one another, and whether some are more fundamental than others. It can also be used retroactively to bring structure to the field, enabling us to identify which subproblems have received the most attention and which have been overlooked. This, in turn, can help us identify blind spots and direct future efforts in the field to subproblems that warrant more attention and that are amenable to progress through technical solutions.

## Acknowledgments

This publication was made possible through the support of Grant 63578 from the John Templeton Foundation. The

---

<sup>13</sup> See Korb (1998) for a somewhat similar analysis of the longstanding AI frame problem. Korb argues that as originally presented, the frame problem is an unsolvable 'fairy tale' of a problem. We don't necessary endorse Korb's view about the frame

problem, but his analysis offers interesting parallels to our present analysis of the AI value alignment problem.

opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

## References

- Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep Learning with Differential Privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS'16). New York: Association for Computing Machinery. doi.org/10.1145/2976749.2978318
- Anwar, U.; Saparov, A.; Rando, J.; Paleka, D.; Turpin, M.; Hase, P.; Lubana, E. S.; Jenner, E.; Casper, S.; Sourbut, O.; Edelman, B., L.; Zhang, Z.; Günther, M.; Korinek, A.; Hernandez-Orallo, J.; Hammond, L.; Bigelow, E.; Pan, A.; Langosco, L.; Korbak, T.; Zhang, H.; Zhong, R.; Ó hÉigearthaigh, S.; Recchia, G.; Corsi, G.; Chan, A.; Anderljung, M.; Edwards, L.; Petrov, A.; Schroeder de Witt, C.; Motwani, S. R.; Bengio, Y.; Chen, D.; Torr, H. S., P.; Albanie, S.; Maharaj, T.; Foerster, J.; Tramèr, F.; He, H.; Kasirzadeh, A.; Choi, Y.; and Krueger, D. 2024. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. arXiv preprint. arXiv:2404.09932 [cs.LG]. Ithaca, NY: Cornell University Library.
- Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; El Showk, S.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv preprint. arXiv:2212.08073 [cs.CL]. Ithaca, NY: Cornell University Library.
- Bengio, Y.; Clare, S.; Prunkl, C.; Murray, M.; Andriushchenko, M.; Bucknall, B.; Bommasani, R.; Casper, S.; Davidson, T.; Douglas, R.; Duvenaud, D.; Fox, P.; Gohar, U.; Hadshar, R.; Ho, A.; Hu, T.; Jones, C.; Kapoor, S.; Kasirzadeh, A.; Manning, S.; Maslej, N.; Mavroudis, V.; McGlynn, C.; Moulange, R.; Newman, J.; Ng, K. Y.; Paskov, P.; Rismeni, S.; Sastry, G.; Seger, E.; Singer, S.; Stix, C.; Velasco, L.; Wheeler, N.; Acemoglu, D.; Conitzer, V.; Dietterich, T. G.; Felten, E. W.; Heintz, F.; Hinton, G.; Jennings, N.; Leavy, S.; Luderemir, T.; Marda, V.; Margetts, H.; McDermid, J.; Munga, J.; Narayanan, A.; Nelson, A.; Neppel, C.; Ramchurn, S. D.; Russell, S.; Schaake, M.; Schölkopf, B.; Soto, A.; Tiedrich, L.; Varoquaux, G.; Yao, A.; Zhang, Y-Q.; Aguirre, L. A.; Ajala, O.; Albalawi, F.; AlMalek, N.; Busch, C.; Collas, J.; de Carvalho, A. C. P. de L. F.; Gill, A.; Hatip, A., H.; Heikkilä, J.; Johnson, C.; Jolly, G.; Katzir, Z.; Kerema, M., N.; Kitano, H.; Krüger, A.; Lee, K. M.; López Portillo, J. R.; McLysaght, A.; Molchanovskiy, O.; Monti, A.; Nemer, M.; Oliver, N.; Pezoa, R.; Plonk, A.; Ravindran, B.; Riza, H.; Rugege, C.; Sheikh, H.; Wong, D.; Zeng, Y.; Zhu, L.; Privitera, D.; and Mindermann, S. 2026. International AI Safety Report 2026, Technical Report DSIT 2026/001. Department for Science, Innovation and Technology.
- Bondarenko, A.; Volk, D.; Volkov, D.; and Ladish, J. 2025. Demonstrating Specifcation Gaming in Reasoning Models. arXiv preprint. arXiv:2502.13295 [cs.AI]. Ithaca, NY: Cornell University Library.
- Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford, U.K.: Oxford University Press.
- Brandt, R. B. 1954. *Hopi Ethics: A Theoretical Analysis*. Chicago, IL: The University of Chicago Press.
- Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, P.; Zeitoff, T.; Filar, B.; Anderson, H.; Roff, H.; Allen, G.; Steinhardt, J.; Flynn, C.; Ó hÉigearthaigh, S.; Beard, S.; Belfield, H.; Farquhar, S.; Belfield, H.; Farquhar S.; Lyle, C.; Crotoft, R.; Evans, O.; Page, M.; Bryson, J.; Yampolskiy, R.; and Amodei, D. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *Apollo – University of Cambridge Repository*. doi.org/10.17863/CAM.22520
- Buckner, C. 2024. *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us About the Future of Artificial Intelligence*. New York, NY: Oxford University Press.
- Carranza, A.; Pai, D.; Schaeffer, R.; Tandon, A., and Koyejo, S. 2023. Deceptive Alignment Monitoring. arXiv preprint. arXiv:2307.10569 [cs.LG]. Ithaca, NY: Cornell University Library.
- Carroll, M.; Foote, D.; Siththaranjan, A.; Russell, S.; and Dragan, A. 2024. AI Alignment with Changing and Influenceable Reward Functions. In Proceedings of the 41st International Conference on Machine Learning.
- Cave, S.; Nyrupe, R.; Vold, K.; and Weller, A. 2019. Motivations and Risks of Machine Ethics. *Proceedings of the IEEE*, 107(3): 562-574. doi.org/10.1109/JPROC.2018.2865996
- Chakraborty, S.; Qiu, J.; Yuan, H.; Koppel, A.; Manocha, D.; Huang, F.; Bedi, A., S.; and Wang, M. 2024. MaxMin-RLHF: Alignment with Diverse Human Preferences. In Proceedings of the 41st International Conference on Machine Learning.
- Chagnon, N. 1996. *Yanomamö: The Fierce People*. 5th ed. New York, NY: Wadsworth Publishing.
- Chen, D.; Chen, Y.; Rege, A.; and Vinayak, R. K. 2024. PAL: Pluralistic Alignment Framework for Learning from Heterogeneous Preferences. arXiv preprint. arXiv:2406.08469 [cs.LG]. Ithaca, NY: Cornell University Library.
- Christian, B. 2020. *The Alignment Problem: Machine Learning and Human Values*. New York, NY: WW Norton and Company.
- Christiano, P. F.; Leike, J.; Brown, T., B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. In Proceedings of the 31st International Conference of Neural Information Processing Systems. Redhook, NY: Curran Associates, Inc.

- Clark, J., and Amodei, D. 2016. Faulty Reward Functions in the Wild. <https://openai.com/index/faulty-reward-functions/>. Accessed: 2026-08-05.
- Cohen, D., and Nisbett, R., E. 1996. *Culture of Honor: The Psychology of Violence in the South*. Boulder, CO: Westview Press.
- Conitzer, V.; Freedman, R.; Heitzig, J.; Holliday, W. H.; Jacobs, B. M.; Lambert, N.; Mosse, M.; Pacuit, E.; Russell, S.; Schoelkopf, H.; Tewolde, E.; and Zwicker, W. S. 2024. Position: Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. In Proceedings of the 41st International Conference on Machine Learning.
- Cummings, R.; Desfontaines, D.; Evans, D.; Geambasu, R.; Huang, Y.; Jagielski, M.; Kairouz, P.; Kamath, G.; Oh, S.; Ohrimenko, O.; Papernot, N.; Rogers, R.; Shen, M.; Song, S.; Su, W.; Terzis, A.; Thakurta, A.; Vassilvitskii, S.; Wang, Y.; Xiong, L.; Yu, D.; Yekhanin, S.; Zhang, H.; and Zhang, W. 2024. Advancing Differential Privacy: Where We are Now and Future Directions for Real-World Deployment. *Harvard Data Science Review* 6(1). doi.org/10.1162/99608f92.d3197524
- Dalrymple, D.; Skalse, J.; Bengio, Y.; Russell, S.; Tegmark, M.; Seshia, S.; Omohundro, S.; Szegedy, C.; Goldhaber, B.; Ammann, N.; Abate, A.; Halpern, J.; Barrett, C.; Zhao, D.; Zhi-Xuan, T.; Wing, J.; and Tenenbaum, J. 2024. Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. arXiv preprint. arXiv:2405.06624 [cs.AI]. Ithaca, NY: Cornell University Library.
- Denison, C.; MacDiarmid, M.; Barez, F.; Duvenaud, D.; Kravec, S.; Marks, S.; Schiefer, N.; Soklaski, R.; Tamkin, A.; Kaplan, J.; Shlegeris, B.; Bowman, S. R.; Perez, E.; and Hubinger, E. 2024. Sycophancy to Subterfuge: Investigating Reward-Tampering in Language Models. arXiv preprint. arXiv:2406.10162 [cs.AI]. Ithaca, NY: Cornell University Library.
- Doris, J. M., and Stich, S., P. 2007. As a Matter of Fact: Empirical Perspectives on Ethics. In *The Oxford Handbook of Contemporary Philosophy*, edited by F. Jackson and M. Smith, 114-152. Oxford, U.K.: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199234769.003.0005>
- Dung, L. 2023. Current Cases of AI Misalignment and their Implications for Future Risks. *Synthese*, 202(138). doi.org/10.1007/s11229-023-04367-0
- Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; Grosse, R.; McCandlish, S.; Kaplan, J.; Amodei, D.; Wattenberg, M.; and Olah, C. 2022. Toy Models of Superposition. arXiv preprint. arXiv:2209.10652 [cs.LG]. Ithaca, NY: Cornell University Library.
- Feng, S.; Sorensen, T.; Liu, Y.; Fisher, J.; Park, C. Y.; Choi, Y.; and Tsvetkov, Y. 2024. Modular Pluralism Alignment via Multi-LLM Collaboration. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida: Association for Computational Linguistics.
- <http://doi.org/10.18653/v1/2024.emnlp-main.240>
- Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; Schafer, B.; Valcke, P.; and Vayena, E. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28: 689-707. doi.org/10.1007/s11023-018-9482-5
- Floridi, L. 2019. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology* 32: 185-193. doi.org/10.1007/s13347-019-00354-x
- Floridi, L., and Cowls, J. 2022. A Unified Framework of Five Principles for AI in Society. In *Machine Learning and the City: Applications in Architecture and Urban Design*, edited by S. Carta, 535-545. Hoboken, NJ: John Wiley & Sons. doi.org/10.1002/9781119815075.ch45
- Future of Life Institute. 2017. Asilomar AI Principles. <https://futureoflife.org/open-letter/ai-principles/>. Accessed: 2026-08-05.
- Gabriel, I., 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30: 411-437. doi.org/10.1007/s11023-020-09539-2
- Gabriel, I., and Ghazavi, V. 2021. The Challenge of Value Alignment: From Fairer Algorithms to AI Safety. In *The Oxford Handbook of Digital Ethics*, edited by C. Véliz, 336-355. Oxford, U.K.: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780198857815.013.18>
- Gabriel, I.; Manzini, A.; Keeling, G.; Hendricks, L.A.; Rieser V.; Iqbal, H.; Tomašev, N.; Ktena, I.; Kenton, Z.; Rodriguez, M.; El-Sayed, S.; Brown, S.; Akbulut, C.; Trask, A.; Hughes, E.; Bergman, A., S.; Shelby, R.; Marchal N.; Griffin, C.; Mateos-Garcia, J.; Weidinger, L.; Street, W.; Lange, B.; Ingerman, A.; Lentz, A.; Enger, R.; Barakat, A.; Krakovna, V.; Siy, J., O.; Kurth-Nelson, Z.; McCroskery, A.; Bolina, V.; Law, H.; Shanahan, M.; Alberts, L.; Balle, B.; de Haas, S.; Ibitoye, Y.; Dafoe, A.; Goldberg, B.; Krier, S.; Reese, A.; Witherspoon, S.; Hawkins, W.; Rauh, M.; Wallace, D.; Franklin, M.; Goldstein, J., A.; Lehman, J.; Klenk, M.; Vallor, S.; Biles, C.; Morris, M., R.; King, H.; Arcas, B., A., Y.; Isaac, W.; and Manyika J. 2024. The Ethics of Advanced AI Assistants. arXiv preprint. arXiv:2404.16244 [cs.CY]. Ithaca, NY: Cornell University Library.
- Gabriel, I., and Keeling, G. 2025. A Matter of Principle? AI Alignment as the Fair Treatment of Claims. *Philosophical Studies* 182: 1951-1973. doi.org/10.1007/s11098-025-02300-4
- Gao, L.; Schulman, J.; and Hilton, J. 2023. Scaling Laws for Reward Model Overoptimization. In Proceedings of the 40th International Conference on Machine Learning.
- Godfrey-Smith, P. 2021. *Theory and Reality: An Introduction to the Philosophy of Science*. 2nd ed. Chicago, IL: University of Chicago Press.
- Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral Foundations Theory: The Prag-

- matic Validity of Moral Pluralism. *Advances in Experimental Social Psychology* 47: 55-130. <https://doi.org/10.1016/B978-0-12-407236-7.00002-4>
- 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. 2024. Alignment Faking in Large Language Models. arXiv preprint. arXiv:2412.14093 [cs.AI]. Ithaca, NY: Cornell University Library.
- Hendrycks, D. 2024. *Introduction to AI Safety, Ethics and Society*. Boca Raton, FL: CRC Press. <https://doi.org/10.1201/9781003530336>
- Hubinger, E.; van Merwijk, C.; Mikulik, V.; Skalse, J.; and Garra-brant, S. 2019. Risks From Learned Optimization in Advanced Machine Learning Systems. arXiv preprint. arXiv:1906.01820 [cs.AI]. Ithaca, NY: Cornell University Library.
- Humphreys, R. 2014. Biocentrism. In *Encyclopedia of Global Bioethics*, edited by ten Have, H., 1-11. Cham, Switzerland: Springer. doi.org/10.1007/978-3-319-05544-2\_41-1
- Ibarz, B.; Leike, J.; Pohlen, T.; Irving, G.; Legg, S.; and Amodei, D. 2018. Reward Learning From Human Preferences and Demonstrations in Atari. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Redhook, NY: Curran Associates Inc.
- Jobin, A.; Ienca, M.; and Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1: 389-399. doi.org/10.1038/s42256-019-0088-2
- Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Vierling, L.; Hong, D.; Zhou, J.; Zhang, Z.; Zeng, F.; Dai, J.; Pan, X.; Ng, K.; Y.; O’Gara, A.; Xu, H.; Tse, B.; Fu, J.; McAleer, S.; Yang, Y.; Wang, Y.; Zhu, S.; Guo, Y.; Gao, W. 2025. AI Alignment: A Contemporary Survey. arXiv preprint. arXiv:2310.19852 [cs.AI]. Ithaca, NY: Cornell University Library.
- Ji, J.; Chen, W.; Wang, K.; Hong, D.; Fang, S.; Chen, B.; Zhou, J.; Dai, J.; Han, S.; Guo, Y.; and Yang, Y. 2025. Mitigating Deceptive Alignment via Self-Monitoring. arXiv preprint. arXiv:2505.18807 [cs.AI]. Ithaca, NY: Cornell University Library.
- Kasirzadeh, A. 2024. Plurality of Value Pluralism and AI Value Alignment. In *Proceedings of the Pluralistic Alignment Workshop at NeurIPS 2024*. OpenReview.
- Klassen, T. Q.; Alamdari, P. A.; and McIlraith, S. A. 2024. Pluralistic Alignment Over Time. In *Proceedings of the Pluralistic Alignment Workshop at NeurIPS 2024*. OpenReview.
- Kenton, Z.; Everitt, T.; Weidinger, L.; Gabriel, I.; Mikulik, V.; and Irving, G. 2021. Alignment of Language Agents. arXiv preprint. arXiv:2103.14659 [cs.AI]. Ithaca, NY: Cornell University Library.
- Kirk, H. R.; Gabriel, I.; Summerfield, C.; Vidgen, B.; and Hale, S. A. 2025. Why Human-AI Relationships Need Socioaffective Alignment. *Humanities and Social Science Communications* 12(728). doi.org/10.1057/s41599-025-04532-5
- Korb, K. B. 1998. The Frame Problem: An AI Fairy Tale. *Minds and Machines* 8: 317-351. doi.org/10.1023/A:1008286921835.
- Krakovna, V.; Uesato, J.; Mikulik, V.; Rahtz, M.; Everitt, T.; Kumar, R.; Kenton, Z.; Leike, Z.; Legg, S. 2020. Specification Gaming: The Flipside of AI Ingenuity. <https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/> Accessed: 2026-08-05.
- Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.
- LaCroix, T. 2025. *Artificial Intelligence and the Value Alignment Problem: A Philosophical Introduction*. Peterborough, ON: Broadview Press.
- Lazar, S., and Nelson, A. 2023. AI Safety on Whose Terms? *Science*, 381(6654): 138-138. doi.org/10.1126/science.adi8982
- May, J., and Holton, R. 2012. What in the World is Weakness of Will?. *Philosophy Studies* 157: 341-360. doi.org/10.1007/s11098-010-9651-8
- Malmqvist, L. 2026. Winning at All Cost: A Small Environment for Eliciting Specification Gaming Behaviors in Large Language Models. In *Applied Cryptography and Network Security Workshops, ACNS 2025. Lecture Notes in Computer Science, vol 15655*, edited by M. Manulis, 52-69. Cham, Switzerland: Springer. doi.org/10.1007/978-3-032-01823-6\_4
- Morley, J.; Floridi, L.; Kinsey, L.; and Elhalal, A. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics* 26: 2141-2168. doi.org/10.1007/s11948-019-00165-5
- Ng, A.Y.; and Russell, S. J. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc.
- Ngo, R.; Chan, L.; and Mindermann, S. 2024. The Alignment Problem from a Deep Learning Perspective. In *Proceedings of the Twelfth Intentional Conference on Learning Representations*. OpenReview.
- Olson, T.; Salas-Damian, R.; and Forbus, K. D. 2024. A Defeasible Deontic Calculus for Resolving Norm Conflicts. arXiv preprint. arXiv:2407.04869 [cs.AI]. Ithaca, NY: Cornell University Library.
- OpenAI. 2025. Sycophancy in GPT-4o: What Happened and What We’re Doing About it. <https://openai.com/index/sycophancy-in-gpt-4o/> Accessed: 2026-08-05.
- Ouadrhiri, A. E.; and Abdelhadi, A. 2022. Differential Privacy for Deep and Federated Learning: A Survey. *IEEE Access*, 10: 22359-22380. doi.org/10.1109/ACCESS.2022.3151670

- 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.; Lowe, R. 2022. Training Language Models to Follow Instructions with Human Feedback. In Proceedings of the 36th International Conference on Neural Information Processing Systems. Redhook, NY: Curran Associates, Inc.
- Peters, D.; Vold, K.; Robinson, D.; and Calvo, R. A. 2020. Responsible AI—Two frameworks For Ethical Design Practice. *IEEE Transactions on Technology and Society*, 1(1): 34–47. doi.org/10.1109/TTS.2020.2974991
- Peterson, M. 2019. The Value Alignment Problem: A Geometric Approach. *Ethics and Information Technology*, 21: 19-28. doi.org/10.1007/s10676-018-9486-0
- Pettit, P. 2003. Akrasia, Collective and Individual. In *Weakness of the Will and Practical Irrationality*, edited by S. Stroud and C. Tappolet, 68-96. London, U.K.: Oxford University Press. doi.org/10.1093/0199257361.003.0004
- Popper, K. 1963. *Conjectures and refutations: The growth of scientific knowledge*. London, England: Routledge.
- Popov, I.; Heess, N.; Lillicrap, T.; Hafner, R.; Barth-Maron, G.; Vecerik, M.; Lampe, T.; Tassa, Y.; Erez, T.; and Riedmiller, M. 2017. Data-efficient Deep Reinforcement Learning for Dexterous Manipulation. arXiv preprint. arXiv:1704.03073 [cs.LG]. Ithaca, NY: Cornell University Library.
- Raper, R. 2025. A Comment on the Pursuit to Align AI: We Do Not Need Value-Aligned AI, We Need AI That is Risk-Averse. *AI and Society* 40: 229-231. doi.org/10.1007/s00146-023-01850-7
- Regan, T. 1983. *The Case for Animal Rights*. Berkley, CA: University of California Press.
- Russell, S. 1998. Learning Agents for Uncertain Environments. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT' 98). New York, NY: Association for Computing Machinery.
- Russell, S.; and P. Norvig. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. New Jersey: Prentice Hall.
- Russell, S. 2019. *Human Compatible: AI and The Problem of Control*. New York, NY: Viking.
- Santos, J. S.; Zahn, J. O.; Silvestre, E. A.; Silva, V. T.; and Vasconcelos, W., W. 2017. Detection and Resolution of Normative Conflicts in Multi-agent Systems: A Literature Survey. *Autonomous Agents and Multi-Agent Systems* 31: 1236–1282. doi.org/10.1007/s10458-017-9362-z
- Sinnott-Armstrong, W. 1988. *Moral Dilemmas*, Oxford, U.K.: Basil Blackwell.
- Skalse, J.; Howe, N., H., R.; Krashenninnikov, D.; and Krueger D. 2022. Defining and Characterizing Reward Hacking. In Proceedings of the 36th Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc.
- Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Miresghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; Althoff, T.; and Choi, Y. 2024. Position: A Roadmap to Pluralistic Alignment. In Proceedings of the 41st International Conference on Machine Learning.
- Statman, D. 1996. Hard Cases and Moral Dilemmas. *Law and Philosophy* 15: 117-148. doi.org/10.1007/BF00144130
- Stole, I. L. 2008. Philanthropy as Public Relations: A Critical Perspective on Cause Marketing. *International Journal of Communications* 2: 20-40.
- Stroud, S., and Svirsky, L. 2025. Weakness of Will. In *The Stanford Encyclopedia of Philosophy* (Fall 2025 Edition), edited by E. N. Zalta and U. Nodelman. <https://plato.stanford.edu/entries/weakness-will/>. Accessed: 2026-08-06.
- Sutton, R. S., and Barto, A., G. 1998. *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
- Tessler, M. H.; Bakker, M.A.; Jarrett, D.; Sheahan, H.; Chadwick, M. J.; Koster, R.; Evans, G.; Campbell-Gillingham, L.; Collins, T.; Parkes, D. C.; Botvinick, M.; and Summerfield C. 2024. AI Can Help Humans Find Common Ground in Democratic Deliberation. *Science* 386 (6719). doi.org/10.1126/science.adq2852
- Turing, A. M. 1951. Intelligent Machinery, A Heretical Theory. In *The Essential Turing*, edited by B. J. Copeland, 465-475. Oxford, U.K.: Oxford University Press, 2004. <https://doi.org/10.1093/oso/9780198250791.003.0018>
- Vamplew, P.; Hayes, C. F.; Foale, C.; Dazeley, R.; and Harland, H. 2024. Multi-objective Reinforcement Learning: A Tool for Pluralistic Alignment. In Proceedings of the Pluralistic Alignment Workshop at NeurIPS 2024. OpenReview.
- Vold, K. and Harris, D. R. 2021. How Does Artificial Intelligence Pose an Existential Risk? In *The Oxford Handbook of Digital Ethics*, edited by C. Véliz, 724-747. Oxford, U.K.: Oxford University Press. doi.org/10.1093/oxfordhb/9780198857815.013.36
- Wiener, N., 1960. Some Moral and Technical Consequences of Automation. *Science* 131(3410): 1355-1358. doi.org/10.1126/science.131.3410.1355
- Whittlestone, J.; Nyrup, R.; Alexandrova, K.; Cave, S. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). New York, NY: Association for Computing Machinery. doi.org/10.1145/3306618.3314289
- Whittlestone, J.; Arulkumaran, A.; and Crosby, M. 2021. The

Societal Implications of Reinforcement Learning. *Journal of Artificial Intelligence Research* 70: 1003-1030.  
doi.org/10.1613/jair.1.12360

Zhi-Xuan, T.; Carroll, M.; Franklin, M. and Ashton, H. 2024. Beyond Preferences in AI Alignment. *Philosophy Studies* 182: 1813–1863. doi.org/10.1007/s11098-024-02249-w

Zhou, W., and Li, W. 2024. Rethinking Inverse Reinforcement Learning: From Data Alignment to Task Alignment. arXiv preprint. arXiv:2410.23680. Ithaca, NY: Cornell University Library.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning Language Models from Human Preferences. arXiv preprint. arXiv:1909.08593 [cs.CL]. Ithaca, NY: Cornell University Library.