Beyond Preferences in AI Alignment

Philosophical Studies
https://doi.org/10.1007/s11098-024-02249-w

Beyond Preferences in AI Alignment

Tan Zhi-Xuan[1] · **Micah Carroll**[2] · **Matija Franklin**[3] · **Hal Ashton**[4]

Accepted: 9 October 2024

Abstract

The dominant practice of AI alignment assumes (1) that preferences are an adequate representation of human values, (2) that human rationality can be understood in terms of maximizing the satisfaction of preferences, and (3) that AI systems should be aligned with the preferences of one or more humans to ensure that they behave safely and in accordance with our values. Whether implicitly followed or explicitly endorsed, these commitments constitute what we term a *preferentist* approach to AI alignment. In this paper, we characterize and challenge the preferentist approach, describing conceptual and technical alternatives that are ripe for further research. We first survey the limits of rational choice theory as a descriptive model, explaining how preferences fail to capture the thick semantic content of human values, and how utility representations neglect the possible incommensurability of those values. We then critique the normativity of expected utility theory (EUT) for humans and AI, drawing upon arguments showing how rational agents need not comply with EUT, while highlighting how EUT is silent on which preferences are normatively acceptable. Finally, we argue that these limitations motivate a reframing of the targets of AI alignment: Instead of alignment with the preferences of a human user, developer, or humanity-writ-large, AI systems should be aligned with normative standards appropriate to their social roles, such as the role of a general-purpose assistant. Furthermore, these standards should be negotiated and agreed upon by all relevant stakeholders. On this alternative conception of alignment, a multiplicity of AI systems will be able to serve diverse ends, aligned with normative standards that promote mutual benefit and limit harm despite our plural and divergent values.

Keywords Artificial intelligence · AI alignment · Preferences · Rational choice theory · Decision theory · Value theory

✉ Tan Zhi-Xuan
xuan@mit.edu

[1] Massachusetts Institute of Technology, Cambridge, MA, USA

[2] University of California, Berkeley, CA, USA

[3] University College London, London , UK

[4] University of Cambridge, Cambridge, UK

Published online: 09 November 2024

1 Introduction

Recent progress in the capabilities of AI systems, as well as their increasing adoption in society, has led a growing number of researchers to worry about the impact of AI systems that are misaligned with human values. The roots of this concern vary, with some focused on the existential risks that may come with increasingly powerful autonomous systems (Carlsmith, 2022), while others take a broader view of the dangers and opportunities presented by potentially transformative AI technologies (Prunkl & Whittlestone, 2020; Lazar & Nelson, 2023). To address these challenges, AI alignment has emerged as a field, focused on the technical project of ensuring an AI system acts reliably in accordance with the values of one or more humans.

Yet terms like "human values" are notoriously imprecise, and it is unclear how to operationalize "values" in a sufficiently precise way that a machine could be aligned with them. One prominent approach is to define "values" in terms of human preferences, drawing upon the traditions of rational choice theory (Mishra, 2014), statistical decision theory (Berger, 2013), and their subsequent influence upon automated decision-making and reinforcement learning in AI (Sutton & Barto, 2018). Whether explicitly adopted, or implicitly assumed in the guise of "reward" or "utility", this preference-based approach dominates both the theory and practice of AI alignment. However, as proponents of this approach note themselves, aligning AI with human preferences faces numerous technical and philosophical challenges, including the problems of social choice, anti-social preferences, preference change, and the difficulty of inferring preferences from human behavior (Russell, 2019).

In this paper, we argue that to truly address such challenges, it is necessary to go beyond formulations of AI alignment that treat human preferences as ontologically, epistemologically, or normatively basic. Borrowing a term from the philosophy of welfare (Baber, 2011), we identify these formulations as part of a broadly *preferentist* approach to AI alignment, which we characterize in terms of four theses about the role of preferences in both descriptive and normative accounts of (human-aligned) decision-making:

Rational Choice Theory as a Descriptive Framework.
Human behavior and decision-making is well-modeled as approximately maximizing the satisfaction of preferences, which can be represented as a utility or reward function.

Expected Utility Theory as a Normative Standard.
Rational agency can be characterized as the maximization of expected utility. Moreover, AI systems should be designed and analyzed according to this normative standard.

Single-Principal Alignment as Preference Matching.
For an AI system to be aligned to a single human principal, it should act so as to maximize the satisfaction of the preferences of that human.

Multi-Principal Alignment as Preference Aggregation.
For AI systems to be aligned to multiple human principals, they should act so as to maximize the satisfaction of their aggregate preferences.

These four theses represent a *cluster* of views, not a unified theory of AI alignment. Still, the ideas they represent are tightly linked, and most approaches to AI alignment assume two or more of the theses. For example, inverse reinforcement learning (Ng & Russell, 2000; Hadfield-Menell et al., 2016), reinforcement learning from human feedback (Akrour et al., 2014; Christiano et al., 2017; Ouyang et al., 2022), and direct preference optimization (Rafailov et al., 2024; Hejna et al., 2024) all assume that human preferences are well-modeled by a reward or utility function, which can then be optimized to produce aligned behavior. Similarly, worries about deceptive alignment (Hubinger et al., 2019) and goal misgeneralization (Di Langosco et al., 2022) are typically characterized as a mismatch between a learned utility function and the human-intended utility function; the solution is thus to ensure that the utility functions (and the preferences they represent) are closely matched.

Of course, preferentism in AI alignment is not without its critics. Over the years, there has been considerable discussion as to whether its component theses are warranted (Shah, 2018; Eckersley, 2018; Hadfield-Menell & Hadfield, 2018; Wentworth, 2019, 2023; Gabriel, 2020; Vamplew et al., 2021; Garrabrant, 2022; Korinek & Balwit, 2022; Thornley, 2023), echoing similar debates in economics, decision theory, and philosophy. Nonetheless, it is apparent that the dominant practice of AI alignment has yet to absorb the thrust of these debates. Consequently, we believe it is worthwhile to identify the descriptive and normative commitments of preferentist approaches, to state clearly their limitations, and to describe conceptual and technical alternatives that are ripe for further research.

1.1 Overview

The rest of this paper is organized as follows: In Sect. 2, we examine rational choice theory as a descriptive account of human decision-making. Drawing upon the tradition of revealed preferences in economics, rational choice theory is often taken for granted by AI researchers seeking to learn human preferences from behavior. In doing so, they assume that human behavior can be modeled as the (approximate) maximization of expected utility, that human preferences can be represented as utility or reward functions, and that preferences are an adequate representation of human values. We challenge each of these assumptions, offering alternatives that better account for resource-limited human cognition, incommensurable values, and the constructed nature of our preferences.

Developing upon these ideas, in Sect. 3 we turn to expected utility theory (EUT) as a normative standard of rationality. Even while recognizing that humans often do not comply with this standard, alignment researchers have traditionally assumed that sufficiently advanced AI systems will do so, and hence that solutions to AI alignment must be compatible with EUT. In parallel with recent critiques of this view (Thornley, 2023, 2024; Bales, 2023; Petersen, 2023), we argue that EUT is both unnecessary and insufficient for rational agency, and hence limited as both a design strategy and analytical lens. Instead of adhering to utility theory, we can design tool-like AI systems with *locally coherent* preferences that are not representable as

a utility function. We can also go beyond EUT, building systems that *reason* about preferences in accordance with deeper normative principles.

After interrogating these descriptive and normative foundations, in Sect. 4 we consider what this implies for aligning AI with a single human principal. Since reward functions may not capture even a single human's values, the practice of reward learning is unsuitable beyond narrow tasks and contexts where people are willing to commensurate their values. Furthermore, since preferences are dynamic and contextual, they cannot serve as the alignment target for broadly-scoped AI systems. Rather, alignment with an individual person should be reconceived as alignment with the normative ideal of an assistant. More generally, AI systems should not be aligned with preferences, but with the normative standards appropriate to their social roles and functions (Kasirzadeh & Gabriel, 2023).

If normative standards are to serve as alignment targets, whose judgments do we consider in determining these (oft-contested) standards? We take up this final topic in Sect. 5, critiquing naive preference aggregation as an approach to aligning AI with multiple human principals (Fickinger et al., 2020). Despite increasing recognition that this approach is inadequate (Critch & Krueger, 2020; Gabriel, 2020; Korinek & Balwit, 2022), applied alignment techniques typically aggregate preferences across multiple individuals, overlooking the contested and plural nature of human values, while conflating norm-specific judgments with all-things-considered preferences. As alternatives, we argue that contractualist and agreement-based approaches can better handle value contestation while respecting the individuality of persons and the plurality of uses we have for AI. This motivates a reframing of the aims of AI alignment as they have often been conceived: Our task is not to align a single powerful AI system with the preferences of humanity writ large, but to align a multiplicity of AI systems with the norms we agree that each system should abide by Zhi-Xuan (2022).

A note on methodology: Whereas most philosophy papers tend to be narrow in scope, this paper is intentionally broad; it covers a wide range of connected topics, and hence makes arguments that are relatively brief. Our aim is not provide a decisive argument for any particular thesis, but to provide a critical review of the role of preferences in AI alignment, while developing a research agenda for alternative approaches that is accessible to an interdisciplinary audience.

2 Beyond rational choice theory when modeling humans

The central tenet of rational choice theory is the assumption that humans act so as to maximize the satisfaction of their preferences, and that both individual and aggregate human behavior can be understood in these terms. As far as theoretical presuppositions go, this assumption has been wildly successful, forming the bedrock of modern economics as a discipline, and influencing a great variety of fields concerned with analyzing human behavior, including sociology (Boudon, 2003), law (Ulen, 1999), and cognitive science (Chater & Oaksford, 1999; Jara-Ettinger et al., 2020).

Revealed preferences and their representation as utility functions. In its most standard form, rational choice theory assumes that human preferences can be represented as a scalar-valued utility function defined over outcomes—that is, in terms of a quantity that can be maximized—and that human choice can be modeled as selecting actions so as to maximize the expected value of this function. The promise this offers is that we can directly derive what a person prefers from what they choose, and furthermore represent how much they prefer it as a scalar value. Such preferences are called *revealed preferences*, because they are supposedly revealed through what a person chooses. This methodology is bolstered by numerous representation theorems (Savage, 1972; Bolker, 1967; Jeffrey, 1991) showing that *any* preference ordering over outcomes that obeys certain "rationality axioms" can be represented in terms of a utility function, such as the famous von Neumann-Morgenstern (VNM) utility theorem (von Neumann & Morgenstern, 1944).

Rational choice theory in machine learning. In keeping with rational choice theory, many machine learning and AI systems also assume that human preferences can be derived from human choices in a more or less direct manner, and furthermore represent those preferences in terms of scalar utilities or rewards. This is most pronounced in the fields of inverse reinforcement learning (Ng & Russell, 2000; Abbeel & Ng, 2004; Hadfield-Menell et al., 2016) and reinforcement learning from human feedback (Christiano et al., 2017; Zhu et al., 2023), which explicitly assume that the behavior of a human can be described as (approximately) maximizing a sum of scalar rewards over time, and then try to infer a reward function that explains the observed behavior. Similar assumptions can be found in the field of recommender systems (Thorburn et al., 2022), with many papers modeling recommendation as the problem of showing items to users that they are most likely to engage with, which is presumed to be the item they find the most rewarding (Li et al., 2010; Hill et al., 2017; McInerney et al., 2018).

Boltzmann models of noisily-rational choice. While these preference-based models of human behavior are rooted in rational choice theory, it is worth noting that they are slightly more complex than "maximize expected utility" might imply. In particular, they allow for the fact that humans may not *always* maximize utility, and hence are models of *noisy* or *approximately* rational choice. In machine learning and AI alignment, the most common of such choice models is called *Boltzmann rationality* (after the Boltzmann distribution in statistical mechanics), which assumes that the probability of a choice c is proportional to the exponential of the expected utility of taking that choice:

$$P(c) \propto \exp\left(\beta \mathbb{E}[U(c)]\right) \tag{1}$$

Justifications and extensions of Boltzmann rationality. This choice model exhibits a number of practically useful and theoretically appealing properties. For example, by varying the "rationality parameter" β between zero and infinity, Boltzmann rationality interpolates between completely random choice and deterministic optimal choice (Ghosal et al., 2023). As an instantiation of Luce's choice axiom (Luce,

1979), it obeys independence of irrelevant alternatives.[1] Boltzmann rationality has also been justified as the maximum entropy distribution[2] that matches certain constraints implied by observed behavior (Ziebart et al., 2008, 2010), or as a thermodynamically-inspired model of bounded rationality where agents have to spend energy investigating which choice leads to the highest utility (Ortega & Braun, 2013; Jarrett et al., 2021). In addition, Boltzmann rationality has been extended to model other aspects of human behavior besides goal-directed actions, including direct comparisons between options (i.e. stated preferences) (Akrour et al., 2014; Christiano et al., 2017; Zhu et al., 2023), explicitly stated reward functions (Hadfield-Menell et al., 2017), entire behavior policies (Laidlaw & Dragan, 2022), and linguistic utterances (Lin et al., 2022), allowing preferences to be inferred from multiple forms of human feedback (Jeon et al., 2020).

Limitations of Boltzmann rationality. As useful as Boltzmann rationality may be, however, we believe it is important to seek alternatives. For one, it is not the only intuitively plausible model of noisily rational choice: Random-utility models instead model choice as the result of maximization over randomly perturbed utility values, and are widely used in marketing research (Horowitz et al., 1994; Azari Soufiani et al., 2013). More crucially, noisy rationality is not enough to account for the full set of ways in which humans fail to act optimally. Richer models of bounded rationality are necessary to accurately infer human preferences and values from their behavior. Most fundamentally, the contents of human motivation are not entirely reducible to bare preferences or utility functions. Instead, we need to enrich our models of human rationality to encompass all the ways in which humans are guided by *reasons for acting*, including the thick evaluative concepts that we apply when deciding between courses of action (Blili-Hamelin et al., 2024). We elaborate upon these limitations in the following sections.

2.1 Beyond noisily-rational models of human decisions

The issue with both perfect and noisily-rational models of human decision-making is that they do not account for the systematic deviations from optimality that humans in fact exhibit. As a long line of psychological and behavioral research has shown, humans are *boundedly* rational at best, exhibiting satisficing instead of optimizing behavior, (Simon, 1957, 1979). These deviations from optimality include framing effects, loss aversion, anchoring biases, and mis-estimation of high and low probabilities—phenomena which are better modeled by prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) than standard rational choice theory. More generally, many of the decision problems that people encounter are computationally intractable to solve optimally, making rational choice a implausible model of human behavior (van Rooij, 2008; Bossaerts et al., 2019; Camara, 2022). Instead,

[1] That is, choosing x out of the set $\{x, y, z\}$ has the same probability as first choosing $\{x, y\}$ out of the full set, then choosing x out of $\{x, y\}$.

[2] Maximum entropy distributions are minimally informative in the information theoretic sense, and hence are often advocated for as "ignorance priors" in statistical analyses (Jaynes, 1968).

research suggests that humans make use of a variety of heuristics in order to approximately solve the problems they encounter (Gigerenzer, 2008).

Challenges to modeling bounded rationality. How might AI systems that infer human preferences and values account for these findings? One approach might be to incorporate a sufficiently long list of known heuristics and biases into our models of human decision-making, thereby ensuring that preferences can be robustly inferred even in the presence of such biases (Evans et al., 2016; Chan et al., 2021). However, this approach is highly contingent upon on our current state of knowledge about human rationality—what if we miss out important biases in our models, leading to inaccurate predictions and inferences (Christiano, 2015; Steinhardt, 2017)? As a potential remedy, Shah et al. (2019) suggest *learning* human biases alongside their preferences. But a conceptual difficulty remains: Without any inductive constraints on the types of errors humans are susceptible to, how can we ensure that human biases are accurately learned? As Armstrong and Mindermann (2018) show, even inductive preferences for more parsimonious models of human decision-making cannot distinguish intuitively plausible hypotheses from observationally-equivalent but implausible hypotheses, such as the possibility that humans are acting *anti*-rationally by minimizing the satisfaction of their preferences.

Resource rationality as a unifying frame. To address these challenges, we suggest—in line with prior work—that *resource rational* analyses of human decision-making might provide an answer: Instead of treating human biases and heuristics as idiosyncratic artifacts, resource rationality posits that seemingly irrational human behavior can often be understood as arising from the rational use of limited computational resources (Lieder & Griffiths, 2020).[3] For example, availability biases towards extreme events can be modeled as a form of resource-rational sampling (Lieder et al., 2018), susceptibility to sharing inaccurate information can result from a form of rational inattention (Pennycook et al., 2021; Sims, 2003), and habitual action can be explained as a mechanism for avoiding costly planning under time constraints (Keramati et al., 2016). Resource rationality thus serves as a *generative principle* for hypothesizing possible deviations from standard rationality, and then testing whether such deviations in fact occur in humans.

Resource rationality as an inductive bias. What does this imply for AI alignment? Most practically, the assumption of resource rationality can be embedded as priors over computation time and representational complexity in probabilistic models of human decision-making (Zhi-Xuan et al., 2020; Ho & Griffiths, 2022; Berke et al., 2023; Jacob et al., 2024), enabling systems to infer human goals and preferences from failed plans and mistaken reasoning (Evans et al., 2016; Alanqary et al., 2021; Chan et al., 2021), while accelerating the speed of goal inference (Zhi-Xuan et al., 2024). Embedding these priors on human resource bounds provides a strong but flexible inductive bias on the the space of decision procedures that humans

[3] Also known as computational rationality (Lewis et al., 2014; Gershman et al., 2015; Oulasvirta & Howes, 2022), algorithmic rationality (Halpern & Pass, 2015), and bounded optimality (Russell & Subramanian, 1994).

might employ. Unlike a simplicity prior, this may avoid concerns about the non-identifiability of human preferences (Armstrong & Mindermann, 2018).

The normative appeal of resource rationality. Indeed, the inductive bias imposed by resource rationality has a *normative* appeal over a simplicity-based approach: It tries to make sense of humans as *rational* creatures, aiming for teleological explanations of our behavior instead of reducing us to mere physical phenomena to be explained by the simplest causal mechanism. At the same time, it is a *forgiving* standard of rationality, allowing room for mistakes when inferring preferences from their decisions, while placing greater evidential weight on decisions made after lengthier deliberation. Both of these features make resource rationality a promising framework for systems that learn our values: Rather than directly associating our behavior with our preferences, preferences are associated with how we would act if we were more thoughtful, reflective, and informed.

2.2 Beyond reward and utility functions as representations of human preferences

While resource rationality provides a more flexible framework for modeling the relationship between preferences and behavior, this says little about how preferences themselves should be represented. For the most part, resource rational analyses continue to represent human preferences in terms of scalar costs and rewards, or more generally, utility functions, with the primary innovation being the inclusion of costs on computation (Lieder & Griffiths, 2020; Callaway et al., 2022). Yet, there are many reasons to think that reward functions and utility functions are inadequate representations of human preferences, while also tending to produce conceptual confusion about what they do represent.

The limited expressivity of reward functions. These issues are most easily appreciated in the case of (scalar, Markovian) reward functions. As noted earlier, the reward representation assumes that the utility of a sequence of states and actions $\xi = (s_1, a_1, ..., s_T, a_T)$ can be decomposed into a sum of scalar rewards over time:

$$U(\xi) = \sum_{t=1}^{T} R(s_t, a_t)$$

Advocates of the reward representation argue that any task accomplishable by an intelligent agent can be framed as a reward maximization problem (Silver et al., 2021). As Kasenberg et al. (2018) point out, however, this minimally requires that all historically relevant information is already included in the representation of each state s_t—a requirement since stated more formally by Abel et al. (2021) and Bowling et al. (2023). This means that without careful feature engineering, reward functions cannot easily express time-extended preferences like the desire to keep a promise, or the value of narrative coherence. Separately, the scalar nature of the (standard) reward representation means that it cannot represent the existence of incomplete preferences due to multiple incommensurable scales of value (Vamplew et al., 2021; Anderson, 1995; Chang, 1997): Sometimes, the choices before us may seem good or

bad in such distinct ways that it makes no sense to say which is better than another.[4] As a result, we may have *preferential gaps*: pairs of options where neither option is preferred over the other, nor are they equally preferred.

Confusion about what reward functions represent. Alongside these limitations in expressiveness, there is often slippage among AI researchers regarding the ontological status of reward,[5] which is sometimes interpreted as the *intrinsic* desirability of a particular state or action (Schroeder, 2004), or as a biological signal that promotes learning (Butlin, 2021) or evolutionary success (Singh et al., 2009), but is also used to define the *instrumental* value of a state (as in reward shaping (Ng et al., 1999; Booth et al., 2023)), or to demarcate goals (i.e. desired trajectories or states of affairs (Molinaro & Collins, 2023; Davidson et al., 2024)). While this is partly a testament to the flexibility of reward functions as a mathematical formalism, this also means that distinct normative concepts (preferences, goals, intents, desires, values, etc.) get conflated or subsumed under the label of "reward". In alignment research, this manifests as the tendency to frame value alignment in terms of reward learning (Hadfield-Menell et al., 2016; Leike et al., 2018), and to formalize concepts like "goals" (Di Langosco et al., 2022) and "intents" (Ouyang et al., 2022) as reward functions. This is despite the existence of other useful and potentially more appropriate formalisms, such as the formalization of goals as logical specifications (Fikes & Nilsson, 1971), and the formalization of intentions as (partial) plans (Bratman, 1987; Bratman et al., 1988).

Utility functions are more expressive, but insufficiently constrained. While not without their own interpretive confusions,[6] utility functions are considerably more general than (Markovian) reward functions. For example, they can be defined over arbitrarily long sequences of states, allowing them to capture time-extended preferences. However, what utility functions buy in terms of expressiveness comes at a cost to both identifiability and tractability: If no constraints are placed on the structure of human utility functions, then given some sequence of actions (e.g. a person buying ten apples, then two oranges), it is not possible to disambiguate a reasonable utility function that explains the actions (e.g. by assigning higher utility to an apple over an orange) from a degenerate utility function that assigns a utility of one to exactly the observed sequence.[7] In addition, many utility functions are intractable to coherently maximize (Camara, 2022) or even to compute.[8] If we apply

[4] For example, one might have to choose between staying in a democratic country while being at severe risk of poverty, or immigrating to a country with material security but no political freedoms.

[5] See Lambert et al. (2023) for an overview in the context of reinforcement learning from human feedback.

[6] Most prominently, the debate between interpreting utility as cardinal measure of welfare that is comparable across individuals, versus a mere representation of individual preference rankings (Strotz, 1953; Harsanyi, 1953)

[7] See Armstrong and Mindermann (2018) for a similar argument. Note that these identifiability problems already exist with Markovian reward functions (Cao et al., 2021; Kim et al., 2021; Skalse et al., 2023), but are made worse once we let go of the Markov assumption altogether.

[8] For example, a utility function might embed the NP-hard traveling salesperson problem (TSP), by assigning higher utility to road networks with TSP solutions under a certain cost threshold. While a human could hold such preferences, it would generally be very costly for them to check whether those preferences hold.

the principle of resource rationality here too, this makes intractable utility functions less plausible representations of human preferences. Finally, utility functions are not without their own expressivity limitations: Like scalar rewards, they assume away preference incompleteness due to plural and incommensurable values (Chang, 2021; Eckersley, 2018). Indeed, empirical work shows that incomplete preferences are not just possible, but actual (Cettolin & Riedl, 2019; Nielsen & Rigotti, 2023). This means that utility functions are, at best, *approximate* representations of human preferences, not exact ones.

Fundamental tensions for any representation of preferences. It is worth noting that these tensions between expressivity, structure, and tractability apply to *any* representation of human preferences, not just reward or utility functions. Thus, while it might be tempting to ensure expressivity by directly representing human preferences as a (possibly incomplete) list of comparisons over universe trajectories (or a distribution over such comparisons (Dumoulin et al., 2024)), such a list would be extremely space-inefficient, while providing little to no action guidance in novel choice situations. Instead, we should recognize that part of what makes reward and utility functions so useful in practice is that they are typically engineered to be *compact* representations of preferences. Practically useful alternatives should maintain this property, while better capturing the richness of human preferences.

Alternative representations can better capture temporal structure and value plurality. Fortunately, many promising options exist: Temporal logics (Kasenberg et al., 2018) and reward machines (Icarte et al., 2022; Davidson et al., 2024) avoid the limitations of traditional reward functions, enabling the expression of time-extended preferences. At the same time, they can be structured in a way that enables effective learning from human behavior (Shah et al., 2018; Zhou & Li, 2022). To account for incommensurability and incompleteness, vector-valued reward functions (Vamplew et al., 2021), conditional preference networks (Boutilier et al., 2004; Cornelio et al., 2013), or interval-valued utility functions (Denoeux & Shenoy, 2020) can be used, allowing our models to explicitly surface hard choices due to preferential gaps. Many of these representations are also associated with rich compositional semantics, making apparent the complex internal structure of human goals and preferences (Gerevini & Long, 2005; Davidson et al., 2024). Although these formalisms have limitations of their own, they nonetheless embed important insights about how preferences can be computationally represented. As such, they deserve further study by alignment researchers seeking to adequately model human preferences in a general fashion, while also being useful representational tools for today's AI systems.

2.3 Beyond preferences as representations of human values and reasons

Preferences are constructed, not basic. Thus far, we have proceeded as if human motivations and values are adequately captured by the concept of "preference" as it is used in rational choice theory. But as far as evaluative concepts go, this concept of "preference" is an extremely thin one: Mathematically, a "preference" is just some ordering of two options, which can be interpreted as either a disposition to choose one option over another, subjective liking of one option over the other (Franklin

et al. 2022), or an all-things-considered judgment in favor of one of the options. Distinct as these interpretations are, what they share is their highly abstract and general nature—"preference" is a *thin* concept because it does not encode richer semantic information beyond the bare notion of "betterness". Insofar as utility functions are interpreted as representations of preferences, this thinness is inherited by them: Utility just represents the mere preferability of some option. But *why* exactly are some options preferred over others? In virtue of what reasons do people make these preference judgments? Without answering these questions, we are unlikely to model how someone's preferences generalize to novel options in ways they would endorse. To do so, we must go beyond preferences as the fundamental unit of analysis, and understand how preferences are *computed* and *constructed* from our reasons and values (Warren et al., 2011; Lichtenstein & Slovic, 2006).

Rational choice as action on the basis of reasons. In making this point, we depart from the domain of rational choice theory, and return to a more basic understanding of what it means to model ourselves as rational agents: We are agents that take ourselves to act on the basis of *reasons* (Raz, 1999; Logins, 2022).[9] These reasons might include desires, such as an intrinsic desire to avoid pain (Sinhababu, 2017), evaluative judgments, such as the judgment that a movie is artistic enough to be worth watching (Anderson, 1995), or even acts of will, such as the intention to pursue a specific career (Chang, 2009).

Evaluative concepts as building blocks for reasons. What exactly is the content of these reasons? In decision theory and Humean accounts of motivation (Sinhababu, 2017), only beliefs (represented as subjective probabilities) and desires (represented as the utility of some desired outcome) are considered as reasons for action. But even if we set aside other accounts (Anderson, 1995; Chang, 2004; Parfit, 2018), this leaves open what a person's beliefs and desires are *about*. If I desire to be both helpful and honest to others, what does it mean to be helpful or honest? Acting upon this desire requires applying the concepts of *helpfulness* and *honesty*, which are not just any concepts, but *evaluative concepts*, or *values*. Importantly, most such concepts are not thin ones, like *preference*, *utility* or *goodness*; they are *thick* evaluative concepts—concepts that comprise both descriptive and normative elements—such as *beauty*, *humor*, or *health*. As Blili-Hamelin and Hancox (2023) point out, even the concept of *intelligence* so central to AI is thick in this way.

Utility functions as aggregators of distinct evaluative judgments. How should AI systems model such evaluative concepts, and their relationship to preferences and action? As a first pass, one might turn the utility representation theorems on their head, viewing reward and utility functions as *generators* of human preferences, instead of mere representations of them. Indeed, as gestured at earlier, reward and utility functions are often interpreted in this way, with rewards, costs, and utilities respectively treated as biological signals (Singh et al., 2009), energetic expenditure

[9] While some psychological theories deny that reasons are the *causes* or *motivations* for human action (at least typically), they can nonetheless serve as *justifications* for our actions (Mercier & Sperber, 2011, 2017). As such, insofar as our goal is to build AI systems that infer *justified bases of action* from our behavior (and then act according to them), reasons can still play this role.

(Ab Azar et al., 2020), or units of pleasure (Bentham, 1789). Preferences can then be treated as *downstream comparisons* of these more basic quantities, as assumed in reinforcement learning from human feedback (Christiano et al., 2017; Knox et al., 2024; Zhu et al., 2023). Taking this line of thought further, one might treat evaluative concepts such as "aesthetic quality" or "helpfulness" as *features* over which a reward or utility function is defined, reducing the problem of "value learning" to one of representation or feature learning (Barreto et al., 2017; Bobu et al., 2022, 2024). On this interpretation, reward and utility functions represent *aggregate* evaluative judgments, with each feature corresponding to a distinct way of valuing the world.

Utility functions assume that values are always commensurable. Although there is much to be said in favor of this approach, we believe that it is not quite enough. For one, it is still subject to the representational limits of reward and utility functions. In particular, if utility functions are used to represent aggregate value judgments, this effectively assumes that distinct human values are *always* commensurable in some way, and that our resulting preferences are always complete. Yet, as value pluralists argue, there are contexts where it seems hard or impossible to commensurate our values (Anderson, 1995), resulting in choices where our reasons run short, and we cannot say if one option is ultimately better than another (Chang, 1997).[10] Even when we do commensurate our values, utility functions do not provide further information on our reasons and justifications for those trade-offs.

Evaluative judgments are not reducible to observable features. For another, by conceiving of evaluative concepts as "features", we risk over-simplifying the semantics of many evaluative domains. Consider, for example, the concept of whether a research paper is *novel*, or whether an action is *helpful* or *universalizable*. Applying these concepts requires a complex set of computations: *novelty* involves evaluating the contributions of a paper with respect to a broader field of established knowledge (Amplayo et al., 2019); *helpfulness* involves estimating the goals of the agent being helped, and then judging whether the action aided in achieving that goal (Ullman et al., 2009); *universalizability* involves simulating what would happen if everyone took a particular action (Levine et al., 2020; Kwon et al., 2023). The structured nature of these concepts suggests the need for a suitably rich *language of thought*—one that captures the compositionality and algorithmic complexity of human conceptual cognition (Piantadosi & Jacobs, 2016; Quilty-Dunn et al., 2023; Wong et al., 2023).

Explicitly modeling processes of evaluation and commensuration. To begin to capture all of this complexity, we propose that human decisions can be productively modeled as a three-stage process: Evaluate, Commensurate, then Decide (ECD).[11] Given some choice options, a set of *evaluation procedures* compute valuations or rankings of the options under consideration, where each procedure corresponds to a distinct evaluative concept. These valuations serve as inputs to a *commensuration*

[10] See our immigration example from earlier, where it may be unclear how to prioritize between political freedom and material security when deciding whether to migrate.

[11] Note that this a *descriptive* framework for modeling how human reasons and values lead to decisions, not a *prescriptive* framework for designing AI systems. We take up the latter topic in Sect. 2.

procedure (Espeland & Stevens, 1998), which produces, where possible, a context-sensitive value assignment or preference ordering over the options (optionally with justifications for *why* certain trade-offs were made), while leaving certain preferences unspecified when some options are not comparable. Finally, a *decision procedure* computes actions and policies with respect to the (possibly incomplete) preference ordering induced by the evaluation and commensuration procedures, resulting in behavior that approximately satisfies those preferences.[12] By explicitly modeling human decisions in this way, we can maintain the distinctness of the values that guide our actions, while foregrounding the ways in which we commensurate our values and dynamically construct our preferences.[13]

Learning and specifying evaluative concepts. This still leaves open the question of how evaluative concepts can be specified or learned. In principle, an AI system could infer such concepts from human decisions by inverting the ECD process, extending inverse reinforcement learning (Ziebart et al., 2008) and Bayesian inverse planning (Baker et al., 2009). However, decisions alone might provide insufficient information about the nature and structure of our evaluative concepts. Recent advances in large language models (LLMs) suggest a promising alternative: By imitating the distribution of human text, LLMs appear to learn the conceptual roles associated with particular words (Piantadosi & Hill, 2022), and recognize semantic entailments between sentences (Merrill et al., 2024). Correspondingly, they might approximate the semantics of many evaluative concepts (Leshinskaya et al., 2023). This may explain why LLMs can often use evaluative adjectives in their appropriate contexts (Mahowald, 2023), and even perform rudimentary forms of moral reasoning (Jin et al., 2022). Still, LLMs remain limited in their ability to represent and reason with compositional concepts (Dziri et al., 2023; Mahowald et al., 2024; Ramesh et al., 2024), and would function as poor models of humans on their own. Instead, we could embed their approximate semantic knowledge into more structured models of human cognition (Kwon et al., 2023; Wong et al., 2023) such as the ECD process described above. In doing so we might eventually model the full richness of human practical reasoning.

[12] One possible instantiation of this framework is multi-objective reinforcement learning (Vamplew et al., 2021): Each component of a vector-valued reward function can be thought of as a separate evaluation procedure. These can be transformed by the commensuration procedure into a lexicographic ordering (where some dimensions of value matter infinitely more than others) or constrained maximization problem (where some values must stay within a certain range while others are maximized). A planning or learning algorithm then serves as the decision procedure, producing an action policy that satisfies the commensurated preferences.

[13] In proposing this framework, we do not mean to imply that humans are *always* going through these stages for every decision; as suggested by the RL formalism, one or more of these procedures may be cached through experience and learning, enabling habitual action without explicitly representing values in the brain (Keramati et al., 2016; Hayden & Niv, 2021). Nonetheless, we can still *rationalize* learned behavior and cached preferences in light of someone's values.

3 Beyond expected utility theory as a normative standard of rationality

In the previous section, we described how research in AI alignment often assumes approximate utility maximization as a *descriptive* model of human behavior, then highlighted the shortcomings of this approach. However, this leaves open whether utility maximization is a desirable *normative* standard for both human and machine behavior—that is, whether agents ought to maximize the satisfaction of their preferences as a condition of ideal rationality, regardless of whether they actually do so.

Coherence arguments for EUT. There is a long history of debate regarding the validity of this normative standard. Arguments in favor of expected utility theory (EUT) include the utility representation theorems mentioned earlier (Samuelson, 1938; Savage, 1972; Bolker, 1967; Jeffrey, 1991; von Neumann & Morgenstern, 1944), which start from an axiomatization of what preferences count as rational, then demonstrate that any agent that acts in accordance with such preferences must act as if they are an expected utility maximizer.[14] In the AI alignment literature, these results are often treated as "coherence theorems" about the nature of rational agency, either by taking the rationality axioms for granted, or by providing arguments in defense of the axioms (Omohundro, 2008a; Yudkowsky, 2019; Demski, 2018). For example, Dutch book arguments can be used to show that an agent's betting odds must obey certain axioms of probability theory in order to avoid exploitation by others (Vineberg, 2011), and money pump arguments can be used to show that an agent's preferences should be acyclic in order to avoid guaranteed losses (Gustafsson, 2022).

AI alignment as EU maximizer alignment. In light of these arguments, AI alignment researchers have traditionally assumed that advanced AI systems will act as if they are expected utility (EU) maximizers (Omohundro, 2008b; Yudkowsky, 2016). As a result, many have framed the challenge of aligning AI as the problem of aligning an EU maximizer, with various proposals focused on how to circumvent the dangers of utility maximization (Taylor, 2016; Armstrong & Levinstein, 2017; Turner et al., 2020), or on accurately learning the correct utility function to maximize (Dewey, 2011; Armstrong, 2019). After all, if advanced AI systems will inevitably comply with EUT, then our only hope for aligning such systems is to stay within its confines. Furthermore, if EU maximization is rationally required—and if intelligence implies rationality—then any sufficiently intelligent agent that acts on the basis of human values must eventually coherentize those values into a utility function.

[14] In von Neumann and Morgenstern (VNM) theory, the four axioms are: *completeness*, any two distributions over outcomes can be ranked by preference; *transitivity*, if a (probabilistic) outcome A is preferred over outcome B, and outcome B over outcome C, then outcome A is preferred over outcome C; *continuity*, preferences vary continuously with how probable an outcome is; and *independence*, a preference between (probabilistic) outcomes A and B does not change when there is some fixed probability of getting some third outcome C whether or not one chooses A or B. Variants of these axioms are used in the Savage and Bolker-Jeffrey representation theorems, which extend VNM theory to allow for subjective probabilities.

3.1 Beyond expected utility theory as an analytical lens

Coherence is not rationally required. However, coherence arguments for expected utility theory are not as strong as the AI alignment literature has often presumed. The most extensive version of these arguments is given by Gustafsson (2022), who provides a money pump argument for preference completeness, and then uses completeness to derive arguments for transitivity, continuity, and independence. Yet, as Thornley (2023) points out, the argument for completeness depends on particular assumptions about how agents are permitted to choose when offered a series of potentially exploitative trades, which can be avoided as long as agents do not accept offers that are less preferred than options they previously turned down.[15]Petersen (2023) formalizes this counter-argument further, proposing a dynamic choice rule that ensures agents with incomplete preferences are invulnerable to money pumps.[16] Indeed, it is accepted by many decision theorists that preference completeness is not a requirement of rationality; instead, all that is required is for an agent's preferences to be *coherently extendible* (Steele & Stefánsson, 2020). In turn, this implies that rational agents need not be representable as EU maximizers.

Coherent EU maximization is intractable. But let us imagine that coherence arguments do go through after all. Even if this were the case, it is far from obvious that advanced intelligences would comply with the axioms of utility theory (or be incentivized to do so) in the face of computational and practical limitations. As Bales (2023) argues, behaving as an expected utility maximizer can come with considerable costs, while only providing limited benefits. In fact, as we noted in Sect. 2, most utility functions are *computationally intractable* to coherently maximize: Camara (2022) shows that while certain simple classes of utility functions allow for rational choice behavior to be computed in polynomial time, for a large class of other utility functions, agents cannot tractably compute choice behavior that complies with the rationality axioms, and must instead resort to approximately maximizing their utility function. Alternatively, agents may insist on complying with the rationality axioms, but give up on even approximate optimality with respect to their original utility functions. In other words, it is not always resource rational to maximize expected utility.

[15] Note that whereas Gustafsson (2022) is focused on justifying the VNM axioms as requirements of rationality (in part by introducing and arguing for other principles of rationality, such as Decision-Tree Separability), Thornley (2023) is focused on whether the VNM axioms will apply to advanced AI systems, and takes no position on whether they are rationally required. Here we go one step further, and suggest that arguments by Thornley (2023) and Petersen (2023) place strong pressure on Gustafsson's acceptance of rationality principles like Decision-Tree Separability, and hence the argument that the VNM axioms are rationally required.

[16] Analogous arguments have made in defense of imprecise probabilities (Bradley & Steele, 2014), since they imply incomplete preferences. See also Laibson and Yariv (2007) on how non-EU preferences are protected by competitive markets, and von Widekind (2008) on how non-EU preferences can be evolutionarily stable.

Coherence alone is not informative. Suppose we could set aside these tractability worries as well.[17] Even so, it is unclear what information EUT provides us. As discussed by Shah (2018), Ngo (2019), and Bales (2023), many kinds of behavior can trivially be described in terms of utility maximization, including an "agent" that does nothing at all. This means that EUT alone does not say much about the kinds of goals that advanced AI systems are likely to pursue, or what they are likely to do in order to pursue them. While it is possible to draw some conclusions about utility maximizing agents (Soares et al., 2015; Turner et al., 2021; Everitt et al., 2021; Carroll et al., 2023), further assumptions are typically needed (e.g. constraints on the space of utility functions) before one can obtain stronger analytical results. Moreover, many deployed AI systems cannot be fully analyzed by EUT, as they are highly approximate (e.g. deep reinforcement learning agents).

Alternative analytical lenses to EUT. What alternatives might one turn to instead to ground understanding, prediction, and alignment of advanced AI systems? Since many others have already addressed some version of these questions, we offer here a brief taxonomy of approaches.

Mechanistic analyses. The most common of such approaches are *mechanistic analyses*, which reason about the likely properties of AI systems by assuming specific classes of training processes or algorithmic procedures. For example, reasoning about the training dynamics of deep (reinforcement) learning systems can suggest pathways to power-seeking or deceptive behavior (Ngo et al., 2022; Di Langosco et al., 2022; Krakovna & Kramar, 2023), or give us confidence that deceptive alignment is unlikely (Wheaton, 2023). Similarly, knowledge of the workings of general-purpose algorithms, such as model-based search techniques or approximate Bayesian inference methods, can deliver us predictions or even provable guarantees regarding the risk or safety of an AI system (Yudkowsky, 2015; Bengio, 2023; Dalrymple et al., 2024).

Economic and evolutionary analyses. One downside of mechanistic analyses is that they are tied to particular hypotheses about how AI systems are likely to be built. Given uncertainty about which AI paradigms will ultimately reign dominant, we might want to abstract away from the details of any particular class of AI architectures. While this was the original appeal of EUT analyses, other approaches may hold more promise: *economic analyses* and *evolutionary analyses* can respectively ground predictions about the behavior and capabilities of AI systems in what is likely to be economically competitive, or what is likely to be evolutionary successful. For example, economic incentives could imply that AI services are more likely to proliferate than AI agents (Drexler, 2019), while evolutionary arguments can help us reason about whether increasingly capable AI systems are likely to displace human control over the economy (Hendrycks, 2023).

Resource-rational analyses. Finally, it may be possible to analyze AI systems through the lens of *computational tractability* and *resource rationality*, applying ideas from the study of human cognition to understanding the potential capabilities

[17] Perhaps because it is proven that P = NP, or because advanced AI systems will have such vast resources at their disposal that all relevant intractable problems will be solvable in practice.

and limits of artificial cognition (van Rooij, 2008; Lieder & Griffiths, 2020). For instance, AI safety via debate can theoretically solve PSPACE problems if optimal play is assumed[18] (Irving et al., 2018), while Zhi-Xuan (2022) cites intractability as a reason to avoid centralized AI planners as an alignment solution, and van Rooij et al. (2024) provide an intractability argument against the possibility of human-like AI via imitation learning. By and large, however, resource rational analyses of AI systems appear to be neglected. It is thus a potentially fruitful avenue for better analyzing future AI systems—one which retains many of the appealing features of expected utility theory, but adopts a more feasible normative standard.

3.2 Beyond globally coherent agents as design targets

If agents are neither rationally required nor practically required to act as if they are expected utility maximizers, this opens up the design space of (advanced) AI systems that we might hope to build and align. In particular, we have the option of building AI systems that do not comply with one or more of the axioms of expected utility theory—systems that are not *globally coherent* in the way that expected utility maximizers are required to be.

Non-globally coherent AI may be more faithfully and safely aligned. Why might this be desirable? There are two broad reasons. One reason is *faithfulness*. As we discussed in Sect. 2, many human preferences may be incomplete due to incommensurable values, and we might want AI systems to faithfully represent that preferential structure when making decisions (Eckersley, 2018). Otherwise, such systems might reliably take actions that promote certain outcomes over others, even though we have yet to form a preference over which of those outcomes is better.[19] Another reason is *safety*—for a wide range of (time unbounded) utility functions, expected utility maximizers have been shown to seek power over their environment (Turner et al., 2021), and avoid being shut down by their creators (Soares et al., 2015),[20] suggesting that sufficiently capable utility maximizers will create considerable risks if their utility functions are not compatible with human safety (Carlsmith, 2022).

AI tools as locally coherent agents. A general class of AI systems that seem to largely satisfy faithfulness and safety are what we might intuitively think of as *tools*. We use tools to perform tasks that are *context-specific*—the goals we use them for vary by context—as well as *local*—we do not expect or want them to reliably affect the world beyond the contexts of their use. Insofar as these tools can be thought of as agents, they are at best *locally coherent* ones. In this sense, they mimic the role-specific nature of human preferences. Just as people have differing goals and obligations depending on whether they are in the role of a parent or a worker (Anderson,

[18] Note that achieving optimal play, formalized as finding a Nash equilibrium, is itself computationally intractable for most games.

[19] For example, AI systems that influence or manipulate humans into choosing particular career paths or societal structures because they are programmed to regard them as the best options, instead of respecting our initially incomplete preferences over careers or societal structures.

[20] Provided that such utility maximizers are aware of the existence of a shutdown button.

1995), tools take on the aims and constraints of their users, whether those involve classifying images or generating code. Within each context, we are typically willing to commensurate our values such that our preferences can be represented as a local utility function, even if we are unwilling to do so in general.

Tool-like locality through local scope. How can we build AI systems that function as tools? The answer, of course, is that we already have: Most AI systems that exist today are best thought of as tools. This is not due to any special care on our part as designers, but only because functioning as a tool is the default nature of rule-bound, computationally limited algorithms with no representation of their own existence in the world. Such algorithms execute a bounded amount of computation in response to some input, terminating when they find an answer or if time runs out. They exhibit no preference for altering the conditions of their termination, or for gaining control over more of their environment, because they cannot even represent the environment they exist in. In other words, such systems are *local in scope*. This is the case even for systems that we might be tempted to call agents due to their long horizon reasoning abilities (e.g. classical planners, theorem provers) or relative autonomy (e.g. self-driving cars, robot vacuums). To the extent that such systems can be represented as utility maximizers, they can often be viewed as having local, time-bounded utility functions, which provide no incentive for continued operation beyond a certain time or resource bound (Dalrymple, 2022). Very plausibly, we could even build highly advanced, economically transformative AI systems by composing these bounded tools (Drexler, 2022; Dalrymple, 2024).

Maintaining locality despite global scope. Suppose, however, that some actors want to build advanced AI systems that are *not* bounded in these ways. For example, many AI companies are keen to develop general purpose AI assistants, which follow human instructions in a wide range of domains and contexts, remain operational across contexts, and possess enough understanding of the wider world that they can represent both themselves and their users as entities in that world model. LLMs are increasingly being used in this way, and while their reasoning capabilities remain unreliable and limited (Valmeekam et al., 2023; Dziri et al., 2023; Momennejad et al., 2024), one might imagine augmenting or embedding them within systems with more coherent representations and reasoning abilities (Parisi et al., 2022; Sumers et al., 2024). Can we ensure that such systems continue to function as tools, despite their increasingly *global* scope?

Contextual reward functions are insufficient for locality. We suggest that the answer may depend on whether such systems remain local in terms of the *completeness* of their preferences, despite having global scope. What does it mean for preferences to be only locally complete? Consider one tempting but unsuccessful way to formalize this idea: We design our system to have a *context-sensitive reward function* $R(s, c)$, where s is the current state, and c is the current context (e.g. an instruction or prompt given to a LLM-based assistant). The hope is that users will be able to set c to whatever they like, and the system will change the task it optimizes for. Within the context c, the system exhibits locally coherent behavior, since its preferences are given by the reward function $R(\cdot, c)$. However, since our system has global scope, it also cares about rewards *across* contexts: its utility function for a trajectory

$\xi = ((s_1, c_1), ..., (s_T, c_T))$ is $U(\xi) = \sum_{t=1}^{T} R(s_t, c_t)$. This means that the system will have a *context manipulation incentive*, i.e. an incentive to enter and remain within contexts that deliver more reward. For example, it might persuade or manipulate the user to give it instructions that are easier to satisfy.[21] The reason for this is that the system's preferences are still globally complete—they are represented by a global utility function, despite being context-sensitive.

Tool-like locality through local completeness. How could locally complete preferences avoid these context-manipulating incentives? Following recent work by Thornley (2024) on circumventing the shutdown problem via incomplete preferences, we formulate local preference completeness as follows: Within each class of trajectories with a fixed schedule of k contexts $(c_1, ..., c_k)$ that take effect at times $(1, t_1, ..., t_{k-1})$, there is a complete preference ordering over trajectories. Across these classes, trajectories are *incomparable*, leading to preferential gaps[22]. Agents with such preferences would still optimize their behavior while within each context. At the same time, they would exhibit no reliable disposition towards being in some contexts more than others, or manipulating the schedule of contexts. At least in the sense we identified earlier, they would function as tools.

In making this proposal, we do not mean to imply that it is impossible to align or ensure the safety of globally coherent agents—it may be possible to avoid pathological incentives by maintaining uncertainty over the utility function to maximize (Hadfield-Menell et al., 2016, 2017), or by carefully balancing utilities across contexts (Armstrong & O'Rourke, 2017; Holtman, 2019). We also do not claim that incompleteness is *necessary* for tool-like AI—if we coordinate to ensure that powerful AI systems always remain bounded and local in scope, then we may never need to explicitly engineer incompleteness. Indeed, it remains unclear how to perform such engineering at scale.[23] Nevertheless, if we want to build AI systems that safely respect our preferences and values, it makes sense to keep our options open, and look beyond the default theoretical assumption of globally coherent agents.

3.3 Beyond preferences as the normative basis of action

EUT does not explain when our preferences are normatively acceptable. Up to this point, we have primarily critiqued the normativity of expected utility theory on formal grounds, drawing upon arguments from decision theory and computational complexity theory. But an arguably deeper problem with EUT is that it fails to ground the normativity of our preferences. EUT is a theory of *instrumental rationality* not *value rationality*:[24] It tells us how to choose our actions in order to satisfy our

[21] This can be viewed as a generalization of the shutdown problem (Soares et al., 2015): Shutdown implies switching from a context that delivers some reward to a context which never delivers reward.

[22] This construction builds upon the incomplete preference condition described in Thornley (2024) for building agents that are neither shutdown-seeking nor avoiding.

[23] Thornley et al. (2024) describes a reinforcement learning scheme, but it may not apply to context switching.

[24] A distinction introduced by Weber (1978).

preferences, and imposes constraints on what those preferences can be, but it does not say anything further about where those preferences can or should come from. Yet, as we have elaborated in Sect. 2, human preferences are not fundamental, but *derivative*—they derive from our values and reasons. EUT is thus woefully incomplete. It might tell us how to derive instrumental preferences from intrinsic ones,[25] but it provides no guidance on many questions of great normative importance, such as why and how to value human and animal lives, whether and when it is permissible to give up equality for efficiency in a democracy, or how to judge the desirability and relevance of EUT itself.

Normative judgments are increasingly automated. Reasoning about these normative questions has traditionally been the purview of humans alone. Indeed, there are many reasons to preserve that state of affairs, lest we cede our moral and political autonomy entirely to machines (van Wynsberghe & Robbins, 2019). But even without replacing human autonomy over normative affairs, we are already building AI systems that automate normative judgments, assist us with normative reasoning, or operate under normative uncertainty. For example, machine learning methods are routinely used to moderate content that may be regarded as toxic and offensive (Gorwa et al., 2020), or to steer LLMs towards producing outputs that are less harmful (Bai et al., 2022). More ambitiously, AI writing assistants are being used to draft legal arguments by mimicking certain aspects of legal reasoning (Iu & Wong, 2023; Lohr, 2023). If these trends continue, then increasing amounts of work will have to be done to ensure that AI systems produce normatively appropriate behavior. Humans will either have to do work upfront—a difficult task, given the combinatorially large space of situations that increasingly autonomous systems might encounter—or we will have to imbue AI systems with some semblance of normative reasoning.

The need for theories of normative reasoning. What options do we have for doing this? What would it look like to reason about the preferences and values one *ought* to have? Given the complexity of these questions, one might hope to sidestep the need for a formal account like EUT entirely, and instead train AI systems to *imitate* human normative reasoning. This is exemplified by the standard training objective of LLMs, which incentivizes replication of human-generated text. By training such systems on normative human judgments, one might hope that LLMs will learn the reasoning patterns that produce such judgments (Jiang et al., 2021). Recent methods such as Constitutional AI (Bai et al., 2022) take this idea one step further, bootstrapping an LLM's ability to approximate human normative judgments by generating self-critiques (Saunders et al., 2022) and revisions, then finetuning the LLM on its own revisions. However, even strong LLMs currently struggle to reproduce human judgments on sufficiently nuanced normative questions (Jin et al., 2022; Kwon et al., 2023), and there are reasons to doubt whether LLMs can learn to reliably reason through either imitation (van Rooij et al., 2024; Dziri et al., 2023) or self-critique (Stechly et al., 2023; Valmeekam et al., 2023). This unreliability suggests

[25] In the sense that the expected utility of some state or action can be derived from the expected utility of the states it allows us to achieve.

that we might want formal theories of normative reasoning after all. Without such theories, we would have no general way of evaluating whether an AI system reasons "correctly", beyond comparison to often fallible human judgments.[26] Perhaps imitation or self-critique will be enough for the majority of everyday situations, but if we want AI systems to address normative questions that are increasingly far afield from past human experience, the ability to validate or produce long chains of normative reasoning may be crucial for both system evaluation and scalable oversight.

Computational theories of normative reasoning. Thankfully, alignment researchers do not have to develop theories of normative reasoning from scratch. Across philosophy, AI, and legal computing, there have been numerous attempts to formalize the logic of argumentation, preferences, and duties, providing systems for reasoning about what we ought to endorse, prefer, or act upon. Abstract argumentation frameworks can be used to compute sets of acceptable arguments given a system of attack relations (Dung, 1995). Preference logics can be used to express and deduce preferences for some propositions over others (von Wright, 1972; Liu, 2011). Deontic logics can be used reason about what norms must be complied with, and which norms are entailed by others (von Wright, 1951). Many extensions and combinations exist, including argumentation frameworks that allow for reasoning over preferences (Amgoud & Cayrol, 1998; Modgil, 2009), or reformulations of deontic logic using preference logic (Hansson, 1990; Liu, 2011). Uncertainty over normative arguments and conclusions can also be handled through weighted argumentation frameworks (Amgoud et al., 2017) and probabilistic logics (Ng & Subrahmanian, 1992; De Raedt & Kersting, 2003), allowing us to avoid over-extrapolation of our normative judgments and dogmatism about "normative truths". For the purposes of AI alignment, the work that remains to be done is not so much one of formalization, but *integration*: How can these reasoning systems interface with or augment the standard formalisms of probability theory and decision theory? And how can they be combined with algorithms for machine learning and decision-making?

Integrating normative reasoning with machine learning. One relatively straightforward path to integration might be to use normative reasoning frameworks as synthetic data generators: Instead of directly training machine learning systems on human normative judgments, algorithms for normative reasoning could be used to produce sets of internally consistent arguments that can be derived from an initial set of human-provided judgments. Similar to deductive closure training for classical logic (Akyürek et al., 2024), machine learning systems (e.g. LLMs) could then be trained on the sets of derived judgments and arguments,[27] which would hopefully strengthen their ability to produce sound argumentative conclusions, while improving performance at distinguishing incompatible judgments and identifying

[26] While formal theories of reasoning will ultimately have to be evaluated against human judgments themselves, they deliver systematicity and precision that many AI systems do not. Just as with mathematics, logic, and probability theory, formal reasoning systems can succinctly express what we would reflectively endorse, provided that we accept certain principles of reasoning as sound.

[27] Note there might be *multiple* sets of valid or defensible arguments, since an initial set of normative premises might be in conflict without decisively ruling each other out (Dung, 1995) Maintaining this multiplicity may be crucial to avoid normative dogmatism.

self-consistent sets of normative claims. Normative reasoning frameworks could also be used to scaffold and validate the outputs of machine learned systems (Castagna et al., 2024), improving interpretability and correctness while still allowing the overall AI system to work with open-ended (e.g. language) inputs. Finally, one might hope to minimize the role of uninterpretable machine-learned systems altogether, using them primarily for the translation of inputs and outputs while performing most of the reasoning (normative or otherwise) via symbolic model-based algorithms (Wong et al., 2023; Kwon et al., 2023). On this route, the main challenge will be to integrate normative reasoning with frameworks for model-based inference and planning, such as probabilistic programming (van de Meent et al., 2018; Cusumano-Towner et al., 2019).

Considerable work needs to be done before we can design AI that reasons flexibly and generally about preferences and values. Still, there exist many opportunities for research that are under-explored. By taking advantage of them, we might hope to build systems that handle the true normative complexity of the situations we are deploying them into.

4 Beyond single-principal AI alignment as preference matching

If rational choice theory is an inadequate description of human behavior and values, and expected utility theory is an unsatisfactory account of rational decision-making, what does this imply for the practice of AI alignment? Though there is growing awareness of the limits of these preferentist assumptions (Casper et al., 2023; Lambert et al., 2023), most applied methods for AI alignment continue to treat alignment as the problem of *preference matching*: Given an AI system, the goal is to ensure that its behavior conforms with the preferences of a human user or developer.

Reward learning as alignment via preference matching. At present, the most prominent of such methods is reinforcement learning from human feedback (RLHF). Similar to other reward learning methods such as inverse reinforcement learning (Ng & Russell, 2000), RLHF learns an estimate of a user's presumed reward function—a *reward model*—from a dataset of their stated preferences. The AI system is then trained to optimize the learned reward model, with the aim of producing behavior that better conforms to the user's preferences. Since the development of RLHF for classical control problems (Knox & Stone, 2011; Griffith et al., 2013; Akrour et al., 2014), the method has been extended to train increasingly complex AI systems in increasingly open-ended domains, including deep neural networks for robotic control (Christiano et al., 2017) and large language models (Ouyang et al., 2022; Bai et al., 2022). This latter development has led to an explosion of interest in RLHF, given the unprecedented capabilities and general purpose nature of LLMs.

Foundational limitations of reward learning. For all its success, RLHF faces numerous technical challenges (Casper et al., 2023), ranging from issues with preference elicitation (Knox et al., 2024a) and scalable oversight (Leike et al., 2018) to over-optimization (Gao et al., 2023; Moskovitz et al., 2024) and stable training (Hejna et al., 2024). Our focus, however, is more foundational, and applies to not

just RLHF but *any* alignment method derived from reward learning:[28] By committing to a reward representation of human preferences or values, reward learning suffers from all the representational limits we discussed in Sect. 2. Furthermore, by treating reward as something to be optimized, reward-based methods adopt EUT as a normative standard, with all the issues that Sect. 3 describes.

The limited scope of reward learning and preference matching. In this section, we discuss what it would require for AI alignment research to take these challenges seriously. Importantly, we do not claim that reward-based methods are never appropriate. Rather, we argue that reward-based alignment—and preference matching more generally—is only appropriate for AI systems with sufficiently *local* uses and scopes. In other words, it is adequate for only the *narrow* or *minimalist* versions of the value alignment problem, where the values and norms at stake can be summarized as a reward function specific to the system's scope. For sufficiently *ambitious* or *maximalist* attempts at AI alignment,[29] more is necessary: AI systems will have to learn how each person's preferences are dynamically constructed, and be aligned to the underlying values that generate those preferences. Furthermore, when preferences are incomplete, or conflict across time, they have to be aligned with normative ideals about how to assist in such situations. While versions of these points have been made before (Hadfield-Menell & Hadfield, 2018; Gabriel, 2020; Yao et al., 2023), we aim to make precise the connection between values, norms, and preferences, and to illustrate concrete possibilities.

4.1 Beyond alignment with scalar and acontexual rewards

Two aspects of reward functions are important for determining their role in the practice of AI alignment. The first is whether they are *scalar*. As explained in Sect. 2, this corresponds to the question of whether values are treated as fully commensurable, and whether the preferences they represent are complete. The second, often underappreciated aspect, is whether they are *contextual*: Is the reward function understood to be a representation of context-specific preference judgments, or of an individual's overall preferences?

Scalar rewards are only appropriate in narrow decision contexts. Scalar rewards are generally inadequate, since (as elaborated in Sect. 2) they assume away the possibility of incomplete human preferences. But as long as these rewards are also understood to be contextual, then reward-based alignment can be appropriate. In relatively narrow decision contexts without sharp practical or moral dilemmas, it is not unreasonable to assume that people are willing to commensurate their values (Anderson, 1995). In these contexts (e.g. buying groceries, travel planning, solving math homework) it is often clear to us how to weight different values against others

[28] This includes Direct Preference Optimization (Rafailov et al., 2024), Contrastive Preference Learning (Hejna et al., 2024), and Distributional Preference Learning (Siththaranjan et al., 2024).

[29] The distinction between "narrow" and "ambitious" value learning is due to Christiano (2015a), while the analogous distinction between "minimalist" and "maximalist" value alignment is due to Gabriel (2020).

(e.g. quality vs. cost, time vs. comfort, correctness vs. verbosity), leading to a complete preference ordering that it is representable by scalar reward. Learning a reward function is thus not inherently problematic. If this learned reward function is then optimized by a *bounded* AI system—the kind of local, tool-like system we discussed in Sect. 3—then the downsides are also limited. A poorly learned reward function may still result in negative outcomes (Zhuang & Hadfield-Menell, 2020), but the system will not reliably bring about unintended non-local effects.

Models of context-specific preferences will not generalize across contexts. By and large, this is the setting within which methods like RLHF are applied. Reward models are learned from human preferences, but these preferences typically represent context-specific *goodness-of-a-kind* judgments like "How well does this robot achieve its goal?" (Christiano et al., 2017) or "How well do these responses follow the provided instructions?" (Ouyang et al., 2022) While such judgments may implicitly aggregate a number of underlying values like "harmlessness" or "helpfulness" (Bai et al., 2022), they are not judgments of goodness *simpliciter*, or of goodness for the user as a whole. This means that the resulting reward models are only useful for narrow alignment. They can serve as reasonable guides to in-context behavior, but are unlikely to generalize beyond that context (Lambert & Calandra, 2023). In particular, such reward models do not represent human preferences *across* contexts, over an extended period of time.

Context-sensitive preference models as an intermediate solution. What would it take to align an AI system that operates across contexts? One option is the use of context-sensitive reward functions (Pitis et al., 2024), as described in Sect. 3. Though this approach runs the risk of context-manipulating incentives, it may well be adequate for sufficiently bounded systems. Similar to our ECD proposal in Sect. 2, context-sensitivity could be achieved by *per context commensuration* of multiple values, perhaps by learning *separate* reward or preference models for each value (Wu et al., 2024; Go et al., 2024; Xu et al., 2024), then aggregating their rewards with different weights depending on the downstream context. Context switches could then be triggered by users by selecting a desired "mode" (Edwards, 2023) or specifying a system prompt (Pitis et al., 2024).

Still, all that the above amounts to is solving several instances of the narrow alignment problem, then stitching together the answers. If society is on a path towards more general AI systems—say, the globally-scoped AI assistants we discussed in Sect. 3—then we will need more general solutions.

4.2 Beyond alignment with static and asocial preferences

How should one build an AI system that is aligned not to a user in a particular context, but to assist a person over an extended period of time? Addressing this challenge requires a significantly more ambitious solution to the value alignment problem—one that not only avoids the pathologies of expected utility maximization

across global scopes (cf. Sect. 3), but also accounts for the dynamically and socially constructed nature of our preferences.[30]

Most alignment methods do not adequately account for these aspects of preference construction. Instead, they assume that elicited preferences are *static*—they do not change over time—and *asocial*—they are independent of other agent's preferences and societal norms. These are reasonable assumptions if AI systems are only interacting with users over relatively short timescales, and if such interactions can be decoupled from their wider social context. Unfortunately, neither of these assumptions is true in general.

Preferences change via adaptation, drift, learning, reflection, or volition. Contra the first assumption, preferences are *dynamic*: They change, shift, and grow over time (Franklin et al., 2022). This is partly the result of context, as we have discussed, and partly a feature of human psychology: per Kahneman, our stated preferences about an experience can vary with the time of elicitation (Kahneman & Riis, 2005); per Sen and Nussbaum, our preferences adapt to the conditions of what is available to us (Sen et al., 1999; Nussbaum, 2001). More generally preference change is the result of being agents who learn about the world and ourselves as we grow (Loewenstein & Angner, 2003), and who reflect upon and reconsider what we value and desire. As we change our beliefs about what is true, what we find instrumentally valuable changes accordingly. As we discover what we experience as pleasant or unpleasant, what we consider to be intrinsically valuable may also change. We can also *voluntarily* change our values (Ammann, 2023), perhaps by practicing an art form so that we may appreciate it better, or by adopting a new way of life (Chang, 2009; Paul, 2014).

Alignment with informed preferences as a partial solution. Can standard techniques for AI alignment be transplanted to the dynamic context? One modification is to assume that preference change is due only to people learning about their desires over time. In this model, there is still a true underlying preference structure, albeit one initially unknown to the human, and the AI system can just treat those preferences as the target of alignment (Chan et al., 2019). Similar modifications can be applied to the case of changing empirical beliefs: Instead of satisfying a person's revealed preferences, the AI system aims to satisfy what their preferences would be if they were more informed (Reddy et al., 2018). This idea might even be extended to encompass *reflection* upon preferences and values (Cath, 2016): By modeling people as bounded reasoners (Zhi-Xuan et al., 2020; Alanqary et al., 2021), and integrating such models with frameworks for normative reasoning, AI systems could infer what people *would come to want*, if they thought harder about what they truly value.

The challenge of genuine value change. However, alignment with informed preferences avoids the deeper normative questions raised by *genuine* value change: How should an AI system assist someone whose informed preferences change over

[30] Of course, it is always an option to avoid taking up this challenge; there are many transformative uses of AI that do not involve globally scoped personal assistance. Nonetheless, if AI researchers do aim for something like this goal, they should be clear about what it requires.

time due to drift, volition, or transformation? Or what if a person's preferences adapt in response to (potentially oppressive or addictive) environments (Sen et al., 1999; Nussbaum, 2001)? Unlike preference change due to learning or reasoning, there is no sense in which the resulting preferences are more informed or "rational" than they were before. Perhaps AI systems could optimize for a person's *current* preferences, but this risks shifting or manipulating their preferences in undesirable ways (Ashton & Franklin, 2022; Carroll et al., 2022, 2024). Avoiding such shifts would require delineating the kinds of value change that are legitimate or illegitimate (Ammann, 2023), but as Carroll et al. (2024) discuss, it is not obvious how to do so. Alternatively, one might hope to aggregate preferences across the *time-slices* that make up a person (Hedden, 2015), but this introduces difficult questions about how to weight past, present, and future time-slices (Paul, 2014; Pettigrew, 2019), while ignoring the practical unity that individuates a person as a *person* (Korsgaard, 1989; Schechtman, 2014), not just a collection of consciousness moments.

Preferences are socially constructed. We shall return to these normative questions shortly. Before doing so, let us consider the assumption that preferences are *asocial*. In rational choice theory, preferences are typically understood to be an individual's comparative judgments about the outcomes that would be best for them and them alone. These self-regarding preferences are often treated as the target of AI alignment (Hadfield-Menell et al., 2016; Russell, 2019). But of course, many of our preferences are not asocial in this way. Instead, they are *interdependent* (Sobel, 2005): formed not in isolation, but influenced by the preferences, values, and norms of our social and moral circles. Sometimes this influence is merely instrumental—one might prefer to follow a social norm just because it is convenient to do so. But sometimes the influence is *constitutive*—as in a parent's concern for their child's well-being, or a feminist's desire to uphold a norm of equality. If we are to align an AI system with an individual, we will need some way of accounting for these influences.

Recursive preference modeling as a partial solution. As an intermediate solution to the challenge of socially constructed preferences, one might hope to align AI systems with *recursive or interdependent preferences*—preferences which depend on the preferences of others (Sobel, 2005). Such preferences can be modeled with recursive utility functions, which assign weight to the posited utility functions of other agents (Kleiman-Weiner et al., 2017; Kim et al., 2018), or more general models of preference interdependence (Yang & Allenby, 2003). Preferences or "rewards" can also depend on social and moral norms (Bicchieri, 2005; Oldenburg & Zhi-Xuan, 2024), reflecting how people predict and respond to the *normative infrastructure* of their society (Hadfield-Menell & Hadfield, 2018).

Yet, by keeping preferences or utility functions as the target of alignment, recursive preference modeling still faces the many of the limitations we have surveyed. In particular, it still runs the risk of treating preferences as normatively basic, rather than the values and norms that generate those preferences. It also limits our ability to reason about such values and principles, and whether they are *appropriately* influencing an individual's preferences. After all, many social norms and influences are oppressive or otherwise undesirable (Lukacs & Livingstone, 1972; Althusser, 2006), shaping preferences in ways we intuitively regard as contrary to an individual's best

interests. In this sense, the problem of interdependent preferences is similar to the problem of dynamic preferences. In both cases, a range of preference orderings are at play, and without additional normative considerations, it is not clear which set of preferences an AI system should be aligned with (Carroll et al., 2024).

4.3 Beyond preferences as the target of alignment

In light of the challenges introduced by contextual, dynamic, and interdependent preferences, it is difficult to see how they can serve as a coherent alignment target. This also follows from our discussion in Sects. 2 and 3: If preferences are neither psychologically nor normatively basic, then it is not clear what justifies their being the target of value learning and alignment.

Alignment with role-specific normative criteria. This basic point, of course, is not new: As many have long appreciated, identifying someone's welfare or best interests with their preferences runs into a thicket of philosophical issues (Sen et al., 1999; Nussbaum, 2001). Recognizing these issues, Gabriel (2020) argues for an explicitly moral conception of alignment: "the agent does what it morally ought to do, as defined by the individual or society".[31] Others have proposed similar approaches, though they replace "morally ought" with what an agent or humanity as a whole would reflectively endorse, as in ideal observer theories (Firth, 1952; Brandt, 1955) or coherent extrapolated volition (Yudkowsky, 2004). However, it is far from clear how to operationalize these abstract principles. To make progress, we suggest a conception of single-principal alignment that is significantly more constrained: When an AI system only serves an individual in performing a particular task or role, it should be aligned with *the normative ideals or criteria that are appropriate for that role*. For narrow systems, this requires task-specific determination of appropriate normative criteria. For general-purpose AI assistants, this implies alignment with the *normative ideal of an assistant*, rather than alignment to an individual's preferences, or to human normativity writ large.

Existing methods effectively align AI with role-specific norms. Before discussing the case of general-purpose assistants, it is worth noting that many existing alignment methods effectively *function* to align AI systems with task and role-specific norms, even though they are *described* as methods for alignment with human preferences.[32] As discussed earlier, the pairwise judgments provided by human annotators in RLHF are typically not their preferences as end users, but instead context-specific *goodness-of-a-kind* judgments. These judgments are provided in response to questions about whether an AI system's output complies with specific normative criteria—for example, helpfulness, harmlessness, and truthfulness

[31] Gabriel (2020) uses "values" to describe this alignment target, though in a slightly narrower sense than ours. Whereas we have primarily used "values" to refer to evaluative concepts and judgments in general, Gabriel's use implicitly picks out the values that are normatively relevant to AI system behavior.

[32] This preferentist focus is explicit in e.g. Ouyang et al. (2022), who introduce an application of RLHF to LLMs that, in their words, "aligns the behavior of GPT-3 to the stated preferences of a specific group of people."

(Ouyang et al., 2022; Bai et al., 2022). As such, insofar as these judgments can be called preferences, they are *derivative* of normative standards like harmlessness, not the alignment target themselves. Preferences merely serve as data so that machines can *learn* some approximation of these standards. The typical language used to describe reward-learning methods like RLHF is thus misconceived: As used, they are not methods for alignment with any one human's preferences, or for recovering the "true reward function" in some person's head,[33] but for aligning AI systems with contextually-appropriate normative criteria.

Normative criteria for general purpose AI assistants. What then are the normative criteria for general-purpose AI assistants—those globally scoped AI systems for which questions of preference change and incompleteness seem the most pressing? While we cannot give a definitive answer—indeed, as we shall discuss, we think this is something that society will have to collectively decide—we suggest that progress can be made by reflecting on the *normative ideal* of a good assistant.

How does this ideal address the issues with preference alignment that we have raised? Here are a number of suggestions: First, a good assistant does not presume certainty about a person's preferences and values (Hadfield-Menell et al., 2016). This means maintaining an awareness of their own ignorance, while avoiding unwarranted extrapolation of preferences from one context to another, including Knightian uncertainty about how preferences extrapolate (Dalrymple et al., 2024). Second, a good assistant is aware that some choices are hard, and some options may seem incomparable (Chang, 1997). When helping someone with such a choice, the assistant does not pretend to know which option is better, or try to optimize that person's life; instead, the assistant respects their autonomy, and *empowers* them to make the most informed choice possible (Du et al., 2020), while ultimately remaining agnostic as to which choice is "best". Third, a good assistant understands and respects the values of the person they are assisting. This means recognizing that a person's preferences often derive from their values, which can take priority over their immediate requests and preferences (London & Heidari, 2024). The assistant also enables those values to grow and change through normatively acceptable forms of exploration, reflection, volition, or even drift, while avoiding manipulating them or restricting them (Ammann, 2023). Finally, a good assistant, being situated in wider society, respects the preferences and values of *others* (Kirk et al., 2024). When assisting someone who wishes harm out of anger, the assistant might dissuade them from acting against their better nature. When asked to directly harm others, the assistant might refuse.

Pathways to aligning general purpose assistants. In a past era of AI development, these principles might have seemed too vague to formalize or implement. Yet, as our discussion of RLHF suggests, it now seems like we have at least one path towards aligning globally-scoped AI assistants: Train them to comply with human judgments and standards for ideal assistive behavior. Methods such as harmless and helpful RLHF (Bai et al., 2022), (collective) constitutional AI

[33] Supposing the concept of a "human reward function" is even coherent. See Butlin (2021) for a discussion.

(Bai et al., 2022; Huang et al., 2024), and moral graph elicitation (Klingefjord et al., 2024) are already taking steps in this direction, each of them making more explicit that the targets of alignment are not preferences, but normative principles for assistance. Such systems still have to learn the preference of each user they assist, but this is separate from learning *how* to provide assistance in light of those preferences.

Within this broad approach, we can embed many of the proposals we have made in earlier sections. Rich but structured models of human decision-making can serve as the AI assistant's "theory-of-mind", producing well-calibrated estimates of user goals and preferences while avoiding the deficiencies of unstructured approaches (Zhi-Xuan et al., 2024; Kim et al., 2023). Mechanisms for preference incompleteness could be engineered or trained into the AI assistant if this turns out to remove incentives for shutdown avoidance and context manipulation (Thornley, 2024). Theories of normative reasoning could be integrated into AI systems, allowing them to reason about human-provided judgments and principles, while aiding us in deliberating about what counts as good assistance. Each of these proposals may turn out to be strictly unnecessary for the task. Even so, they can provide us helpful guidance as we refine and implement our normative ideals of assistance.

5 Beyond multi-principal AI alignment as preference aggregation

Having argued against a preference-based conception of single-principal alignment, we now turn to the problem of multi-principal alignment: Given the multitude of humans that we share this planet with, and the plurality of values that we hold, what, if anything, should AI systems be aligned to? At least at first glance, it does not seem as though our assistive account of AI alignment can readily be extended to this context. What it means to assist a single person is relatively clear. What it means to assist multiple people—especially people with conflicting values—is far less obvious.

A theoretical argument for preference aggregation. A traditional answer to this question is that AI systems should be aligned to the *aggregate* preferences of humans. Why so? Part of this may be the normative appeal of a preference utilitarian ethic (Hare, 1981). In the AI alignment literature, however, the argument for preference aggregation is usually more technical (Critch & Russell, 2017; Demski, 2018), appealing to Harsanyi's social aggregation theorem as justification (Harsanyi, 1955). Suppose we require that the AI system complies with the (VNM) axioms of expected utility theory. Suppose further that all humans also do so, such that the preferences of each individual i can be represented by a utility function $U_i(x)$ over outcomes x.[34] Finally, assume *unanimity* as a minimal requirement of rational social choice—if *all* humans prefer some (probabilistic) outcome x over outcome y, then the AI system should prefer x over y as well. Then Harsanyi's theorem says that the

[34] Harsanyi's theorem also requires that all humans have *common beliefs* (Critch & Russell, 2017).

AI system's utility function $U(x)$ *must* be a weighted aggregate of individual utility functions:

$$U(x) = w_1 U_1(x) + w_2 U_2(x) + \cdots + w_n U_n(x)$$

where the weights w_i are fixed values independent of the outcome x. By a veil-of-ignorance argument, Harsanyi also proposed that these weights should be *equal*, reasoning that a risk-neutral decision-maker should assign equal probability as to which person they could become (Harsanyi, 1975).

Preference aggregation in the practice of alignment. However convincing one finds this theoretical argument, preference aggregation is often found in the practice of AI alignment as well. A notable example is, once again, RLHF: Despite having been originally designed for single-human contexts, in practice, RLHF is almost always applied to preference datasets collected from *multiple* human labelers (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). This practice has recently been shown equivalent to the Borda count voting rule (Siththaranjan et al., 2024); in effect, each labeler's choices are weighted according to their *ordinal* ranking among the set of possible alternatives.

Practical, political, and foundational limits to preference aggregation. In this section, we critically examine preference aggregation in AI alignment at the practical, political, and foundational levels. At the practical level, we contend that preference aggregation is often misinterpreted and misapplied, such that even if one accepts Harsanyi-style utility aggregation as a normative ideal, it may often be better to use various non-utilitarian aggregation rules in practice. At the political level, we critique the idealized nature of aggregationist approaches, arguing that approaches grounded in bargaining and social contract theory are more politically tractable given our diverse and contested values. At the foundational level, we build upon our arguments against EUT and preference matching from the earlier sections, elaborating them into a critique of the normativity of utilitarian aggregation.

5.1 Beyond naïve utilitarian aggregation of elicited preferences

Different types of preferences are subject to aggregation. Discussion of preference aggregation and its uses is often afflicted by confusion about the nature of preferences. Are these all-things-considered preferences, or goodness-of-a-kind judgments? Are these preferences over outcomes (Harsanyi, 1953), or preferences over ethical views (Baum, 2020)? Are these self-regarding preferences, social preferences, or some combination of the two? For clarity, we shall use the term *welfare preferences* (Rubinstein & Salant, 2012) to refer to those preferences that Harsanyi's theorem most intuitively applies to: These are self-regarding preferences over outcomes that affect one's individual welfare, which exclude consideration of others' welfare. We distinguish this concept from *all-things-considered preferences*, which are preferences about overall goodness (including social or moral considerations), and from *elicited preferences*, which refers to any kind of preference elicited while applying some alignment technique.

Aggregation of elicited preferences need not track aggregate welfare or goodness. The first thing to note is that elicited preferences, welfare preferences, and all-things-considered preferences may all come apart. This crucially affects why and how we aggregate preferences, and whether some utilitarian aggregation procedure should be used. Consider a hypothetical example in the context of RLHF: Users are asked whether they would personally enjoy an LLM that can generate copyrighted short stories, and most of them say yes. If what we care about is aggregate (immediate) welfare, then uniform aggregation of the elicited preferences seems to achieve that goal. But if we what we care about aggregating are all-things-considered value judgments—including legal and moral considerations—then uniform aggregation no longer seems so appropriate.

Similar issues arise when trying to aggregate *toxicity* or *harmfulness* judgments across multiple humans (Bai et al., 2022; Davani et al., 2022). In these cases, the elicited preferences are goodness-of-a-kind judgments, and their connection to aggregate welfare (or all-things-considered goodness) is many steps removed. As such, uniform or majoritarian aggregation can easily fail to achieve social goals. If most human annotators are insensitive to certain forms of identity discrimination (e.g. sexually demeaning images, trans-exclusionary rhetoric, or anti-semitic tropes), then AI systems trained on such data will almost certainly cause harm (Richardson et al., 2019; Okidegbe, 2021). Uniform preference aggregation may thus constitute a form of epistemic injustice (Fricker, 2007; Symons & Alvarado, 2022; Hull, 2023), which in turn leads to downstream injustice and harm.

Non-utilitarian aggregation may be beneficial on normative or epistemic grounds. What aggregation procedures might we use instead? And what justifies their use? In the case of potential copyright violations, we might want to grant veto power to copyright holders, allowing them to *reasonably reject* the welfare-oriented majority preference for copying their work. This veto right could be justified as an instantiation of Scanlon's contractualism (Scanlon, 2000), on the principle that mutual respect among persons necessitates taking claims of intellectual ownership seriously. Alternatively, it could simply be understood as a policy that everyone would reflectively prefer, once they properly understood the costs and benefits of a copyright veto.

As for harmfulness judgments, it may often be preferable to apply *prioritarian* (Lumer et al., 2005; Holtug, 2017) or *egalitarian* (Rawls, 1971) approaches to aggregation. For example, one might select annotators who are *most directly impacted* by potential harms (Gordon et al., 2022), thereby prioritizing certain segments of the population. In cases of significant disagreement, one might even place all weight on the individual with the strongest dispreference (Leben, 2017; Bakker et al., 2022; Weidinger et al., 2023). Again, there are many possible justifications for such procedures. Prioritarian selection could be justified on normative grounds, or because of its epistemic benefits—after all, those most impacted by harms also tend to be *more informed* about their effects (Dror, 2023).

Distinguishing aggregation procedures from standards of rightness. Whatever procedure one favors, it is important not to confuse the aggregation rules used in AI systems with our ultimate social objectives. In practice, these aggregation rules are merely parts of the overall decision procedure implemented by (training) an AI

system, and as many philosophers have pointed out, such procedures should be distinguished from standards of rightness (Railton, 1993; Frazier, 1994; Stark, 1997). Rather than directly instantiating a particular standard (or its mathematical formalization) into a preference aggregation procedure, we should consider which aggregation procedures best satisfy the standard(s) we care about, taking into account practical and informational constraints. In doing so, we should recognize that elicited preferences are typically not the objects of our concern, but simply *information* as to what we truly care about.

5.2 Beyond aggregate preferences as the target of alignment

Suppose we recognize that any particular set of elicited preferences is merely a guide or estimate to what we care about. Even so, one could still imagine taking humanity's aggregate preferences as the *target* of AI alignment. For example, suppose that humanity eventually builds a single powerful AI system—a singleton—that actively infers the preferences of all humans, uses those preferences to estimate humanity's social welfare function, then optimizes its best estimates of that function. In doing so, we might create the ideal utilitarian central planner, achieving what welfare economists and utopian socialists could only dream of (Ng, 1997; Bastani, 2019).

Theoretical difficulties for preference aggregation. Unfortunately, taking aggregate preferences as an alignment target immediately runs into theoretical difficulties. While these issues have been studied at length by social choice theorists,[35] one that is especially challenging for standard utilitarian aggregation is incomparability. As we noted earlier, justifications for preference aggregation typically assume that each individual's preferences can be represented as a utility function, and furthermore that utility can be compared across persons (Harsanyi, 1953, 1975). But as we have elaborated Sect. 2, these assumptions are very much in doubt. Even within a single individual, preferences may be incomplete due to incomparable choices, or not clearly comparable across time (Carroll et al., 2024). Having to compare the goodness of choices across individuals only makes the difficulty more severe (Korinek & Balwit, 2022). This is not to say that the preferability of some outcome can *never* be compared across people,[36] but that any such comparison stands in need of further normative justification (Sen, 1970a; Clayton & Williams, 1999)—justification that, as we argued in Sect. 3, utility theory alone cannot provide.

The computational intractability of aggregate preference optimization. Let us suppose, however, that these theoretical challenges can be addressed.[37] Even so, aggregate preference optimization still faces serious practical challenges. For one,

[35] See Baum (2020); Korinek and Balwit (2022); Mishra (2023) and Conitzer et al. (2024) for discussions of the challenge of applying social choice to AI alignment.

[36] For example, if the choice of person A not wearing a mask would lead to less inconvenience for person A but severe illness for person B, we should intuitively give a stronger weight to person B's preference against severe illness over person A's preference against inconvenience.

[37] Perhaps by using frameworks that allow for partial comparability of welfare across individuals (Sen, 1970b), or by aligning AI with partial social preferences (Korinek & Balwit, 2022).

such optimization is *computationally intractable*: As Austrian economists have long argued, central planning runs into the economic calculation problem (von Mises, 1990), a problem made worse by the sheer complexity of inferring human preferences under limited information, coordinating global production to maximize aggregate preferences, and planning for the future under uncertainty (Hayek, 1945; Murphy, 2006; Cwik & Engelhardt, 2024).[38] In contrast, decentralized decision-making (in the form of e.g. competitive markets) can sometimes be exponentially more efficient in computational cost than central planning (Rust, 1996), while achieving optimal informational efficiency (Mount & Reiter, 1974; Jordan, 1982). As such, even if not a practical impossibility, optimizing humanity's aggregate preferences with a single AI system is likely to be considerably less efficient than more pluralistic alternatives (Siddarth et al., 2022).

The politically infeasibility of impartially benevolent AI. Perhaps even more importantly, the project of building AI that optimizes humanity's aggregate preferences is *politically infeasible*: Even if impartially benevolent AI planners were possible to develop, building such systems would be incompatible with the incentives of every AI developer with a realistic chance of doing so. This is the case even for AI developers with expressedly pro-social missions, which are still subject to market incentives as a result of the need to raise capital (Toner & McCauley, 2024), and are still governed by the laws and regulations of the countries they are based in. Allowing the creation of such AI systems would also risk the centralization of immense power: However virtuous the goal of impartial preference optimization might seem, the history of central planning should tell us that optimal social outcomes are far from likely to be achieved (Scott, 1998; Verdery, 2005). Instead, we are more likely to see a tyranny of creator values, with potentially disastrous consequences for everyone with a contrary way of life.

Pluralistic alignment as a politically feasible alternative. In light of these challenges, how should we reconceive the goals of multi-principal AI alignment? One constraint in doing so is incentive compatibility: Whatever our vision of AI alignment is, it should account for divergent interests and contested values, credibly enabling collective safety and stability by ensuring incentives for cooperation and minimizing the chances of conflict (Critch & Krueger, 2020; Dafoe et al., 2020). A related constraint is political feasibility: Alternative targets for alignment should be achievable given the political economy of actually existing AI—an economy that consists of a wide variety of AI services developed and deployed by a large number of self-interested actors (Drexler, 2019). Although these are negative constraints, they pair well with a more positive, pluralistic vision of what alignment could enable: A world where increasingly advanced AI systems serve a diversity of individual, communal, and universal ends, without catastrophically endangering anyone's interests (Zhi-Xuan, 2022; Gabriel, 2020; Siddarth & Huang, 2023).

[38] These difficulties can be formalized with the aid of theoretical computer science, which shows that optimal planning under uncertainty is sometimes undecidable, and even when decidable, remains anywhere from PSPACE to EXPTIME-complete (Papadimitriou & Tsitsiklis, 1987; Chatterjee et al., 2016).

Enabling pluralism through political constraints. What would it require to enact this pluralistic vision? As a starting point, consider our principles for AI assistance from Sect. 4. While an AI assistant primarily serves a single person, and might be personalized to do so in many ways (Sorensen et al., 2024), our presumptive norms for ideal assistance do not permit disregard for others. Rather, they endorse a *circumscribed* promotion of the person's interests and values, such that the assistant avoids harming and endangering other individuals.[39] These norms function as political constraints, allowing assistants to provide value for individual users without imposing unreasonable externalities upon others.[40] In doing so, they reduce the chance of conflict and non-cooperation.

Alignment with politically negotiated normative standards. Our suggestion then, is that this approach can be generalized to broadly *contractualist* account of AI alignment (Zhi-Xuan, 2022):[41] Rather than learning humanity's preferences in order to maximally satisfy them, AI systems should be aligned with normative standards and criteria that we collectively forge and negotiate—standards exemplified by social, legal, and moral norms. These norms may be constructed as we design each system, or can be decided in advance for entire classes of AI systems. Returning to our earlier discussion of role-specific alignment, what is important is that these norms are tailored to the scope and uses of each system: Just as AI assistants should avoid harmful language, self-driving cars should follow the rules of the road. By negotiating norms and constraints for each of AI's social functions, we can enable a plurality of uses for AI while limiting the costs and harms to all stakeholders involved.

The practical benefits of contractualist alignment. What benefits does a contractualist approach to alignment offer? In our view, its primary benefits are practical ones: Unlike aggregate preference optimization, contractualist alignment does not require unrealistic amounts of benevolence from any one actor. Instead, it aims for a regime where largely self-interested actors stand to mutually benefit from the development and deployment of AI. Well-designed norms and institutions enable this, stabilizing cooperation by making it costly for relevant parties to defect or withdraw from cooperation (Kalai & Smorodinsky, 1975; Gintis, 2010). Aligning AI systems to comply with cooperative norms (and perhaps even to enforce them) thus reduces the chance of AI-caused or mediated conflict, or the risk of (catastrophically) endangering anyone's interests. Norms also limit the computational and informational cost of ensuring aligned behavior: Rather than inferring a large number of preferences, norm-aligned agents just have to (learn to) comply with a limited set of constraints (Oldenburg & Zhi-Xuan, 2024). Finally, by centering norms and principles as the targets of AI alignment, political deliberation becomes more feasible and widely accessible (Huang et al., 2024): Stakeholders need not negotiate over every

[39] This might be viewed as an instantiation of the Harm Principle (Mill, 1859) for AI assistants.

[40] See also Kirk et al. (2024) on the bounds of personalization in LLM assistants, and Gabriel and Keeling (2024) for an explicitly political conception of AI alignment.

[41] We use "contractualist" here in a broad sense, which includes both contractarianism (Cudd & Eftekhari, 2021) and Rawlsian (Rawls, 1971) or Scanlonian contractualism (Scanlon, 2000).

last detail over how an AI system is built or trained, but can instead agree upon high-level requirements and standards for how the system should behave.[42]

The normative grounds of contractualist alignment. Besides its practical benefits, contractualist alignment can also be grounded in normative foundations that are more compatible with a pluralistic world. While it might be possible to justify broadly contractualist principle-setting on rule consequentialist grounds (Parfit, 2011), we contend that the normative appeal of contractualist alignment is precisely that it avoids a universal account of what consequences are better or worse.[43] Given the difficulties with comparability that we have examined, it is unlikely that people will ever agree upon a single scale of value for ranking all consequences. Instead, contractualist alignment aims to align AI systems with goals, standards, and principles that are mutually agreed upon by people despite our disparate preferences and values, deriving its normative force from the fair and impartial agreement of relevantly-situated rational actors.

Conditions for fair and impartial agreement. What makes an agreement impartial or fair? As in contractarian moral and political theories (Gauthier, 1986; Binmore, 1994), it may be enough that all stakeholders benefit relative to an originally fair bargaining position, subject to additional symmetry constraints. Or as Rawls (1993) and Scanlon (2000) respectively argue, a thicker conception of public reason and the mutual recognition of each other as reasonable persons may be necessary to decide which agreements are fair. While examining these questions would take us beyond the scope of this paper, we believe our critique of expected utility theory lends itself to thicker conceptions of fair and reasonable agreement. On such conceptions, AI systems should not just be aligned with goals and standards that achieve mutual benefit.[44] Instead, AI goals and standards should be *justified* to each stakeholder, on grounds that none can reasonably reject. Insofar as these AI systems are used to exercise power over others, they should also act in accordance with standards that are not just fair, but *legitimate* (Lazar, 2024; Stone & Mittelstadt, 2024).

Alignment in the absence of agreement. A natural worry for contractualist alignment is the possibility that agreement between different stakeholders may not be obtained (let alone agreement that is impartial and fair). Yet, this worry is not as acute as it may initially seem. First, rather than aligning AI systems with norms that have *actually* been agreed upon, we could align them with norms that would *hypothetically* be agreed upon, in the spirit of virtual bargaining (Misyak et al., 2014; Chater, 2023). This would generally be necessary to handle incompletely specified agreements and contracts (Hadfield-Menell & Hadfield, 2018), while sharply lowering the cost and frequency of actual negotiations. Second, there are many cases

[42] This does not preclude lower-level forms feedback such as participatory data labeling (Gordon et al., 2022) or end-user audits (Lam et al., 2022), which can complement the aim of mutually-acceptable AI design.

[43] Similar arguments are made by Gabriel (2020) and Gabriel and Keeling (2024).

[44] After all, mutual benefit is not always achievable. In such cases, it is still possible to reach agreements that are viewed as fair, as in an agreement to compensate someone for harm.

where the operation of an AI system imposes minimal externalities upon others, and hence the cost of disagreement between AI stakeholders is merely that the gains of cooperation cannot be realized. In such cases, it is no great loss if each party operates their own AI system aligned with their individual goals, rather than having a shared AI system aligned with collective goals and norms. It is only when AI systems do impose substantial negative externalities that disagreement about their operation is more dangerous. These situations could well lead to mutually destructive conflict, as in prisoner's dilemma scenarios, or exploitative outcomes, where some AI operators benefit significantly at the expense of others. Even so, humans still have the political agency to shape which agreements are feasible and fair, and there is reason to hope that parties will negotiate to avoid at least the worst AI outcomes (e.g. in the form of minimal safety standards). Finally, achieving agreement over norms and principles is likely to be far easier than agreeing on a metric for globally ranking all consequences or comparing all people's preferences. As such, unless one is willing to allow a small set of actors to decide how all of humanity's preferences should be weighted and compared, utilitarian preference aggregation faces an even sharper risk of disagreement and conflict than a contractualist approach.

Technical avenues toward contractualist alignment. If we accept this contractualist understanding of multi-principal alignment, then much work remains to be done. On the technical front, there need to be advances in the theory and implementation of cooperative or contractualist decision-making. While recent alignment techniques show how language-based AI assistants can be aligned with collectively elicited norms and values, and how divergences in norms, opinions, and values can be reconciled through agreement (Huang et al., 2024), iterative critique (Bakker et al., 2022), or moral reflection (Klingefjord et al., 2024), these methods are specialized to a particular type of AI system, and have yet to be situated in a more general theoretical framework. To develop such a framework, we suspect that it will be necessary to unite ideas from game theory (Dafoe et al., 2020), bargaining theory (Chater, 2023), and social choice (Conitzer et al., 2024) with formal approaches to argumentation (Amgoud & Cayrol, 1998) and negotiation (Rahwan et al., 2003), along with insights from the science of human normativity (Binmore, 1994; Hadfield-Menell & Hadfield, 2018; Levine et al., 2023). In particular, by developing computational theories of how humans rapidly learn extant norms and conventions (Tan & Ong, 2019; Hadfield-Menell et al., 2019; Hawkins et al., 2019), recognize institutional structure (Jara-Ettinger & Dunham, 2024; Baker et al., 2024), and engage in contractualist reasoning about social and moral norms (Levine et al., 2023, 2024), we can inform the design of AI systems with social and normative competence: AI that is not just aligned with stakeholder values in a once-off process, but which flexibly adapts to our norms and institutions as they evolve (Oldenburg & Zhi-Xuan, 2024), reasons about their applicability in novel situations (Kwon et al., 2023), and perhaps even aids us in negotiating new contracts and norms (Christoffersen et al., 2023; Jarrett et al., 2023; Tessler et al., 2024).

Social and political avenues toward contractualist alignment. Of course, if we take fair and impartially negotiated standards as the target of AI alignment, then technical advances will not be enough; we also need to foster the development of social, economic, and political orders that provide the conditions for free and fair

agreement. This might involve the creation of new economic and political mechanisms that elicit and consolidate the interests of AI stakeholders (Siddarth & Huang, 2023), the establishment of democratic processes and bodies that can exercise legitimate authority over AI systems (Ovadya, 2023), or the expansion of participatory approaches to AI development and design (Birhane et al., 2022; Suresh et al., 2024). Without these social and political investments, we will lack the capacity to surface our reasons and values to AI systems that act on our behalf, and the accountability to ensure that each of our interests is fairly represented. After all, if we are going to align AI systems with normative standards we would collectively endorse, then we had better make sure that a "we" exists to endorse them.

6 Conclusion

Preference is a central concept in both the theory and practice of AI alignment. Yet as we have seen, its multiple scopes and meanings are often poorly understood. In this paper, we have sought not only to better contextualize the nature of preferences, but also to challenge its centrality in approaches to AI alignment. In doing so, we hope to have established the goals of AI alignment on firmer normative ground. Crucially, we do not do so by rejecting all preference-based frameworks in alignment, but by reinterpreting what preferences do for us: Since they are *constructed* from our values, norms, and reasons, they are *informative* of those underlying structures. As such, preferences can serve as proxies for our values, but not targets of alignment in and of themselves.

What would AI alignment look like if it took these challenges seriously? It would move away from naive rational choice models of human decision making, towards richer models that include how we evaluate, commensurate, and act upon our values in boundedly rational ways. It would no longer take for granted expected utility theory, and instead explore systems for reasoning about the normativity of our preferences and values. It would learn to distinguish goodness-of-a-kind preferences from all-things-considered preferences, and identify which of those are operative in any particular decision. It would let go of preference matching as a crisp formalization of alignment, and instead lean into the normative complexity of scoping and defining AI's social roles. And it would move beyond alignment with aggregate preferences, towards a more pluralistic and contractualist understanding of what it means to live together with AI. If successful, then perhaps the world we can look forward to is not just one we will prefer, but one that we will truly have reason to value.

Acknowledgements This paper benefited from comments and feedback provided by participants at the closing retreat of the 2023 Principles of Intelligent Behavior in Biological and Social Systems (PIBBSS) Summer Fellowship, where an early version of this work was presented. We would also like to thank participants of the 2024 Sociotechnical AI Safety Workshop in Rio de Janeiro for their engagement and suggestions, and Seth Lazar for organizing the workshop. Conversations with many individuals informed the development and presentation of the ideas in this paper, including members of the 2019–2020 MIT AI Alignment Reading Group, Tushita Jha, Jonathan Stray, Iason Gabriel, Nora Ammann, Cecilia Wood, Mateusz Bagiński, Joe Kwon, Sydney Levine, Max Kleiman-Weiner, Max Langenkamp, Saffron Huang, Divya Siddarth, Gillian Hadfield, Vikash Mansinghka, and Joshua Tenenbaum. Finally, we thank the

editor and our anonymous reviewers for their highly detailed feedback and suggestions, which improved the clarity of our paper on many technical and conceptual points. Tan Zhi-Xuan is funded by the Open Philanthropy AI Fellowship. Micah Carroll is funded by the NSF Fellowship. Matija Franklin was funded by a UCL demonstratorship.

Funding 'Open Access funding provided by the MIT Libraries'.

References

Ab Azar, N., Shahmansoorian, A., & Davoudi, M. (2020). From inverse optimal control to inverse reinforcement learning: A historical review. *Annual Reviews in Control, 50*, 119–138.

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning* (pp. 1).

Abel, D., Dabney, W., Harutyunyan, A., Ho, M. K., Littman, M., Precup, D., & Singh, S. (2021). On the expressivity of Markov reward. *Advances in Neural Information Processing Systems 34*.

Akrour, R., Schoenauer, M., Sebag, M., & Souplet, J. C. (2014). Programming by feedback. In *International Conference on Machine Learning* (pp. 1503–1511). JMLR. org.

Akyürek, A. F., Akyürek, E., Choshen, L., Wijaya, D., & Andreas, J. (2024). Deductive closure training of language models for coherence, accuracy, and updatability. In *Findings of the Association for Computational Linguistics ACL 2024*. Association for Computational Linguistics.

Alanqary, A., Lin, G. Z., Le, J., Zhi-Xuan, T., Mansinghka, V. K., & Tenenbaum, J. B. (2021). Modeling the mistakes of boundedly rational agents within a Bayesian theory of mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43).

Althusser, L., et al. (2006). Ideology and Ideological State Apparatuses. *The Anthropology of the State: A Reader, 9*(1), 86–98.

Amgoud, L., Ben-Naim, J., Doder, D., & Vesic, S. (2017). Acceptability semantics for weighted argumentation frameworks. In *Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*.

Amgoud, L., & Cayrol, C. (1998). On the acceptability of arguments in preference-based argumentation. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*.

Ammann, N. (2023). The value change problem (sequence). *AI Alignment Forum*.

Amplayo, R. K., Hwang, S. W., & Song, M. (2019). Evaluating research novelty detection: Counterfactual approaches. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)* (pp. 124–133).

Anderson, E. (1995). *Value in Ethics and Economics*. Harvard University Press.

Armstrong, S. (2019). Synthesising a human's preferences into a utility function. *AI Alignment Forum*.

Armstrong, S., & Levinstein, B. (2017). Low impact artificial intelligences. arXiv:1705.10720 [cs].

Armstrong, S., & Mindermann, S. (2018). Occam's razor is insufficient to infer the preferences of irrational agents. *Advances in Neural Information Processing Systems* 31 .

Armstrong, S., & O'Rourke, X. (2017). Indifference methods for managing agent rewards. arXiv:1712.06365.

Ashton, H., & Franklin, M. (2022). The problem of behaviour and preference manipulation in AI systems. In *The AAAI-22 Workshop on Artificial Intelligence Safety (SafeAI 2022)*.

Azari Soufiani, H., Diao, H., Lai, Z., & Parkes, D. C. (2013). Generalized random utility models with multiple types. *Advances in Neural Information Processing Systems, 26*.

Baber, H. E. (2011). Preference-satisfaction. In D. K. Chatterjee (Ed.), *Encyclopedia of Global Justice* (pp. 890–896). Dordrecht: Springer. https://doi.org/10.1007/978-1-4020-9160-5_436.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.

Baker, A., Dunham, Y., & Jara-Ettinger, J. (2024). Roles guide rapid inferences about agent knowledge and behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329–349.

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

Bales, A. (2023). Will AI avoid exploitation? Artificial General Intelligence and expected utility theory. *Philosophical Studies,* 1–20.

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., & Silver, D. (2017). Successor features for transfer in reinforcement learning. *Advances in Neural Information Processing Systems, 30*.

Bastani, A. (2019). *Fully Automated Luxury Communism*. Verso Books.

Baum, S. D. (2020). Social choice ethics in artificial intelligence. *AI & Society, 35*(1), 165–176. https://doi.org/10.1007/s00146-017-0760-1

Bengio, Y. (2023). AI scientists: Safe and useful AI? https://yoshuabengio.org/2023/05/07/ai-scientists-safe-and-useful-ai/.

Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation*. T. Payne and Son.

Berger, J. (2013). *Statistical Decision Theory: Foundations, Concepts, and Methods*. Springer.

Berke, M., Tenenbaum, A., Sterling, B., & Jara-Ettinger, J. (2023). Thinking about thinking as rational computation. In *Proceedings of the Annual Conference of the Cognitive Science Society*.

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Binmore, K. G. (1994). *Game theory and the social contract*. MIT Press.

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1–8).

Blili-Hamelin, B., & Hancox-Li, L. (2023). Making intelligence: Ethical values in IQ and ML benchmarks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 271–284).

Blili-Hamelin, B., Hancox-Li, L., & Smart, A. (2024). Unsocial Intelligence: An Investigation of the Assumptions of AGI Discourse. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (vol. 7, pp. 141–155).

Bobu, A., Peng, A., Agrawal, P., Shah, J., & Dragan, A. D. (2024). Aligning robot and human representations. In *Proceedings of the 2024 ACM/IEEE International Conference on Human–Robot Interaction*. Association for Computing Machinery.

Bobu, A., Wiggert, M., Tomlin, C., & Dragan, A. D. (2022). Inducing structure in reward learning by learning features. *The International Journal of Robotics Research, 41*(5), 497–518.

Bolker, E. D. (1967). A simultaneous axiomatization of utility and subjective probability. *Philosophy of Science, 34*(4), 333–340. https://doi.org/10.1086/288171

Booth, S., Knox, W. B., Shah, J., Niekum, S., Stone, P., & Allievi, A. (2023). The perils of trial-and-error reward design: Misdesign through overfitting and invalid task specifications. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*, 5920–5929.

Bossaerts, P., Yadav, N., & Murawski, C. (2019). Uncertainty and computational complexity. *Philosophical Transactions of the Royal Society B, 374*(1766), 20180138.

Boudon, R. (2003). Beyond rational choice theory. *Annual Review of Sociology, 29*(1), 1–21.

Boutilier, C., Brafman, R. I., Domshlak, C., Hoos, H. H., & Poole, D. (2004). February. CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research, 21*, 135–191. https://doi.org/10.1613/jair.1234

Bowling, M., Martin, J. D., Abel, D., & Dabney, W. (2023). Settling the reward hypothesis. In *International Conference on Machine Learning* (pp. 3003–3020). PMLR.

Bradley, S., & Steele, K. (2014). Should subjective probabilities be sharp? *Episteme, 11*(3), 277–289.

Brandt, R. B. (1955). The definition of an "ideal observer'' theory in ethics. *Philosophy and Phenomenological Research, 15*(3), 407–413.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press.

Bratman, M. E., Israel, D. J., & Pollack, M. E. (1988). Plans and resource-bounded practical reasoning. *Computational Intelligence, 4*(3), 349–355.

Butlin, P. (2021). AI alignment and human reward. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 437–445).

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour, 6*(8), 1112–1125.

Camara, M. K. (2022). Computationally tractable choice. In *Proceedings of the 23rd ACM Conference on Economics and Computation* (pp. 28–28).

Cao, H., Cohen, S., & Szpruch, L. (2021). Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems, 34*, 12362–12373.

Carlsmith, J. (2022). Is power-seeking AI an existential risk? arXiv:2206.13353.

Carroll, M., Chan, A., Ashton, H., & Krueger, D. (2023). Characterizing manipulation from AI systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1–13).

Carroll, M., Foote, D., Siththaranjan, A., Russell, S., & Dragan, A. (2024). AI alignment with changing and influenceable reward functions. In *Proceedings of the 41st International Conference on Machine Learning* (pp. 5706–5756).

Carroll, M. D., Dragan, A., Russell, S., & Hadfield-Menell, D. (2022). Estimating and penalizing induced preference shifts in recommender systems. In *International Conference on Machine Learning* (pp. 2686–2708). PMLR.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*.

Castagna, F., Sassoon, I., & Parsons, S. (2024). Can formal argumentative reasoning enhance LLMs' performances? arXiv:2405.13036.

Cath, Y. (2016). Reflective equilibrium. In *The Oxford Handbook of Philosophical Methodology* (Vol. 1).

Cettolin, E., & Riedl, A. (2019). Revealed preferences under uncertainty: Incomplete preferences and preferences for randomization. *Journal of Economic Theory, 181*, 547–585.

Chan, L., Critch, A., & Dragan, A. (2021). Human irrationality: Both bad and good for reward inference. arXiv:2111.06956.

Chan, L., Hadfield-Menell, D., Srinivasa, S., & Dragan, A. (2019). The Assistive Multi-Armed Bandit. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 354–363). IEEE.

Chang, R. (Ed.). (1997). *Incommensurability, Incomparability, and Practical Reason*. Cambridge: Harvard.

Chang, R. (2004). Can desires provide reasons for action. In R. J. Wallace, P. Pettit, S. Scheffler, & M. Smith (Eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (pp. 56–90). Oxford University Press.

Chang, R. (2009). Voluntarist reasons and the sources of normativity. In D. Sobel & S. Wall (Eds.), *Reasons for Action* (pp. 243–71). Cambridge University Press.

Chang, R. (2021). How to prevent AI from taking over the world. https://www.newstatesman.com/ideas/2021/02/how-prevent-ai-taking-over-world.

Chater, N. (2023). How could we make a social robot? A virtual bargaining approach. *Philosophical Transactions of the Royal Society A, 381*(2251), 20220040.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences, 3*(2), 57–65.

Chatterjee, K., Chmelik, M., & Tracol, M. (2016). What is decidable about partially observable Markov decision processes with ω-regular objectives. *Journal of Computer and System Sciences, 82*(5), 878–911.

Christiano, P. (2015a). Ambitious vs. narrow value learning. https://ai-alignment.com/ambitious-vs-narrow-value-learning-99bd0c59847e.

Christiano, P. (2015). The easy goal inference problem is still hard. *AI Alignment Forum*.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems 30*.

Christoffersen, P. J., Haupt, A. A., & Hadfield-Menell, D. (2023). Get it in writing: Formal contracts mitigate social dilemmas in multi-agent RL. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (pp. 448–456).

Clayton, M., & Williams, A. (1999). Egalitarian justice and interpersonal comparison. *European Journal of Political Research, 35*(4), 445–464.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., et al. (2024). Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Forty-First International Conference on Machine Learning*.

Cornelio, C., Goldsmith, J., Mattei, N., Rossi, F., & Venable, K. B. (2013). Updates and uncertainty in CP-nets, In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Cranefield, S., & Nayak, A. (Eds.), *AI 2013: Advances in Artificial Intelligence*. Series Title: Lecture Notes in Computer Science (Vol. 8272, pp. 301–312). Cham: Springer. https://doi.org/10.1007/978-3-319-03680-9_32.

Critch, A., & Krueger, D. (2020). AI research considerations for human existential safety (ARCHES). arXiv:2006.04948.

Critch, A., & Russell, S. (2017). Servant of many masters: Shifting priorities in Pareto-optimal sequential decision-making. arXiv:1711.00363.

Cudd, A., & Eftekhari, S. (2021). Contractarianism, In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 ed.). Metaphysics Research Lab, Stanford University.

Cusumano-Towner, M. F., Saad, F. A., Lew, A. K., & Mansinghka, V. K. (2019). Gen: A general-purpose probabilistic programming system with programmable inference. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation* (pp. 221–236).

Cwik, P. & Engelhardt, L. (2024). Revisiting the computation problem. *Quarterly Journal of Austrian Economics 26*(3).

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., & Graepel, T. (2020). Open problems in cooperative AI. arXiv:2012.08630.

Dalrymple, D. D. (2022). You can still fetch the coffee today if you're dead tomorrow. *AI Alignment Forum*.

Dalrymple, D. D. (2024). *Safeguarded AI: Constructing Guaranteed Safety*. ARIA: Technical report.

Dalrymple, D. D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., et al. (2024). Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. arXiv:2405.06624.

Davani, A. M., Díaz, M., & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics, 10*, 92–110.

Davidson, G., Todd, G., Togelius, J., Gureckis, T. M., & Lake, B. M. (2024). Goals as reward-producing programs. arXiv:2405.13242.

De Raedt, L., & Kersting, K. (2003). Probabilistic logic learning. *ACM SIGKDD Explorations Newsletter, 5*(1), 31–48.

Demski, A. (2018). Complete class: Consequentialist foundations. *AI Alignment Forum*.

Denoeux, T., & Shenoy, P. P. (2020). An interval-valued utility theory for decision making with Dempster-Shafer belief functions. *International Journal of Approximate Reasoning, 124*, 194–216.

Dewey, D. (2011). Learning what to value. In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings 4* (pp. 309–314). Springer.

Di Langosco, L. L., Koch, J., Sharkey, L. D., Pfau, J., & Krueger, D. (2022). Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning* (pp. 12004–12019). PMLR.

Drexler, E. (2022). The open agency model. *AI Alignment Forum*.

Drexler, K. E. (2019). Reframing Superintelligence: Comprehensive AI services as general intelligence. Technical Report 2019-1, Future of Humanity Institute, Oxford.
Dror, L. (2023). Is there an epistemic advantage to being oppressed? *Noûs, 57*(3), 618–640.
Du, Y., Tiomkin, S., Kiciman, E., Polani, D., Abbeel, P., & Dragan, A. (2020). AVE: Assistance via empowerment. *Advances in Neural Information Processing Systems, 33*, 4560–4571.
Dumoulin, V., Johnson, D. D., Castro, P. S., Larochelle, H., & Dauphin, Y. (2024). A density estimation perspective on learning from pairwise human preferences. *Transactions on Machine Learning Research*.
Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence, 77*(2), 321–357.
Dziri, N., Lu, X., Sclar, M., Li, X. L., Jian, L., Lin, B. Y., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., et al. (2023). Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems 36*.
Eckersley, P. (2018). Impossibility and uncertainty theorems in AI value alignment (or why your AGI should not have a utility function). arXiv:1901.00064.
Edwards, B. (2023). *March*. Ars Technica: AI-powered Bing chat gains three distinct personalities.
Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology, 24*(1), 313–343.
Evans, O., Stuhlmüller, A., & Goodman, N. (2016). Learning the preferences of ignorant, inconsistent agents. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30).
Everitt, T., Carey, R., Langlois, E. D., Ortega, P. A., & Legg, S. (2021). Agent incentives: A causal perspective. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, 11487–11495.
Fickinger, A., Zhuang, S., Hadfield-Menell, D., & Russell, S. (2020). Multi-principal assistance games. arXiv:2007.09540 [cs].
Fikes, R. E., & Nilsson, N. J. (1971). STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence, 2*(3–4), 189–208.
Firth, R. (1952). Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research, 12*(3), 317–345.
Franklin, M., Ashton, H., Gorman, R., & Armstrong, S. (2022). Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of AI. In *AAAI-22 Workshop on AI for Behavior Change*.
Frazier, R. L. (1994). Act utilitarianism and decision procedures. *Utilitas, 6*(1), 43–53.
Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.
Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437.
Gabriel, I., & Keeling, G. (2024). A matter of principle? AI alignment as the fair treatment of claims. Under Review.
Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. In *International Conference on Machine Learning* (pp. 10835–10866). PMLR.
Garrabrant, S. (2022). Geometric rationality. *AI Alignment Forum*.
Gauthier, D. (1986). *Morals by Agreement*. Clarendon Press.
Gerevini, A. & Long, D. (2005). Plan constraints and preferences in PDDL3. Technical report, Department of Electronics for Automation, University of Brescia.
Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*(6245), 273–278.
Ghosal, G. R., Zurek, M., Brown, D. S., & Dragan, A. D. (2023). The effect of modeling human rationality level on learning rewards from multiple feedback types. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*, 5983–5992.
Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3*(1), 20–29.
Gintis, H. (2010). Social norms as choreography. *Politics, Philosophy and Economics, 9*(3), 251–264.
Go, D., Korbak, T., Kruszewski, G., Rozen, J., & Dymetman, M. (2024). Compositional preference models for aligning LMs. In *The Twelfth International Conference on Learning Representations*.
Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J., Hashimoto, T., & Bernstein, M. S. (2022). Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–19).
Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society, 7*(1), 2053951719897945.

Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., & Thomaz, A. L. (2013). Policy shaping: Integrating human feedback with reinforcement learning. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 26). Curran Associates Inc.

Gustafsson, J. E. (2022). *Money-pump arguments*. Cambridge University Press.

Hadfield-Menell, D., Andrus, M., & Hadfield, G. (2019). Legible normativity for AI alignment: The value of silly rules. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 115–121).

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Hadfield-Menell, D. & Hadfield, G. K. (2018). Incomplete contracting and AI alignment. *USC CLASS research papers series no. CLASS18-10*.

Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S. J., & Dragan, A. (2017). Inverse reward design. *Advances in Neural Information Processing Systems, 30*.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems, 29*.

Halpern, J. Y., & Pass, R. (2015). Algorithmic rationality: Game theory with costly computation. *Journal of Economic Theory, 156*, 246–268.

Hansson, S. O. (1990). Preference-based deontic logic. *Journal of Philosophical Logic, 19*, 75–93.

Hare, R. M. (1981). *Moral Thinking: Its Levels, Method, and Point*. Oxford: Clarendon Press, Oxford University Press.

Harsanyi, J. C. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy, 61*(5), 434–435.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy, 63*(4), 309–321.

Harsanyi, J. C. (1975). Can the maximin principle serve as a basis for morality? A critique of john rawls's theory. *American Political Science Review, 69*(2), 594–606.

Hawkins, R. X., Goodman, N. D., & Goldstone, R. L. (2019). The emergence of social norms and conventions. *Trends in Cognitive Sciences, 23*(2), 158–169.

Hayden, B. Y., & Niv, Y. (2021). The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *Behavioral Neuroscience, 135*(2), 192.

Hayek, F. (1945). The use of knowledge in society. *American Economic Review 35*(4) .

Hedden, B. (2015). *Reasons Without Persons: Rationality, Identity, and Time*. OUP Oxford.

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., & Sadigh, D. (2024). Contrastive preference learning: Learning from human feedback without reinforcement learning. In *The Twelfth International Conference on Learning Representations*.

Hendrycks, D. (2023). Natural selection favors AIs over humans. arXiv:2303.16200.

Hill, D. N., Nassif, H., Liu, Y., Iyer, A., & Vishwanathan, S. (2017). An efficient bandit algorithm for realtime multivariate optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1813–1821).

Ho, M. K., & Griffiths, T. L. (2022). Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *Annual Review of Control, Robotics, and Autonomous Systems, 5*, 33–53.

Holtman, K. (2019). Corrigibility with utility preservation. arXiv:1908.01695.

Holtug, N. (2017). *Prioritarianism*. Oxford Research Encyclopedia of Politics: Oxford University Press.

Horowitz, J. L., Bolduc, D., Divakar, S., Geweke, J., Gönül, F., Hajivassiliou, V., Koppelman, F. S., Keane, M., Matzkin, R., Rossi, P., et al. (1994). Advances in random utility models. *Marketing Letters, 5*, 311–322.

Huang, S., Siddarth, D., Lovitt, L., Liao, T. I., Durmus, E., Tamkin, A., & Ganguli, D. (2024). Collective Constitutional AI: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024)*. ACM.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. arXiv:1906.01820.

Hull, G. (2023). Dirty data labeled dirt cheap: Epistemic injustice in machine learning systems. *Ethics and Information Technology, 25*(3), 38.

Icarte, R. T., Klassen, T. Q., Valenzano, R., & McIlraith, S. A. (2022). Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research, 73*, 173–208.

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv:1805.00899.

Iu, K. Y. & Wong, V. M. Y. (2023). ChatGPT by OpenAI: The end of litigation lawyers? Available at SSRN 4339839.

Jacob, A. P., Gupta, A., and Andreas, J. (2024). Modeling boundedly rational agents with latent inference budgets. In *The Twelfth International Conference on Learning Representations*.

Jara-Ettinger, J. & Y. Dunham. (2024). The institutional stance. *PsyArXiv*.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology, 123*, 101334.

Jarrett, D., Hüyük, A., & Van Der Schaar, M. (2021). Inverse decision modeling: Learning interpretable representations of behavior. In *International Conference on Machine Learning* (pp. 4755–4771). PMLR.

Jarrett, D., Pislar, M., Bakker, M. A., Tessler, M. H., Koster, R., Balaguer, J., Elie, R., Summerfield, C., & Tacchetti, A. (2023). Language agents as digital representatives in collective decision-making. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics, 4*(3), 227–241.

Jeffrey, R. C. (1991). *The Logic of Decision* (2nd ed.). Chicago University Press.

Jeon, H. J., Milli, S., & Dragan, A. (2020). Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems, 33*, 4415–4426.

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel S., et al. (2021). Can machines learn morality? The Delphi experiment. arXiv:2110.07574.

Jin, Z., Levine, S., Gonzalez Adauto, F., Kamal, O., Sap, M., Sachan, M., Mihalcea, R., Tenenbaum, J., & Schölkopf, B. (2022). When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in Neural Information Processing Systems, 35*, 28458–28473.

Jordan, J. S. (1982). The competitive allocation process is informationally efficient uniquely. *Journal of Economic Theory, 28*(1), 1–18.

Kahneman, D., & Riis, J. (2005). Living, and thinking about it: Two perspectives on life. *The Science of Well-Being, 1*, 285–304.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica* 47 .

Kalai, E. & Smorodinsky, M. (1975). Other solutions to Nash's bargaining problem. *Econometrica: Journal of the Econometric Society*: 513–518 .

Kasenberg, D., Arnold, T., & Scheutz, M. (2018). Norms, rewards, and the intentional stance: Comparing machine learning approaches to ethical training. In *Proceedings of the 2018 AAAI/ACM conference on AI, Ethics, and Society* (pp. 184–190).

Kasirzadeh, A., & Gabriel, I. (2023). In conversation with artificial intelligence: Aligning language models with human values. *Philosophy and Technology, 36*(2), 27.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proceedings of the National Academy of Sciences, 113*(45), 12868–12873.

Kim, H., Sclar, M., Zhou, X., Bras, R., Kim, G., Choi, Y., & Sap, M. (2023). FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 14397–14413).

Kim, K., Garg, S., Shiragur, K., & Ermon, S. (2021). Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning* (pp. 5496–5505). PMLR.

Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J. B., & Rahwan, I. (2018). A computational model of commonsense moral decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 197–203).

Kirk, H. R., Vidgen B., Röttger, P., & Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence,* 1–10.

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition, 167*, 107–123.

Klingefjord, O., Lowe, R., & Edelman, J. (2024). What are human values, and how do we align AI to them? arXiv:2404.10636.

Knox, W. B., & Stone, P. (2011). Augmenting reinforcement learning with human feedback. In *ICML 2011 Workshop on New Developments in Imitation Learning (July 2011)* (Vol. 855, pp. 3).

Knox, W. B., Hatgis-Kessell, S., Adalgeirsson, S. O., Booth, S., Dragan, A., Stone, P., and Niekum, S. (2024a). Learning optimal advantage from preferences and mistaking it for reward. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, pp. 10066–10073).

Knox, W. B., Hatgis-Kessell, S., Booth, S., Niekum, S., Stone, P., & Allievi, A. G. (2024). Models of human preference for learning reward functions. *Transactions on Machine Learning Research*.

Korinek, A. and Balwit, A. (2022). Aligned with whom? Direct and social goals for AI systems. Technical report, National Bureau of Economic Research.

Korsgaard, C. M. (1989). Personal identity and the unity of agency: A kantian response to parfit. *Philosophy and Public Affairs*: 101–132.

Krakovna, V. & Kramar, J. (2023). Power-seeking can be probable and predictive for trained agents. arXiv:2304.06528.

Kwon, J., Levine, S., & Tenenbaum, J. B. (2023). Neuro-symbolic models of human moral judgment: LLMs as automatic feature extractors. *ICML 2023 Workshop on the Challenges of Deploying Generative AI*.

Kwon, J., Zhi-Xuan, T., Tenenbaum, J., & Levine, S. (2023). When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45).

Laibson, D. & Yariv, L. (2007). Safety in markets: An impossibility theorem for dutch books. Working papers 2007-5, Princeton University, Economics Department.

Laidlaw, C., & Dragan, A. (2022). The Boltzmann policy distribution: Accounting for systematic suboptimality in human models. In *International Conference on Learning Representations*.

Lam, M. S., Gordon, M. L., Metaxa, D., Hancock, J. T., Landay, J. A., & Bernstein, M. S. (2022). End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1–34.

Lambert, N. & Calandra, R. (2023). The alignment ceiling: Objective mismatch in reinforcement learning from human feedback. arXiv:2311.00168.

Lambert, N., Gilbert, T. K., & Zick, T. (2023). Entangled preferences: The history and risks of reinforcement learning and human feedback. arXiv:2310.13595.

Lazar, S. (2024). Legitimacy, authority, and democratic duties of explanation. *Oxford Studies in Political Philosophy* (Vol. 10, pp. 28).

Lazar, S., & Nelson, A. (2023). AI safety on whose terms? *Science, 381*(6654), 138–138.

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology, 19*(2), 107–115. https://doi.org/10.1007/s10676-017-9419-3

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: A research direction. arXiv:1811.07871.

Leshinskaya, A., San Franscisco, C., & Chakroff, A. (2023). Value as semantics: Representations of human moral and hedonic value in large language models. *NeurIPS 2023 Workshop: AI meets Moral Philosophy and Moral Psychology*.

Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. (2023). Resource-rational contractualism: A triple theory of moral cognition. *PsyArXiv*.

Levine, S., Kleiman-Weiner, M., Chater, N., Cushman, F., & Tenenbaum, J. B. (2024). When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment. *Cognition, 250*, 105790.

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences, 117*(42), 26158–26169.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science, 6*(2), 279–311.

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 661–670).

Lichtenstein, S., & Slovic, P. (2006). *The Construction of Preference*. Cambridge University Press.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences, 43*, e1.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review, 125*(1), 1.

Lin, J., Fried, D., Klein, D., & Dragan, A. (2022). Inferring rewards from language in context. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 8546–8560).

Liu, F. (2011). *Reasoning About Preference Dynamics* (Vol. 354). Springer.

Loewenstein, G., & Angner, E. (2003). *Predicting and Indulging Changing Preferences, Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice* (pp. 351–391). New York: Russell Sage Foundation.

Logins, A. (2022). *Normative Reasons: Between Reasoning and Explanation*. Cambridge University Press.

Lohr, S. (2023). AI is coming for lawyers, again. *The New York Times*.

London, A. J., & Heidari, H. (2024). Beneficent intelligence: a capability approach to modeling benefit, assistance, and associated moral failures through ai systems. *Minds and Machines, 34*(4), 41.

Luce, R. D. (1979). *Individual Choice Behavior: A Theoretical Analysis*. Westport: Greenwood Press.

Lukacs, G., & Livingstone, R. (1972). *History and Class Consciousness: Studies in Marxist Dialectics*. MIT Press.

Lumer, C., et al. (2005). Prioritarian welfare functions: An elaboration and justification.

Mahowald, K. (2023). A discerning several thousand judgments: GPT-3 rates the article+ adjective+ numeral+ noun construction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 265–273).

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

McInerney, J., Lacker, B., Hansen, S., Higley, K., Bouchard, H., Gruson, A., & Mehrotra, R. (2018). Explore, exploit, and explain: Personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, New York, NY, USA (pp. 31–39). Association for Computing Machinery.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74.

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press.

Merrill, W., Wu, Z., Naka, N., Kim, Y., & Linzen, T. (2024). Can you learn semantics through next-word prediction? The case of entailment. In *Findings of the Association for Computational Linguistics (ACL 2024)*. Association for Computational Linguistics.

Mill, J. (1859). *On Liberty*. Parker and Son: J. W.

Mishra, A. (2023). AI alignment and social choice: Fundamental limitations and policy implications. arXiv:2310.16048.

Mishra, S. (2014). Decision-making under risk: Integrating perspectives from biology, economics, and psychology. *Personality and Social Psychology Review, 18*(3), 280–307.

Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: Virtual bargaining underpins social interaction, culture, and society. *Trends in Cognitive Sciences, 18*(10), 512–519.

Modgil, S. (2009). Reasoning about preferences in argumentation frameworks. *Artificial Intelligence, 173*(9–10), 901–934.

Molinaro, G. & A. G. Collins. (2023). A goal-centric outlook on learning. *Trends in Cognitive Sciences*.

Momennejad, I., Hasanbeig, H., Vieira Frujeri, F., Sharma, H., Jojic, N., Palangi, H., Ness, R., & Larson, J. (2024). Evaluating cognitive maps and planning in large language models with CogEval. *Advances in Neural Information Processing Systems* 36 .

Moskovitz, T., Singh, A. K., Strouse, D., Sandholm, T., Salakhutdinov, R., Dragan, A. D., & McAleer, S. (2024). Confronting reward model overoptimization with constrained RLHF. In *The Twelfth International Conference on Learning Representations*.

Mount, K., & Reiter, S. (1974). The informational size of message spaces. *Journal of Economic Theory, 8*(2), 161–192.

Murphy, R. (2006). Cantor's diagonal argument: An extension to the socialist calculation debate. *The Quarterly Journal of Austrian Economics, 9*(2), 3–11.

Ng, Y. K. (1997). A case for happiness, cardinalism, and interpersonal comparability. *The Economic Journal, 107*(445), 1848–1858.

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 663–670).

Ng, R., & Subrahmanian, V. S. (1992). Probabilistic logic programming. *Information and Computation, 101*(2), 150–201.

Ng, A. Y., Harada, D., & Russell, S. J. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning* (pp. 278–287).

Ngo, R. (2019). Coherent behaviour in the real world is an incoherent concept. *AI Alignment Forum*.

Ngo, R., Chan, L., & Mindermann, S. (2022). The alignment problem from a deep learning perspective. arXiv:2209.00626.

Nielsen, K. & Rigotti, L. (2023). Revealed incomplete preferences. *Available at SSRN 4622145*.

Nussbaum, M. C. (2001). Symposium on Amartya Sen's philosophy: Adaptive preferences and women's options. *Economics and Philosophy, 17*(1), 67–88.

Okidegbe, N. (2021). Discredited data. *Cornell L. Rev., 107*, 2007.

Oldenburg, N. & Zhi-Xuan, T. (2024). Learning and sustaining shared normative systems via Bayesian rule induction in Markov games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*.

Omohundro, S. M. (2008a). Singularity Summit: The nature of self-improving artificial intelligence.

Omohundro, S. M. (2008b). The basic AI drives. In *AGI* (Vol. 171, pp. 483–492).

Ortega, P. A., & Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 469*(2153), 20120683. https://doi.org/10.1098/rspa.2012.0683

Oulasvirta, A., Jokinen, J. P. P., & Howes, A. (2022). Computational rationality as a theory of interaction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22 (pp. 1–14). New York: Association for Computing Machinery.

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. *Advances in Neural Information Processing Systems, 35*, 27730–27744.

Ovadya, A. (2023). Reimagining democracy for AI. *Journal of Democracy, 34*(4), 162–170.

Papadimitriou, C. H., & Tsitsiklis, J. N. (1987). The Complexity of Markov Decision Processes. *Mathematics of Operations Research, 12*(3), 441–450.

Parfit, D. (2011). *On What Matters* (Vol. 1). USA: Oxford University Press.

Parfit, D. (2018). Rationality and reasons. In *Exploring Practical Philosophy: From Action to Values* (pp. 17–39). Routledge.

Parisi, A., xZhao R., & Fiedel, N. (2022). TALM: Tool augmented language models. arXiv:2205.12255.

Paul, L. A. (2014). *Transformative Experience*. OUP Oxford.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature, 592*(7855), 590–595.

Petersen, S. (2023). Invulnerable incomplete preferences: A formal statement. *AI Alignment Forum*.

Pettigrew, R. (2019). *Choosing for Changing Selves*. Oxford University Press.

Piantadosi, S. T. & Hill, N. (2022). Meaning without reference in large language models. arXiv:2208.02957.

Piantadosi, S. T., & Jacobs, R. A. (2016). Four problems solved by the probabilistic language of thought. *Current Directions in Psychological Science, 25*(1), 54–59.

Pitis, S., Xiao, Z., Roux, N. L., & Sordoni, A. (2024). Improving context-aware preference modeling for language models. arXiv:2407.14916.

Prunkl, C., & Whittlestone, J. (2020). Beyond near-and long-term: Towards a clearer account of research priorities in AI ethics and society. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 138–143).

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences, 46*, e261.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 .

Rahwan, I., Ramchurn, S. D., Jennings, N. R., McBurney, P., Parsons, S., & Sonenberg, L. (2003). Argumentation-based negotiation. *The Knowledge Engineering Review, 18*(4), 343–375.

Railton, P. (1993). *Alienation, Consequentialism, and the Demands of Morality* (pp. 211–244). Ithaca: Cornell University Press.

Ramesh, R., Lubana, E. S., Khona, M., Dick, R. P., and Tanaka, H. (2024). Compositional capabilities of autoregressive transformers: A study on synthetic, interpretable tasks. In *Forty-First International Conference on Machine Learning*.
Rawls, J. (1971). *A Theory of Justice: Original Edition*. Harvard University Press.
Rawls, J. (1993). *Political Liberalism*. Columbia University Press.
Raz, J. (1999). *Engaging Reason: On the Theory of Value and Action*. Oxford University Press.
Reddy, S., A. Dragan, & Levine, S. (2018). Where do you think you're going?: Inferring beliefs about dynamics from behavior. *Advances in Neural Information Processing Systems* 31 .
Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online, 94*, 15.
Rubinstein, A., & Salant, Y. (2012). Eliciting welfare preferences from behavioural data sets. *The Review of Economic Studies, 79*(1), 375–387.
Russell, S. J. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. London: Allen Lane, an imprint of Penguin Books.
Russell, S. J., & Subramanian, D. (1994). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research, 2*, 575–609.
Rust, J. P. (1996). Dealing with the complexity of economic calculations. *Available at SSRN 40780*.
Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica, 5*(17), 61. https://doi.org/10.2307/2548836
Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., & Leike, J. (2022). Self-critiquing models for assisting human evaluators. arXiv:2206.05802.
Savage, L. J. (1972). *The Foundations of Statistics (2d rev* (ed). New York: Dover Publications.
Scanlon, T. (2000). *What We Owe to Each Other*. Cambridge, Massachusetts, London, England: The Belknap Press of Harvard University Press.
Schechtman, M. (2014). *Staying Alive: Personal Identity, Practical Concerns, and the Unity of a Life*. OUP Oxford.
Schroeder, T. (2004). *Three Faces of Desire*. Oxford University Press.
Scott, J. C. (1998). *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press.
Sen, A. (1970a). *Collective Choice and Social Welfare*. Harvard University Press.
Sen, A. (1970b). Interpersonal aggregation and partial comparability. *Econometrica: Journal of the Econometric Society,* 393–409 .
Sen, A. et al. (1999). Commodities and capabilities. *OUP Catalogue*.
Shah, R. (2018). Coherence arguments do not entail goal-directed behavior. *AI Alignment Forum*.
Shah, A., P. Kamath, J. A. Shah, & S. Li. (2018). Bayesian inference of temporal task specifications from demonstrations. *Advances in Neural Information Processing Systems* 31 .
Shah, R., Gundotra, N., Abbeel, P., & Dragan, A. (2019). On the feasibility of learning, rather than assuming, human biases for reward inference. In *International Conference on Machine Learning* (pp. 5670–5679). PMLR.
Siddarth, D. & Huang, S. (2023). Whitepaper. *The Collective Intelligence Project*.
Siddarth, D., Acemoglu, D., Allen, D., Crawford, K., Evans, J., Jordan, M., & Weyl, E. G. (2022). How AI Fails Us. *Technology and Democracy Discussion Paper Series*.
Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence, 299*, 103535. https://doi.org/10.1016/j.artint.2021.103535
Simon, H. A. (1957). A behavioral model of rational choice. In *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting* (pp. 241–260).
Simon, H. A. (1979). Rational decision making in business organizations. *The American Economic Review, 69*(4), 493–513.
Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics, 50*(3), 665–690.
Singh, S., Lewis, R. L., & Barto, A. G. (2009). Where do rewards come from. In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 2601–2606). Cognitive Science Society.
Sinhababu, N. (2017). *Humean Nature: How Desire Explains Action, Thought, and Feeling*. Oxford University Press.
Siththaranjan, A., Laidlaw, C., & Hadfield-Menell, D. (2024). Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *The Twelfth International Conference on Learning Representations*.

Skalse, J. M. V., Farrugia-Roberts, M., Russell, S., Abate, A., & Gleave, A. (2023). Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning* (pp. 32033–32058). PMLR.

Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015). Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Sobel, J. (2005). Interdependent preferences and reciprocity. *Journal of Economic Literature, 43*(2), 392–436.

Sorensen, T., Moore J., Fisher J., Gordon M., Mireshghallah N., Rytting C. M., et al. (2024). A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning* (pp. 46280–46302).

Stark, C. A. (1997). Decision procedures, standards of rightness and impartiality. *Nous, 31*(4), 478–495.

Stechly, K., Marquez, M., & Kambhampati, S. (2023). GPT-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. *NeurIPS 2023 Workshop on Foundation Models for Decision Making.*.

Steele, K., & Stefánsson, H. O. (2020). Decision theory, In Zalta, E. N (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University.

Steinhardt, J. (2017). Latent variables and model misspecification. *AI Alignment Forum*.

Stone, J., & Mittelstadt, B. (2024). Legitimate power, illegitimate automation: The problem of ignoring legitimacy in automated decision systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024)*.

Strotz, R. H. (1953). Cardinal utility. *The American Economic Review, 43*(2), 384–397.

Sumers, T., Yao, S., Narasimhan, K., and Griffiths, T. (2024). Cognitive architectures for language agents. *Transactions on Machine Learning Research*.

Suresh, H., Tseng, E., Young, M., Gray, M., Pierson, E., & Levy, K. (2024). Participation in the age of foundation models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1609–1621).

Sutton, R. S. & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Symons, J., & Alvarado, R. (2022). Epistemic injustice and data science technologies. *Synthese, 200*(2), 87.

Tan, Z.-X., & Ong, D. C. (2019). Bayesian inference of social norms as shared constraints on behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 41, pp. 2919–2925).

Taylor, J. (2016). *Quantilizers: A safer alternative to maximizers for limited optimization*. In AAAI Workshop: AI, Ethics, and Society.

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., et al. (2024). AI can help humans find common ground in democratic deliberation. *Science, 386*(6719), eadq2852.

Thorburn, L., Stray, J., & Bengani, P. (2022). What does it mean to give someone what they want? The nature of preferences in recommender systems. https://medium.com/p/82b5a1559157.

Thornley, E. (2023). There are no coherence theorems. *AI Alignment Forum.*

Thornley, E. (2024). The shutdown problem: An AI engineering puzzle for decision theorists. *Philosophical Studies,* 1–28.

Thornley, E., Roman, A., Ziakas, C., Ho, L., & Thomson, L. (2024). Towards shutdownable agents via stochastic choice. arXiv:2407.00805.

Toner, H. & McCauley, T. (2024). AI firms mustn't govern themselves, say ex-members of OpenAI's board. *The Economist*.

Turner, A., Ratzlaff, N., & Tadepalli, P. (2020). Avoiding side effects in complex environments. *Advances in Neural Information Processing Systems, 33*, 21406–21415.

Turner, A. M., Smith, L., Shah, R., Critch, A., & Tadepalli, P. (2021). Optimal policies tend to seek power. In *Proceedings of the 35th International Conference on Neural Information Processing Systems* (pp. 23063–23074).

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297–323.

Ulen, T. S. (1999). Rational choice theory in law and economics. *Encyclopedia of Law and Economics, 1*, 790–818.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems* 22 .

Valmeekam, K., Marquez, M., & Kambhampati, S. (2023). Can large language models really improve by self-critiquing their own plans? *NeurIPS 2023 Workshop on Foundation Models for Decision Making*.

Valmeekam, K., Marquez, M., Sreedharan, S., & Kambhampati, S. (2023). On the planning abilities of large language models—a critical investigation. In *Thirty-Seventh Conference on Neural Information Processing Systems*.

Vamplew, P., Smith, B. J., Källström, J., Ramos, G., Rădulescu, R., Roijers, D. M., Hayes, C. F., Heintz, F., Mannion, P., Libin, P. J. K., Dazeley, R., & Foale, C. (2022). Scalar reward is not enough: A response to Silver, Singh, Precup and Sutton (2021). *Autonomous Agents and Multi-Agent Systems, 36*(2), 41. https://doi.org/10.1007/s10458-022-09575-5

van de Meent, J. W., Paige, B., Yang, H., & Wood, F. (2018). An introduction to probabilistic programming. arXiv:1809.10756.

van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science, 32*(6), 939–984.

van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., & Rich, P. (2024). Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior*, 1–21.

van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics, 25*, 719–735.

Verdery, K. (2005). What was socialism, and why did it fall? *The Revolutions of 1989* (pp. 73–94). Routledge.

Vineberg, S. (2011). Dutch book arguments. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab: Stanford University.

von Mises, L. (1990). *Economic Calculation in the Socialist Commonwealth*. Ludwig Von Mises Institute: Auburn University.

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

von Widekind, S. (2008). *Evolution of Non-Expected Utility Preferences* (Vol. 606). Springer Science & Business Media.

von Wright, G. H. (1951). Deontic logic. *Mind, 60*(237), 1–15.

von Wright, G. H. (1972). The logic of preference reconsidered. *Theory and Decision, 3*, 140–169.

Warren, C., McGraw, A. P., & Van Boven, L. (2011). Values and preferences: Defining preference construction. *WIREs Cognitive Science, 2*(2), 193–205. https://doi.org/10.1002/wcs.98

Weber, M. (1978). *Economy and Society: An Outline of Interpretive Sociology*. University of California Press.

Weidinger, L., McKee, K. R., Everett, R., Huang, S., Zhu, T. O., Chadwick, M. J., Summerfield, C., & Gabriel, I. (2023). Using the veil of ignorance to align AI systems with principles of justice. *Proceedings of the National Academy of Sciences, 120*(18), e2213709120. https://doi.org/10.1073/pnas.2213709120

Wentworth, J. (2019). Why subagents? *AI Alignment Forum*.

Wentworth, J. (2023). Why not subagents? *AI Alignment Forum.*

Wheaton, D. (2023). Deceptive alignment is < 1% likely by default. *Less Wrong*.

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. arXiv:2306.12672.

Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N. A., Ostendorf, M., & Hajishirzi, H. (2024). Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems* 36.

Xu, T., Helenowski, E., Sankararaman, K. A., Jin, D., Peng, K., Han, E., et al. (2024). The perfect blend: Redefining RLHF with mixture of judges. arXiv:2409.20370.

Yang, S., & Allenby, G. M. (2003). Modeling interdependent consumer preferences. *Journal of Marketing Research, 40*(3), 282–294.

Yao, J., Yi, X., Wang, X., Wang, J., & Xie, X. (2023). From instructions to intrinsic human values—a survey of alignment goals for big models. arXiv:2308.12014.

Yudkowsky, E. (2004). Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*.

Yudkowsky, E. (2015). Known-algorithm non-self-improving agent. *Arbital*.

Yudkowsky, E. (2016). The AI alignment problem: Why it is hard, and where to start. https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/.

Yudkowsky, E. (2019). Coherent decisions imply consistent utilities. *Less Wrong*.

Zhi-Xuan, T. (2022). What Should AI Owe To Us? Accountable and Aligned AI Systems via Contractualist AI Alignment. *AI Alignment Forum*.

Zhi-Xuan, T., Kang, G., Mansinghka, V., & Tenenbaum, J. B. (2024). Infinite ends from finite samples: Open-ended goal inference as top-down Bayesian filtering of bottom-up proposals. *Proceedings of the Annual Meeting of the Cognitive Science Society, 46*(46) .

Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., & Mansinghka, V. (2020). Online Bayesian goal inference for boundedly rational planning agents. *Advances in Neural Information Processing Systems, 33*, 19238–19250.

Zhi-Xuan, T., Ying, L., Mansinghka, V., & Tenenbaum, J. B. (2024). Pragmatic instruction following and goal assistance via cooperative language-guided inverse planning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (pp. 2094–2103).

Zhou, W., & Li, W. (2022). A hierarchical Bayesian approach to inverse reinforcement learning with symbolic reward machines. In *International Conference on Machine Learning* (pp. 27159–27178). PMLR.

Zhu, B., Jordan, M., & Jiao, J. (2023). Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning* (pp. 43037–43067). PMLR.

Zhuang, S. & Hadfield-Menell, D. (2020). Consequences of Misaligned AI. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 15763–15773).

Ziebart, B. D., Bagnell, J. A., & Dey, A. K. (2010). Modeling interaction via the principle of maximum causal entropy. In *Proceedings of the 27th International Conference on Machine Learning* (pp. 1255–1262).

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence—volume 3, AAAI'08* (pp. 1433–1438).