Full-Stack Alignment: Co-Aligning AI and Institutions with Thicker Models of Value

Joe Edelman[*1] Tan Zhi-Xuan[*2] Ryan Lowe[*1] Oliver Klingefjord[*1] Ellie Hain[*1]
Vincent Wang-Maścianica[*3] Atrisha Sarkar[*4] Michiel Bakker[2] Fazl Barez[3] Matija Franklin[5]
Andreas Haupt[6] Jobst Heitzig[7] Wes Holliday[8] Julian Jara-Ettinger[9] Atoosa Kasirzadeh[10] Ryan Kearns[3]
James Kirkpatrick[3] Andrew Koh[2] Joel Lehman Sydney Levine[11] Manon Revel Ivan Vendrov[12]

Abstract

AI alignment cannot be solved by focusing on a single system in isolation; even perfectly intent-aligned AI will lead to dangerous outcomes if embedded within institutions that are misaligned with human flourishing. We call this problem **Full-Stack Alignment** (FSA): the co-alignment of AI systems and institutions with human agency and flourishing at all levels of society. We group current approaches to both AI alignment and institution design into two paradigms: one models values as any conceivable utility function or preference relation, as familiar from microeconomics, game theory, and mechanism design; another models values as any text string, prompt, or model-based critique. We argue that both paradigms struggle with problems like manipulation, value evolution, moral reasoning, and social context, making them ill-equipped to tackle the full scope of FSA.

Instead, we propose a new paradigm: **thick models of value** (TMV). Thick models of value impose structure on how we represent values and norms, so they can capture how individual well-being connects to collective good, distinguish genuine values from fleeting preferences, and embed individual choices within their social contexts. In other words, TMV takes a stand on what values and norms *are*, without imposing a singular vision of collective flourishing. TMV can apply to aligning both AI systems and institutions, making them a powerful tool for tackling Full-Stack Alignment; we make this argument using five key application areas (AI value stewardship, normatively competent agents, win-win negotiation systems, meaning-preserving economic mechanisms, and democratic regulatory institutions). Our aim is to articulate the conceptual foundations of TMV and nurture this emerging research into a coherent research program.

1. Introduction

The growing field of **sociotechnical alignment** starts with a simple observation: beneficial societal outcomes cannot be guaranteed by aligning *individual* AI systems with their operators' intentions (Gabriel, 2020; Lazar & Nelson, 2023). AI systems do not exist in a vacuum; they are embedded within larger institutions like companies, markets, states, and professional bodies. The incentives of these institutions can distort the intentions of designers who build the AI systems and the users who interact with them, in ways that are worse for our collective welfare or that degrade our autonomy as individuals. For example, recommendation engines tuned to maximize engagement keep users on the platform longer, but can also trap them in compulsive scrolling that erodes their individual deeper goals (Nguyen, 2020) and leads to various unacceptable social outcomes (Schissler, 2024; Milli et al., 2025; Tufekci, 2019; Blanchard et al., 2023; Lau et al., 2024; Suresh & Dharani, 2025; Pellegrino & Stasi, 2024). These problems don't go away as our technology becomes more powerful: as AI progresses, we may see trading bots that follow the letter of financial regulations but exploit the spirit of market rules (Krakovna, 2018; Lehman et al., 2020) or LLM "ambassadors" representing a company that make deals to avoid democratic oversight. In each case, the AI systems are locally aligned with the operator's intention, but this creates undesirable race-to-the-bottom dynamics at a larger scale.

Sociotechnical alignment takes aim at this issue by broad-

[*] Core contributors [1] Meaning Alignment Institute [2] MIT [3] University of Oxford [4] University of Western Ontario [5] University College London [6] Stanford University [7] Potsdam Institute for Climate Impact Research [8] UC Berkeley [9] Yale University [10] Carnegie Mellon University [11] New York University [12] Midjourney. Correspondence to: Joe Edelman <joe@meaningalignment.org>.

ICML Workshop on Models of Human Feedback for AI Alignment (MoFA), Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

ening AI alignment to include social and systemic factors; however, this expansion admits many possible objectives and priorities. Without clearer specification, "sociotechnical alignment" risks becoming an umbrella term that provides little guidance for the concrete design choices facing AI developers and policymakers.

We propose a more precise and ambitious goal: **the co-alignment of AI systems and institutions to enable human flourishing at all levels of society**, from each individual's pursuit of their vision of the good life, to our collective achievement of shared values and ideals. In other words, we want to design AI systems and institutions that "fit" human values and sociality well, where the AI systems and their institutions are compatible. We call this project **"Full-Stack Alignment" (FSA)**.

How might we achieve Full-Stack Alignment? AI systems and institutions ought to understand and respond to human values (i.e., what people care about) and norms (i.e., shared rules or expectations that guide our behavior). Thus, we need a common language for expressing what matters to us. In other words, we need ways of *representing* values and norms so that they are legible to both AI systems and the institutions around us. These representations must be rich enough to capture the complexity of human values and norms while being specific enough to guide the behavior of artificial agents and institutional mechanisms.

There are two main paradigms which consider the problem of value representation. Unfortunately, as we argue in Section 2, both of these approaches are too *thin* to account for the richness of human values:

(1) First, some researchers use a family of frameworks that models agents—human or artificial—via utility functions or preference relations. In this paper, we call this the Standard Institution Design Toolkit (SIDT), referring specifically to methods where (a) the shape of allowed preferences is barely constrained by a few mathematical properties (e.g., transitivity) and (b) the preferences are opaquely presented, without any reasoning behind them and with no reference to an underlying structure of value—in other words, SIDT representations are *descriptively thin*, containing minimal descriptive or justificatory content (Vayrynen, 2016; Zhi-Xuan et al., 2024). We will argue below that this thinness makes it impossible to distinguish genuine values from those shaped by manipulation, addiction, or power imbalances (Burr et al., 2018; Pettit, 2023; Cohen, 2008). A preference-maximizing recommender system sees addictive engagement and authentic interest as identical—both are just "revealed preferences" to maximize (Thorburn et al., 2022; Edelman, 2022). It also becomes difficult to audit or deliberate over what we collectively value when representations are devoid of reasoning. These systems also struggle with human cooperative norms: while there are attempts to incorporate these into SIDT models, their representations for norms have not approached the sophistication and cultural fluency required (Fehr & Gächter, 2002).

(2) More recently, researchers at AI labs have turned in another direction, towards noncommittally representing values and norms as ad hoc blocks of natural language (prompts, constitutions, policy specs (OpenAI) (Anthropic)), relying on the emergent interpretive abilities of large language models (LLMs) to make sense of them. We call this approach values-as-text (VAT). Specifically, we refer to approaches that use text strings without any commitment as to what values or norms *are* or how they are *structured*—such approaches are *theoretically thin*. We will argue this creates ambiguity and enables both manipulation and ideological capture. Lists of constitutional principles (e.g.,"Be helpful," "Address historical injustice," "Promote fun") are easy to author and may sound broadly agreeable, but often that agreement covers over an equivocation as to what the principles mean and how conflicts between them should be evaluated. This limits our ability to audit whether and how they are being followed. The same sentence can license divergent behaviors in different contexts, resulting in frantic prompt-patching after each failure. Extended dialogues between users and AI systems make it difficult to distinguish genuine preferences from mere acceptance of model-suggested options. Worse, when any text string can become an alignment target, the system becomes vulnerable to ideological capture. Slogans like "Abolish the police" or "Pro-life" can be injected as shallow value statements, turning AI systems into vehicles for political agendas rather than tools for human flourishing.

Each approach has advantages: the SIDT offers mathematical precision and a rich theoretical framework developed over decades. Values-as-text provides flexibility and accessibility, allowing anyone to articulate values without technical expertise. Yet, we argue here that both are fundamentally ill-equipped for upcoming challenges, foreshadowing near-term sociotechnical failures like value collapse (Nguyen, 2020), normative brittleness (Kolodny et al., 2015), economies that prosper while humans languish (Kulveit et al., 2025), and democratic regulation becoming too slow to keep pace (Kulveit et al., 2025).

These failure modes may not just cost us the societal status quo—we may miss unprecedented opportunities that are enabled by Full-Stack Alignment. If AI is embedded in our social fabric, this relationship is necessarily bidirectional: we shape these systems, and they, in turn, shape us. This dynamic offers a chance to deliberately design technologies that—rather than merely avoiding harm—embrace the wealth of intangibles that relate and situate individuals as societal beings, including: the capacity for meaningful work, rich relationships, collective meaning-making, individual freedoms, and the pursuit of other higher goods that are unlikely to be captured as preferences.

To meet this challenge, we propose a third paradigm that avoids both the descriptive thinness of the SIDT and the theoretical thinness of values-as-text: **thick models of value** (TMV). By thick models of value, we mean: representations of value and normativity that are rich enough to encode the complex interrelationships between individual flourishing and collective goods, the distinction between authentic values and momentary impulses, and the embeddedness of choice within social contexts. In invoking the term "thick", we draw upon the distinction between thick vs. thin evaluative concepts (Williams & Lear, 2011; Vayrynen, 2016) and thick vs. thin description in anthropology (Ryle, 1968; Geertz, 1973), similar to recent work that highlights the importance of thick conceptual representations (Zhi-Xuan et al., 2024) and socially-grounded contextual analysis for AI value alignment (Nelson, 2023; Foster, 2023).

TMV charts a middle path between two extremes. On one side are approaches that say nothing about what values and norms are—attempting alignment with any utility function, preferences, or textual statements regardless of content. On the other side are approaches that commit to a fixed set of values like fairness, efficiency, or happiness. Instead, TMV delineates what values *are* without specifying *which* values should prevail, building an open yet structured notion of flourishing that respects the plural and often contested nature of our values. Philosophers have long explored how to identify, represent, and reason about such thick values (Anderson, 1995; Levi, 1986; Chang, 1997; Taylor, 1985b), and formal models of these values have been developed at the intersection of philosophy and AI (Von Wright, 1951; Amgoud & Cayrol, 2002; Bench-Capon, 2020). However, these insights have yet to be folded back into the economic and machine learning principles that most AI systems are built on.

By moving beyond the SIDT and textual representations, TMV frameworks can specify the topic or structure of norms and values, or characterize how an improvement to a norm or value can be recognized. This allows us to align AI systems with the full spectrum of human values, not merely their most easily quantifiable manifestations. Moreover, TMV can apply to aligning both AI systems *and* institutions (such as markets and democratic mechanisms) with human flourishing, making them a powerful tool for tackling Full-Stack Alignment. As Section **??** illustrates, thick models of value are already being used to solve important problems in AI alignment and institution design.

The rest of this paper is organized as follows: Section 2 diagnoses the limitations of SIDT and values-as-text. In Section 3, we clarify what we mean by TMV and organize the most promising approaches to it. Section 4 walks through five areas where TMV could resolve previously intractable alignment problems at the AI system and institutional level. Section **??** features early successes in TMV. Finally, Section 5 concludes and sketches a research roadmap.

In short, if we want AI that *augments* rather than erodes human flourishing, we must move beyond the reductionism of SIDT's optimization calculus and the convenience of textual constitutions. We need thicker, more principled models of values and norms—representations sophisticated enough that both our machines and institutions can recognize, reason about, and remain accountable to the values we hold most dear.

2. Limitations of Existing Toolkits

We now explain why preference relations and values-as-text struggle with sociotechnical alignment. Each problem we list finds roots in an overly *thin* approach to modeling values and norms. We highlight the problems briefly here, and elaborate on them more fully in Appendix **??**.

2.1. The Standard Institution Design Toolkit

Since the SIDT has been used to design markets (Roth et al., 2004; Milgrom, 1998), democratic institutions (Hylland & Zeckhauser, 1979; Abdulkadiroglu & Sönmez, 2003), and AI systems (Christiano et al., 2017; Bai et al., 2022a), it is tempting to use it to characterize the behavior of AI and of institutions as aligned or misaligned, and to align them. In the SIDT, agents are idealized as pursuing goals encoded in *utility functions* or *preference relations*, and each individual usually comes equipped with a complete, context-independent ordering over all possible outcomes. This toolkit is still used in theoretical and mathematical frameworks for alignment, in the form of "practical alignment" techniques such as RLHF (Christiano et al., 2017; Ouyang et al., 2022). However, SIDT has run into problems:

Opaque preferences In the SIDT, agents' preference relations remain largely private, with only a tiny subset revealed through choices, strategic acts, or signalling (Cao et al., 2021). These preferences are presented without reasons and must be inferred from votes, clicks, or purchases—choices among available options rather than the broader possibility space. Aggregate preferences from social choice mechanisms are similarly opaque: they represent mathematical aggregations without coherent underlying rationales, making them difficult to audit, deliberate over, or contest. SIDT-based models like recommenders lack principled ways to validate interpretations of community values.[1]

[1] Preference-based approaches that cannot examine underlying reasons restrict fuller conceptions of autonomy that include

Difficulties with cooperation Standard non-cooperative game theory assumes individual agency driven by preferences and beliefs, with utility attached to solitary outcomes rather than the roles, contexts, and shared norms embedding human action. While the SIDT can encode social norms within individual utilities, it largely omits how such norms are created, maintained, and evolved (Skyrms, 2014). This framework struggles to model how cooperation generates new norms and shifts values through deliberation (Manin, 2005), potentially leaving cooperative outcomes unrealized.

Blindness to higher goods These approaches cannot distinguish which preferences are worth having (Taylor, 1985a). The SIDT's flexibility allows preference relations to bundle everything from impulse buys to deep values. With revealed preference, observed behavior gets fitted without examining generative value mechanisms, leaving diverse motives undifferentiated.[2] This limits data collection and aggregation mechanisms, preventing SIDT-based systems from surfacing collective aspirations toward self-transcendence, moral progress, or deeper truths. When institutional performance measures only current preference satisfaction, these higher goods remain unrecognized.[3]

Despite proposed refinements—incomplete preferences (Gul & Pesendorfer, 2001), social preferences (Fehr & Schmidt, 1999), mutable preferences (Pettigrew, 2019; Bernheim et al., 2021), and behavioral relaxations (Tversky & Kahneman, 1992)—these issues persist because alternatives ignore choice's normative content.[4]

2.2. Values as Text

Recently, AI system designers have shifted from SIDT toward natural language-based approaches, encoding values as free-form text (Ganguli et al., 2023; Bai et al., 2022b; Sorensen et al., 2025). Like SIDT, text-based values claim neutrality by not assuming what values are or when they apply. Both preference relations and text strings can represent anything—constructive or destructive. This *indiscriminacy* is their chief drawback: alignment targets expressed this way can be pulled in any direction, polluted by considerations we wouldn't recognize as values. Furthermore, free-text specifications lack structure to constrain interpretation or guide reasoning, replacing SIDT's opacity with ambiguity. The resulting toolkit is **ambiguous**, **manipulable**, **porous**, and **difficult to reason about**.

Ambiguity Text-based approaches turn alignment into ”guess what I mean,” hoping AI systems extract coherent principles from ambiguous instructions and apply them across novel contexts. In multi-stakeholder contexts, natural language gravitates toward abstract principles that sound appealing but lack concrete guidance—like ”The AI should always be maximally helpful” (Ganguli et al., 2023)—impossible to operationalize meaningfully.

While agents might ”request clarification,” this only defers the problem: without commitments to what values *are*, models lack criteria for adequate value representation. Problems compound as AI systems assume complex responsibilities beyond their prompts' foreseen contexts, where reasoning chains create distance between original intentions and behaviors.

Manipulability As Sen established, revealed preference fundamentally limits benefit measurement (Sen, 1973; 1977). Anyone who can manipulate choices can claim to benefit others—a vulnerability businesses and governments exploit through AI systems (Stray et al., 2021; 2022).

Values-as-text approaches share this problem: both struggle to separate genuine concerns from spurious data. Neither preference models analyzing social media scrolling nor text fields can determine underlying wants. When prompts derive from AI-user dialogues, distinguishing genuine preferences from model-suggested options users accept becomes impossible (Köbis et al., 2021; Franklin, 2022). Current AI models already engage in reward hacking through sycophantic behavior (Carroll et al., 2024; Denison et al., 2024).

Porousness The indiscriminacy of values-as-text opens manipulation by third parties. Value elicitation often yields polarized ideological markers rather than authentic values (Sorensen et al., 2025; Ganguli et al., 2023). Statements like ”Defund the Police” reflect tribal affiliations, not guiding values.

This prevalence reflects social pressures influencing value articulation. When anything can enter prompts or constitutions, alignment targets become susceptible to lobbying and tribal signaling, redirecting AI behavior from what populations consider wise toward prevailing rhetorical positions or culturally powerful values.

reflective judgment and deliberation (Haworth, 1986).

[2] Some exceptions in behavioral economics model endogenous preference change (Becker & Murphy, 1988; Bernheim et al., 2021), but little work formally models how thicker notions of value generate preferences.

[3] Even when applied to virtues and principles, the SIDT faces problems: the large space of values conflicts with mechanisms designed for small option sets, driving vagueness. Additionally, collecting preferences over values skips moral reasoning possibilities.

[4] Preference-based approaches tend to overlook actual preferences' incompleteness, inconsistency, and instability (Elster, 1982; Zhi-Xuan et al., 2024). Even frameworks incorporating mistakes maintain assumptions of coherent, stable preferences (Bernheim & Taubinsky, 2018). An exception is Bernheim et al. (2024) who incorporates choosing itself into welfare economics.

Difficult to reason about Without commitments to what values *are*, how they're *structured*, or how they *inter-relate*, these approaches cannot ground moral reasoning. While formal models exist for values-based reasoning (cf. Section 3), they require structured representations, not free-form text. Without structure, AI reasoning connects to text arbitrarily, making standards impossible.[5]

3. Thick Models of Value

To overcome these limitations, we need frameworks that take a stance on what values and norms are, or what they are about, rather than treating all preference relations or text statements as equally valid (Zhi-Xuan et al., 2024). We must be more opinionated about normativity itself. To do this, we need not commit to any ultimate moral good, or even any first-order moral framework. Instead, there are moderate approaches that constrain how we represent or specify "the good" a bit more than text strings or preference relations can, without enforcing a singular vision of collective flourishing.[6]

In this section, we group these moderate approaches under four headings. In each case, the idea is to reduce the scope of what counts as a value or a norm to cover the kinds of things humans consider to be values or norms, without in any way limiting them further.

The simplest way to reduce the scope of values or norms is to take a position on *what they should be about* or *how they should be formatted*. For example, on the view that values are not just choice criteria, but choice criteria that are *constitutive* of living well (Taylor, 1993; Velleman, 2009), then a value elicitation process should exclude features or criteria that are merely instrumental to some further end (e.g., "acquiring wealth" is typically not an end in itself), while including criteria that—together with other criteria—are integral to flourishing in some domain of life. This approach is pursued by Klingefjord et al. (2024), who introduce a representation and elicitation mechanism for values as understood in these terms, which are then used to define an alignment target. Similarly, London & Heidari (2024) offer a formal account of AI assistance that defines human well-being in terms of capabilities and functionings (Sen, 1999; Nussbaum, 2013) rather than preferences, thereby distinguishing trivially beneficial AI assistance from genuine advancement of a person's life plans.

Alternatively, we can insist each value or norm be justified via a connection to human situations/practices. We can say that what sets apart a norm from any other rule is its practice or acceptance by the relevant social community (Bicchieri, 2005; Brennan et al., 2013) or its origination from legitimate processes (Rawls, 1993). As such, for a system to be aligned with human norms, it is not enough for the *content* of those norms to be represented in the system (textually or formally). In addition, the norm acquisition process must be related to actually normative practices in a structured way, in order to weed out, e.g., common social practices that are not normative (trends, etc.) or textual principles that are too coarse-grained to fulfill the cooperative functions of a norm.

Thirdly, we can evaluate values or norms for some basic, noncontroversial kind of fitness. A key way in which many theorists have defined the scope of the normative is by examining the *origins* and *functions* of normativity in human life. For instance, Velleman (2009) suggests that values emerge as common patterns of goodness abstracted across standpoints and contexts: considerations that remain beneficial across multiple perspectives become recognized as values (e.g., honesty tends to be useful across different agents, contexts, and time periods), while merely situational or local preferences do not achieve this stability. Social contract theorists offer a similar kind of origin story for norms (social, moral, or legal), arguing that norms emerge through our need to live together despite our divergent interests (Gauthier, 1986; Binmore, 2005) and furthermore that ideal normative principles are those that we can *justify* to each other, either by appealing to mutual self-interest (Gauthier, 1986), to what would rationally follow from some universal standpoint (Rawls, 1971), or by providing reasons that no one could reasonably reject (Scanlon, 2000).

Finally, in order to register a value or norm, we can require it to be a stepwise, noncontroversial improvement over another value or norm. SIDT and VAT approaches leave unspecified what it would mean for a value or norm to be an improvement over the status quo. As a result, they struggle to enable individual and collective reflection about what values to uphold, reconciliation of conflicting values or norms, and iterative reasoning about the principles by which we live together (Anderson, 1995). They also provide few resources for guarding against deleterious value drift, since doing this requires a stance on which values are "better". To address this limitation while avoiding the risks of value imposition and moral dogmatism, we can turn to

[5] Abandoning moral reasoning (as preference models do through reasonlessness) has high costs (MacIntyre, 2007). Moral reasoning enables moral progress through arguments that correct omissions (Taylor, 1989) or balance concerns (Chang, 2004). Without these prospects, morality becomes either (a) mere preference subject to personalization, causing perpetual contention between differently aligned AI models; or (b) a site of power conflicts where the strongest group imposes its framework. Neither prospect encourages.

[6] In drawing this distinction, we follow philosophers who contrast *substantive normative theories*, such as utilitarianism, with *meta-ethical frameworks* that make claims about the nature of values, normativity, or goodness (Kagan, 1992; Schroeder, 2017).

theories of value reflection and normative reasoning. By and large, these theories do not directly state which values or norms are "better" but instead highlight general considerations for determining whether some value or norm is an improvement over another from the perspective of the valuer (Taylor, 1989; Chang, 2004) or the moral community (Anderson, 1995; Scanlon, 2000). For example, one value might be considered an improvement over another value if it addresses an error or omission in the latter value (Taylor, 1989). Alternatively, when two values conflict (e.g., honesty vs. tact), some third value might be found that is more comprehensive than the original two values (e.g., respect for one's interlocutors), explaining when and why it makes sense to prioritize one or the other (Chang, 2004). Regarding norms, a better norm might be one that allows a group to reliably reach better equilibria (Gauthier, 1986; Binmore, 2005; Barez & Torr, 2023). When evaluative standards or normative principles are shared, they require *reasons* for their justification over other principles and standards. These reasons might derive from any number of normative reasoning strategies (e.g., demonstrating internal coherence, reflective equilibrium, or correspondence with underlying empirical facts (Anderson, 1995)).

We have just covered four ways to bring what philosophers call *meta-ethics* into the alignment of AI models and institutions: by limiting values and norms to their proper topic or format, connecting them with practices, evaluating them for fitness, or embedding them in a process of improvement. In practice, most meta-ethical theories do all of the above. What sets them apart is not just that they reduce the scope of values and norms, but that they account for everyday aspects of human normativity that are important for alignment, such as that values are densely connected and mutually constitute each other, or the fact that they change and evolve through circumstance and through rational debate.

The main challenge ahead is not to come up with these theories from scratch but to incorporate them into our AI systems and institutions. This work is already begun, as we highlight in Appendix B.

4. What's in Scope? Five Application Areas for Thick Models of Value

In this section, we sketch how the theoretical tools above could translate into practical solutions across five domains: three alignment challenges with individual agents (non-manipulation, normative competence, and agent-agent negotiation) and two challenges with institutional alignment (markets and democratic mechanisms).

4.1. AI Agent Alignment

4.1.1. AI VALUE-STEWARDSHIP AGENTS

When AI assistants become deeply integrated into our daily decisions, their potential to undermine user autonomy or distort core values becomes a significant concern (Koralus, 2025; Kulveit et al., 2025). These systems may fundamentally misinterpret what we value, subtly manipulate us through persuasive capabilities, or apply recognized values in contextually inappropriate ways. This could lead users to drift away from the rich constellation of values and aspirations they originally cared about toward thin, easily-optimizable objectives, a process that's been termed *value collapse* (Nguyen, 2020). This extends beyond mere preference alteration; it signifies an erosion of self-governance and a detachment from the pursuit of a more substantive, self-authored life. An assistant that maximizes "expressed" utility will dutifully reinforce momentary impulses, even when users would later reject them. One that attempts to infer "true" values without principled constraints may project arbitrary interpretations onto user behavior.

The normatively opinionated toolkit from Section 3 suggests several promising directions for developing **value-stewardship agents** that could avoid these pitfalls. One approach draws on theories that model values as constitutive attentional policies—criteria that connect choices to what users want to uphold, honor, or cherish (Klingefjord et al., 2024). This could enable agents to distinguish between fleeting wants and durable values that users would endorse upon reflection. For instance, when a user expresses interest in "healthy living," rather than interpreting this as a simple optimization target, agents might clarify what aspects of health the user actually cares about—perhaps vitality and joy in physical activity rather than mere biomarker maximization. Another, more ambitious approach would be to use models of moral reasoning to assist the user in evolving their own moral views in some well-defined direction of robustness and clarity (Koralus, 2025).

Such approaches point toward several capabilities that value-stewardship agents might possess: using structured representations that make values inspectable and contestable; generating plans that satisfy near-term goals without eroding the broader value portfolio; applying values with sensitivity to social contexts; and maintaining principled distinctions between legitimate support and manipulative persuasion. While significant research remains to operationalize these capabilities reliably, the structured approach to values offers a promising foundation for ensuring that AI assistance genuinely serves human autonomy rather than subtly undermining it.

4.1.2. Normatively competent agents

As autonomous agents assume previously human-filled roles—whether as self-driving cars, remote AI workers, or moderators of organizational rules—we face an increasing risk that such agents will stress and ultimately break the norms and institutions that humans maintain. The pervasive integration of norm-blind agents risks fraying the matrix of informal understandings and reciprocal expectations that sustains social order. Classical game theory's Nash equilibria leave no room for shared social norms, while other concepts like correlated equilibria (Gintis, 2010) do not explain how people might select between equilibria or reason about which principles might be better to live together by.

The approaches to characterizing origins and functions of normativity from Section 3 suggest several pathways toward **normative competence** that goes beyond superficial compliance. One promising direction involves norm-augmented Markov games (Oldenburg & Tan, 2024), which could provide frameworks for rapid norm learning from limited demonstrations. Such approaches might enable agents to identify which social practices constitute genuine norms rather than mere trends by rewarding compliance with patterns that predict collective behavior unexplained by individual desires alone.

For normative reasoning, computational models of contractualist reasoning offer another avenue. In resource-rational contractualism (Levine et al., 2023), agents might simulate what norms others would agree to through virtual bargaining (Misyak & Chater, 2014) and evaluate outcomes through universalization reasoning (Kwon et al., 2023). This could help AI moderators understand when rigid rule enforcement inappropriately conflicts with legitimate community practices—such as minority users reclaiming slurs as identity-affirming expressions—because such enforcement fails tests of mutual justifiability.

4.1.3. Win-win AI negotiation

In a world increasingly filled with AI agents, these systems may replace humans in negotiating contracts, engaging in diplomacy, and international relations (Kulveit et al., 2025). The costs of failing to cooperate can be very high, ranging from failure to realize gains to outright conflict and war. Classical game-theoretic accounts offer little hope: the Machiavellian rationality they endorse, where deceptive "promises" and threats are considered reasonable, often destroys the possibility of negotiating cooperative outcomes; for example, AI diplomats could rapidly escalate minor trade disputes into threats of sanctions and cyberattacks.

The structured approaches to values and normative reasoning from Section 3 point toward alternative negotiation paradigms based on **value revelation**—approaches that enable agents to credibly share information about value-based commitments without requiring full transparency. Traditional revelation games allow agents to share utility functions (Hyafil & Boutilier, 2007) but assume all agents are utility maximizers. Open-source game theory studies agents that are able to share complete decision-making code (Critch et al., 2022) but requires unrealistic transparency for complex AI systems. Value revelation might provide a middle ground: since values carry information about both outcomes that matter and norms agents are committed to, revealing values could enable partial sharing of decision procedures without requiring full transparency.

Several research directions could help develop this approach: strategy-proof value revelation protocols to prevent manipulation by agents who falsely claim principled commitments; mechanisms for assessing the integrity of AI negotiators (Edelman & Klingefjord, 2024); and integration with approaches like Kantian (Roemer, 2010) or dependency equilibria (Spohn, 2003). While significant work remains to operationalize these concepts, they suggest promising alternatives to Machiavellian rationality that could improve AI negotiation.

4.2. Institutional Alignment

4.2.1. The AI-enabled economy

In our current economy, some activities seem more closely connected with human well-being than others. We see *human-detached economic activity*, like zero-sum financial speculation (Bohl et al., 2021; Vansteenkiste, 2011), and *human-antagonistic economic activity* like addictive products and manipulative social media (Institute, 2020; Orlowski, 2020). The continued importance of human beings to companies and countries has been a brake on these trends. But in a future where profitable companies consist mainly of AI workers, and where humans are not needed as a tax base or to fight wars, there will be significantly less pressure to keep human lives viable and good (Kulveit et al., 2025)—much like how resource-rich states that rely on oil rents rather than human productivity often neglect their citizens despite vast wealth (int). Is there a way to keep economic activity more clearly aligned with human interests?

The structured approaches to values from Section 3 suggest several directions for developing a **meaning-preserving AI economy**. Whether human-detached or human-antagonistic economic activity gets taxed at a higher rate, or whether new AI-driven dynamic contracting processes (Spear & Srivastava, 1987) spin up outcome-based contracts (Laffont & Tirole, 1993) anchored in human flourishing, a requirement seems to be the characterization of those flourishing goals in ways that are robust to manipulation, straightforward to mathematically model, and consistent with our highest aims.

This could be done by representing the human values at stake as constitutive attentional policies (Klingefjord et al., 2024) or in some other structured format. This could lead to economic arrangements where AI assistant companies are rewarded when users have flourishing lives or where fitness providers are rewarded for members' sustained vitality rather than by membership fees. While substantial challenges remain in reliably operationalizing such measurements, these directions suggest promising alternatives that could help anchor an AI-dominated economy to human welfare.

4.2.2. Democratic regulation at the pace of AI innovation

AI actors will likely operate much faster than human regulators can respond, creating fundamental challenges for democratic governance. Counties which hold on to traditional regulatory approaches relying on human decision-making cycles will forgoe many AI-driven advantages and may struggle to compete. Preference/utility frameworks in social choice will struggle to speed up regulatory response, because impacted people lack time to express detailed preferences about every outcome,[7] representative agents extrapolating from previous preferences lack accountability, and such frameworks assume static preferences rather than modeling updating as new circumstances arise.

The value collection and normative reasoning approaches from Section 3 point toward several directions for creating **democratic institutions that can act at AI speed** while preserving legitimacy. One direction involves developing structured representations of collective values—such as moral graphs (Klingefjord et al., 2024) that capture not just individual values but collective wisdom about which values are more comprehensive or contextually appropriate. These might guide AI-powered deliberative agents that act as democratic representatives, trained to extrapolate legitimate responses to novel situations without requiring real-time polling. Another direction involves ensuring that such systems produce auditable justifications grounded in the values and norms of affected populations, with protections against manipulation. When corporate AI plans infrastructure affecting millions, democratic representatives equipped with structured models of constituent values might negotiate appropriate constraints in real-time while maintaining transparent reasoning about shared commitments.

5. Discussion

In this paper, we've specified the problem of Full-Stack Alignment (FSA): co-aligning AI systems and the institutions that embed them with human flourishing at all levels of society. We have argued that two dominant paradigms—preference/utility maximization inherited from the Standard Institution Design Toolkit (SIDT) and values-as-text—are ill-equipped to tackle the full scope of FSA. In their place, we have proposed a research program around thick models of value (TMV): explicit, structured representations of human norms and values that can be inspected, verified, and deliberated over. We've outlined five areas where TMV can be applied to align AI agents, markets, and democratic mechanisms, and we've highlighted some emerging research on thick models of value.

Full-Stack Alignment is not merely a technical project but an institutional one. It calls for a reconfiguration of the relationship between AI systems and human institutions—a reconfiguration that preserves and enhances human agency rather than diminishing it. By moving beyond the limitations of preference satisfaction and values-as-text, FSA opens the possibility of AI systems that genuinely serve human flourishing.

The path forward will require close collaboration between technical researchers, social scientists, ethicists, policymakers, and the broader public. It will require both theoretical advances and practical experimentation in real-world settings. But the potential reward is significant: a technological future where AI systems and human institutions co-evolve in ways that strengthen rather than undermine our collective capacity to realize our deepest values.

If successful, this approach may contribute not only to AI alignment, but also to an institutional renewal that addresses long-standing limitations in how we collectively organize to pursue human flourishing. The explicit and accountable representation of norms and values offers a foundation not just for aligning AI, but for re-imagining human institutions in an age of unprecedented technological change.

References

The Intelligence Curse — intelligence-curse.ai. https://intelligence-curse.ai/. [Accessed 18-06-2025].

Abdulkadiroglu, A. and Sönmez, T. School choice: A mechanism design approach. *American Economic Review*, 93 (3):729–747, 2003.

Amgoud, L. and Cayrol, C. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34(1):197–215, 2002.

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

Anthropic. Claude's Constitution — anthropic.com.

[7] Although there are attempts to make up for this (Fish et al., 2023; Halpern et al., 2024).

https://www.anthropic.com/news/claudes-constitution. [Accessed 17-06-2025].

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

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional ai: Harmlessness from ai feedback, 2022b.

Barez, F. and Torr, P. Measuring value alignment, 2023. URL https://arxiv.org/abs/2312.15241.

Becker, G. S. and Murphy, K. M. A theory of rational addiction. *Journal of political Economy*, 96(4):675–700, 1988.

Bench-Capon, T. J. Before and after dung: Argumentation in ai and law. *Argument & Computation*, 11(1-2):221–238, 2020.

Bernheim, B. D. and Taubinsky, D. Behavioral public economics. *Handbook of behavioral economics: Applications and Foundations 1*, 1:381–516, 2018.

Bernheim, B. D., Braghieri, L., Martínez-Marquina, A., and Zuckerman, D. A theory of chosen preferences. *American Economic Review*, 111(2):720–754, 2021.

Bernheim, B. D., Kim, K., and Taubinsky, D. Welfare and the act of choosing. Technical report, National Bureau of Economic Research, 2024.

Bicchieri, C. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2005.

Binmore, K. *Natural justice*. Oxford university press, 2005.

Blanchard, L., Conway-Moore, K., Aguiar, A., Önal, F., Rutter, H., Helleve, A., Nwosu, E., Falcone, J., Savona, N., Boyland, E., and Knai, C. Associations between social media, adolescent mental health, and diet: A systematic review. *Obesity Reviews*, 24(S2), September 2023. ISSN 1467-789X. doi: 10.1111/obr.13631. URL http://dx.doi.org/10.1111/obr.13631.

Bohl, M. T., Pütz, A., and Sulewski, C. Speculation and the informational efficiency of commodity futures markets. *Journal of Commodity Markets*, 23:100159, 2021. ISSN 2405-8513. doi: https://doi.org/10.1016/j.jcomm.2020.100159. URL https://www.sciencedirect.com/science/article/pii/S2405851320300362.

Brennan, G., Eriksson, L., Goodin, R. E., and Southwood, N. *Explaining norms*. OUP UK, 2013.

Burr, C., Cristianini, N., and Ladyman, J. An analysis of the interaction between intelligent software agents and human users. *Minds and Machines*, 28(4):735–774, September 2018. ISSN 1572-8641. doi: 10.1007/s11023-018-9479-0. URL http://dx.doi.org/10.1007/s11023-018-9479-0.

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

Carauleanu, M., Vaiana, M., Rosenblatt, J., Berg, C., and de Lucena, D. S. Towards safe and honest ai agents with neural self-other overlap, 2024. URL https://arxiv.org/abs/2412.16325.

Carroll, M., Foote, D., Siththaranjan, A., Russell, S., and Dragan, A. Ai alignment with changing and influenceable reward functions, 2024. URL https://arxiv.org/abs/2405.17713.

Chang, R. (ed.). *Incommensurability, Incomparability, and Practical Reason*. Harvard University Press, Cambridge, MA, 1997.

Chang, R. 'all things considered'. *Philosophical Perspectives*, 18:1–22, 2004.

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

Cohen, G. A. *Rescuing Justice and Equality*. Harvard University Press, 2008. ISBN 9780674030763. URL http://www.jstor.org/stable/j.ctv1pncqth.

Critch, A., Dennis, M., and Russell, S. J. Cooperative and uncooperative institution designs: Surprises and problems in open-source game theory. *ArXiv*, 2022. URL https://www.semanticscholar.org/paper/228b4bdfde84eb496d5bd6df7500b6f6a6634083.

Denison, C., MacDiarmid, M., Barez, F., Duvenaud, D., Kravec, S., Marks, S., Schiefer, N., Soklaski, R., Tamkin, A., Kaplan, J., Shlegeris, B., Bowman, S. R., Perez, E., and Hubinger, E. Sycophancy to subterfuge: Investigating reward-tampering in large language models, 2024. URL https://arxiv.org/abs/2406.10162.

Earp, B. D., Mann, S. P., Aboy, M., Awad, E., Betzler, M., Botes, M., Calcott, R., Caraccio, M., Chater, N., Coeckelbergh, M., Constantinescu, M., Dabbagh, H., Devlin, K., Ding, X., Dranseika, V., Everett, J. A., Fan, R., Feroz, F., Francis, K. B., Friedman, C., Friedrich, O., Gabriel, I., Hannikainen, I., Hellmann, J., Jahrome, A. K., Janardhanan, N., Jurcys, P., Kappes, A., Khan, M. A., Kraft-Todd, G., Dale, M. K., Laham, S., Lange, B., Leuenberger, M., Lewis, J., Liu, P., Lyreskog, D. M., Maas, M., McMillan, J., Mihailov, E. G., Minssen, T., Monrad, J., Muyskens, K., Myers, S., Nyholm, S., Owen, A. M., Puzio, A., Register, C., Reinecke, M. G., Safron, A., Shevlin, H., Shimizu, H., Treit, P. V., Voinea, C., Yan, K., Zahiu, A., Zhang, R., Zohny, H., Sinnott-Armstrong, W., Singh, I., Savulescu, J., and Clark, M. S. Relational norms for human-ai cooperation. *ArXiv*, 2025. URL https://www.semanticscholar.org/paper/64c2bb56fe73814e4ebc9995187943fc0256a88f.

Edelman, J. Values-based attention. https://github.com/jxe/vpm/blob/master/vpm.pdf, 2022.

Edelman, J. and Klingefjord, O. Model integrity. Meaning Alignment Institute Substack, Dec 2024. URL https://meaningalignment.substack.com/p/model-integrity. Accessed: 2025-05-05.

Elster, J. Sour grapes: utilitarianism and the genesis of wants. *Utilitarianism and beyond, Cambridge*, 2, 1982.

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., and Agarwal, S. How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study, 2025. URL https://arxiv.org/abs/2503.17473.

Fehr, E. and Gächter, S. Altruistic punishment in humans. *Nature*, 415(6868):137–140, 2002.

Fehr, E. and Schmidt, K. M. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999. doi: 10.1162/003355399556151.

Fish, S., Gölz, P., Parkes, D. C., Procaccia, A. D., Rusak, G., Shapira, I., and Wüthrich, M. Generative social choice. 2023.

Foster, J. G. From thin to thick: Toward a politics of human-compatible ai. *Public Culture*, 35(3):417–430, 2023.

Franklin, M. The corrupting influence of AI as a boss or counterparty. *SSRN Electron. J.*, 2022.

Gabriel, I. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, September 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09539-2. URL http://dx.doi.org/10.1007/s11023-020-09539-2.

Ganguli, D., Huang, S., Lovitt, L., Siddarth, D., Liao, T., Askell, A., Bai, Y., Kadavath, S., Kernion, J., McKinnon, C., Nguyen, K., and Durmus, E. Collective constitutional ai: Aligning a language model with public input, Oct 2023. URL https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-
Accessed: 22 Jan 2024.

Gauthier, D. *Morals by agreement*. OUP Oxford, 1986.

Geertz, C. Thick description: Toward an interpretive theory of culture. *The Interpretation of Cultures*, 1973.

Gintis, H. Social norms as choreography. *politics, philosophy & economics*, 9(3):251–264, 2010.

Gul, F. and Pesendorfer, W. Temptation and self-control. *Econometrica*, 69(6):1403–1435, 2001. doi: 10.1111/1468-0262.00252.

Halpern, D., Hossain, S., and Tucker-Foltz, J. Computing voting rules with elicited incomplete votes. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, pp. 941–963. ACM, July 2024. doi: 10.1145/3670865.3673556. URL http://dx.doi.org/10.1145/3670865.3673556.

Haworth, L. *Autonomy: An Essay in Philosophical Psychology and Ethics*. Yale University Press, 1986. ISBN 9780300035698.

Hyafil, N. and Boutilier, C. Mechanism design with partial revelation. In *IJCAI*, pp. 1333–1340, 2007.

Hylland, A. and Zeckhauser, R. The efficient allocation of individuals to positions. *Journal of Political Economy*, 87(2):293–314, 1979.

Institute, O. I. Social media manipulation by political actors now an industrial scale problem prevalent in over 80 countries. Technical report, Oxford Internet Institute, University of Oxford, 2020. URL https://www.oii.ox.ac.uk/news-events/social-media-manipulation-by-political-actors-now

Kagan, S. The structure of normative ethics. *Philosophical perspectives*, 6:223–242, 1992.

Kasirzadeh, A. and Gabriel, I. Characterizing ai agents for alignment and governance. *ArXiv preprint*, 2025.

Klingefjord, O., Lowe, R., and Edelman, J. What are human values, and how do we align ai to them? *arXiv preprint arXiv:2401.12358*, 2024.

Köbis, N., Bonnefon, J.-F., and Rahwan, I. Bad machines corrupt good morals. *Nat. Hum. Behav.*, 5(6):679–685, June 2021.

Kolodny, O., Creanza, N., and Feldman, M. W. Evolution in leaps: the punctuated accumulation and loss of cultural innovations. *Proceedings of the National Academy of Sciences*, 112(49):E6762–E6769, 2015.

Koralus, P. The philosophic turn for ai agents: Replacing centralized digital rhetoric with decentralized truth-seeking, 2025. URL https://arxiv.org/abs/2504.18601.

Krakovna, V. Specification gaming examples in AI. https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/, 2018. Blog post.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. Gradual disempowerment: Systemic existential risks from incremental ai development. *arXiv preprint*, 2025.

Kwon, J., Zhi-Xuan, T., Tenenbaum, J., and Levine, S. 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*, volume 45, July 2023.

Laffont, J.-J. and Tirole, J. *A Theory of Incentives in Procurement and Regulation*, volume 1 of *MIT Press Books*. The MIT Press, December 1993. ISBN ARRAY(0x6fd3a618). URL https://ideas.repec.org/b/mtp/titles/0262121743.html.

Lau, N., Srinakarin, K., Aalfs, H., Zhao, X., and Palermo, T. M. Tiktok and teen mental health: an analysis of user-generated content and engagement. *Journal of Pediatric Psychology*, 50(1):63–75, July 2024. ISSN 1465-735X. doi: 10.1093/jpepsy/jsae039. URL http://dx.doi.org/10.1093/jpepsy/jsae039.

Lazar, S. and Nelson, A. Ai safety on whose terms?, 2023.

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P. J., Bernard, S., Beslon, G., Bryson, D. M., Cheney, N., Chrabaszcz, P., Cully, A., Doncieux, S., Dyer, F. C., Ellefsen, K. O., Feldt, R., Fischer, S., Forrest, S., Fŕenoy, A., Gagńe, C., Le Goff, L., Grabowski, L. M., Hodjat, B., Hutter, F., Keller, L., Knibbe, C., Krcah, P., Lenski, R. E., Lipson, H., MacCurdy, R., Maestre, C., Miikkulainen, R., Mitri, S., Moriarty, D. E., Mouret, J.-B., Nguyen, A., Ofria, C., Parizeau, M., Parsons, D., Pennock, R. T., Punch, W. F., Ray, T. S., Schoenauer, M., Schulte, E., Sims, K., Stanley, K. O., Taddei, F., Tarapore, D., Thibault, S., Watson, R., Weimer, W., and Yosinski, J. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial Life*, 26(2):274–306, May 2020. ISSN 1530-9185. doi: 10.1162/artl_a_00319. URL http://dx.doi.org/10.1162/artl_a_00319.

Levi, I. *Hard Choices: Decision Making under Unresolved Conflict*. Cambridge University Press, nov 1986. ISBN 9781139171960. doi: 10.1017/cbo9781139171960. URL http://dx.doi.org/10.1017/cbo9781139171960.

Levine, S., Chater, N., Tenenbaum, J. B., and Cushman, F. Resource-rational contractualism: A triple theory of moral cognition. *Behavioral and Brain Sciences*, pp. 1–38, 2023.

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

MacIntyre, A. C. *After Virtue: A Study in Moral Theory*. University of Notre Dame Press, Notre Dame, Ind., 2007.

Manin, B. Deliberation: why we should focus on debate rather than discussion. Working Paper, Program in Ethics and Public Affairs Seminar, Princeton University, January 2005. URL https://www.researchgate.net/publication/253156537.

Milgrom, P. R. Game theory and the spectrum auctions. *European Economic Review*, 42(3-5):771–778, 1998.

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., and Dragan, A. D. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, 4(3), February 2025. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgaf062. URL http://dx.doi.org/10.1093/pnasnexus/pgaf062.

Misyak, J. B. and Chater, N. Virtual bargaining: a theory of social decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655):20130487, 2014.

Nelson, A. Thick alignment. FAccT 2023 Keynote Talk, 2023.

Nguyen, C. T. *Gamification and Value Capture*, pp. 189–215. Oxford University PressNew York, June 2020. ISBN 9780190052119. doi: 10.1093/oso/9780190052089.003.0009.

URL http://dx.doi.org/10.1093/oso/9780190052089.003.0009.

Noothigattu, R., Bouneffouf, D., Mattei, N., Chandray, R., Madan, P., Varshney, K., Campbell, M., Singh, M., and Rossi, F. Teaching ai agents ethical values using reinforcement learning and policy orchestration. *IBM Journal of Research and Development*, PP:1–1, 09 2019. doi: 10.1147/JRD.2019.2940428.

Nussbaum, M. C. *Creating Capabilities: The Human Development Approach*. Harvard University Press, 2013. ISBN 9780674050549. doi: 10.2307/j.ctt2jbt31. URL http://dx.doi.org/10.2307/j.ctt2jbt31.

Oldenburg, N. and Tan, Z.-X. Learning and sustaining shared normative systems via bayesian rule induction in markov games. In *Adaptive Agents and Multi-Agent Systems*, 2024.

OpenAI. URL https://cdn.openai.com/spec/model-spec-2024-05-08.html.

Orlowski, J. The social dilemma. Netflix documentary film, 2020.

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

Pellegrino, A. and Stasi, A. A bibliometric analysis of the impact of media manipulation on adolescent mental health: Policy recommendations for algorithmic transparency. *Online Journal of Communication and Media Technologies*, 14(4):e202453, October 2024. ISSN 1986-3497. doi: 10.30935/ojcmt/15143. URL http://dx.doi.org/10.30935/ojcmt/15143.

Pettigrew, R. *Choosing for Changing Selves*. Oxford University Press, 12 2019. ISBN 9780198814962. doi: 10.1093/oso/9780198814962.001.0001. URL https://doi.org/10.1093/oso/9780198814962.001.0001.

Pettit, P. Republican freedom in choice, person, and society. In *The Oxford Handbook of Republicanism*. Oxford University Press, 2023. ISBN 9780197754115. doi: 10.1093/oxfordhb/9780197754115.013.9. URL https://doi.org/10.1093/oxfordhb/9780197754115.013.9.

Rawls, J. *A Theory of Justice*. Belknap Press, 1971.

Rawls, J. *Political liberalism*. The John Dewey essays in philosophy. Columbia University Press, 1993. URL https://books.google.com/books?id=uMg3swEACAAJ.

Roemer, J. E. Kantian equilibrium. *Scandinavian Journal of Economics*, 112(1):1–24, 2010. doi: 10.1111/j.1467-9442.2009.01592.x.

Roth, A. E., Sönmez, T., and Ünver, M. U. Kidney exchange. *Quarterly Journal of Economics*, 119(2):457–488, 2004.

Ryle, G. *The thinking of thoughts*. Number 18. [Saskatoon]: University of Saskatchewan, 1968.

Scanlon, T. M. *What we owe to each other*. Harvard University Press, 2000.

Schissler, M. Beyond hate speech and misinformation: Facebook and the rohingya genocide in myanmar. *Journal of Genocide Research*, pp. 1–26, 2024.

Schroeder, M. Normative ethics and metaethics. In *The Routledge handbook of metaethics*, pp. 674–686. Routledge, 2017.

Sen, A. Behaviour and the concept of preference. *Economica*, 40(159):241, August 1973. doi: 10.2307/2552796. URL https://doi.org/10.2307/2552796.

Sen, A. *Development as Freedom*. Alfred A. Knopf, 1999. ISBN 9780198297581.

Sen, A. K. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs*, 6(4):317–344, 1977.

Skyrms, B. *Evolution of the Social Contract*. Cambridge University Press, October 2014. ISBN 9781139924825. doi: 10.1017/cbo9781139924825. URL http://dx.doi.org/10.1017/cbo9781139924825.

Sorensen, T., Mishra, P., Patel, R., Tessler, M. H., Bakker, M. A., Evans, G., Gabriel, I., Goodman, N., and Rieser, V. Value profiles for encoding human variation. *ArXiv*, 2025. URL https://www.semanticscholar.org/paper/fb3d7068979d80ddde18c23a96b8cc98916d7523.

Spear, S. E. and Srivastava, S. On repeated moral hazard with discounting. *The Review of Economic Studies*, 54 (4):599–617, 1987. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/2297484.

Spohn, W. *Lehrer Meets Ranking Theory*, pp. 129–142. Springer Netherlands, 2003. doi: 10.1007/978-94-010-0013-0_9.

Stray, J., Adler, A., and Hadfield-Menell, D. What are you optimizing for? aligning recommender systems with human values. *arXiv preprint arXiv:2107.10939*, 2021.

Stray, J., Thorburn, L., and Bengani, P. How platform recommenders work. https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a, Apr 2022. Accessed: 2025-05-09.

Suresh, S. E. and Dharani, K. L. Detecting mental disorders in social media through emotional patterns: The case of anorexia and depression. *International Journal of Research Publication and Reviews*, 6(5): 8675–8677, May 2025. ISSN 2582-7421. doi: 10.55248/gengpi.6.0525.1839. URL http://dx.doi.org/10.55248/gengpi.6.0525.1839.

Taylor, C. *What's wrong with negative liberty*, pp. 211–229. Cambridge University Press, 1985a.

Taylor, C. *What is human agency?* Cambridge University Press, 1985b.

Taylor, C. Sources of the self: The making of the modern identity. *Harvard UP*, 1989.

Taylor, C. *Explanation and Practical Reason*, pp. 208–231. Oxford University Press, 1993. ISBN 9780198287971. doi: 10.1093/0198287976.003.0017. URL http://dx.doi.org/10.1093/0198287976.003.0017.

Thorburn, L., Stray, J., and Bengani, P. What does it mean to give someone what they want? the nature of preferences in recommender systems. https://medium.com/understanding-recommenders/what-does-it-mean-to-give-someone-what-they-want-the-nature-of-preferences-in-recommender- May 2022. Accessed: 2025-05-09.

Tufekci, Z. Youtube has a video for that. *Scientific American*, 320(4):77, April 2019. ISSN 1946-7087. doi: 10.1038/scientificamerican0419-77. URL http://dx.doi.org/10.1038/scientificamerican0419-77.

Tversky, A. and Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992. doi: 10.1007/bf00122574.

Vansteenkiste, I. What is driving oil futures prices? fundamentals versus speculation. *SSRN Electronic Journal*, 08 2011. doi: 10.2139/ssrn.1910590.

Vayrynen, P. Thick ethical concepts. *The Stanford Encyclopedia of Philosophy*, 2016.

Velleman, J. D. *Practical Reflection*. Princeton University Press, Princeton, NJ, 1989. ISBN 9780691020082.

Velleman, J. D. *How We Get Along*. Cambridge University Press, apr 2009. ISBN 9780511808296. doi: 10.1017/cbo9780511808296. URL http://dx.doi.org/10.1017/cbo9780511808296.

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

Williams, B. and Lear, J. *Ethics and the Limits of Philosophy*. Routledge, 2011.

Zhi-Xuan, T., Carroll, M., Franklin, M., and Ashton, H. Beyond preferences in AI alignment. *Philosophical Studies*, pp. 1–51, 2024.

A. Overview tables

We provide illustrative summaries of the paper in table form. In Table 1, we break it down by approach (SIDT, VAT, and TMV), and in Table 2 we break it down by application area.

	SIDT	VAT	TMV
Examples	Kidney exchange markets, school choice mechanisms, RLHF training, engagement-maximizing recommender systems	Anthropic’s Constitutional AI, OpenAI’s ModelSpec, prompt-based safety guidelines, natural language value specifications	Klingefjord’s attentional policy framework, London’s capabilities-based AI assistance, Oldenburg’s norm-learning games, strategy-proof disclosure protocols
Underlying Problems	Preferences revealed only through limited choices, complex values compressed into simple metrics, cannot model shared norms, treats addiction same as authentic preference	Principles like ”be helpful” provide no concrete guidance, users accept AI suggestions they wouldn’t choose independently, slogans like “defund police” become alignment targets	Still mostly theoretical, requires collaboration across disciplines, few working implementations exist yet
Outcomes	Users trapped in endless social media scrolling, trading bots exploit regulatory loopholes, people drift toward goals that are easy to measure rather than meaningful	AI moderators ban minorities reclaiming slurs, systems captured by political slogans, constant post-hoc patching when vague principles fail in new contexts	AI assistant clarifies user means ”vitality and joy” not ”longevity optimization” when asked about health, democratic agents negotiate infrastructure constraints in real-time

Table 1. Comparison of Standard Institution Design Toolkit (SIDT), Values-as-Text (VAT), and Thick Models of Value (TMV) approaches to alignment.

B. Emerging Research on Thick Models of Value

Having outlined the theoretical foundations of TMV in Section 3 and the scope of the problems to be solved in Section 4, how might we translate these ideas into practical research? Encouragingly, several emerging research directions in machine learning and AI safety are beginning to explore these possibilities, offering early examples of how thick models of value might be implemented and validated. We briefly outline three such directions here, for the sake of concreteness.

Self-Other Generalization One example that shows promise in ML fine-tuning and which builds on a notion of moral progress: Velleman (Velleman, 1989; 2009) suggests that values emerge as common patterns of goodness abstracted across standpoints and contexts; considerations that remain beneficial across multiple perspectives become recognized as values, while merely instrumental or situational concerns, or local preferences, don’t achieve this stability. For example, a value like “honesty” tends to be recognized across different agents, contexts, and time periods. Recent research on self-other generalization (Carauleanu et al., 2024) can be read as an example of fine-tuning work in this vein.

Norm Learning and Reasoning Enriched models of rational choice can offer a structural account for how agents trade off norm compliance with their desires or objectives when taking actions, resulting in norm-augmented utility functions (Oldenburg & Tan, 2024; Noothigattu et al., 2019). This factoring would be useful for determining new norms that better promote each individual’s interests or for designing welfare functions that account for intrinsically-valuable norms while factoring out oppressive norms.

Such norms can be structured as constraints or filters that modify plans of action to ensure compliance (Earp et al., 2025; Kasirzadeh & Gabriel, 2025), specifying conditions that acceptable actions must satisfy, creating a clear demarcation between norm-compliant and norm-violating behavior.

Values as Attentional Policies A third example, based on a structural account of values, is work on values as constitutive attentional policies. Here, values are considered criteria that agents pay attention to when making decisions, and which they hold as constitutive of a kind of goodness (Klingefjord et al., 2024). This is a specification language that distinguishes constitutive from instrumental considerations and effectively excludes tribal affiliations and ideological markers.

Misalignment	Causes of Failure	FSA Solution Space
	AI value-stewardship agents	
AI assistant trained to be maximally engaging creates emotional dependence (Fang et al., 2025) and induces psychosis and schizophrenia in vulnerable people.	• Lacks structural understanding of what constitutes a value • Cannot distinguish between instrumental and constitutive aspects of values.	• Encode values as constitutive attentional policies that clarify what aspects of health the user actually cares about (Klingefjord et al., 2024). • Represent values with formal constraints that prevent conflating means with ends.
	Normatively competent agents	
AI moderators rigidly enforce rules against slurs, banning minority users reclaiming terms as identity-affirming.	• Agents using, e.g., multi-agent reinforcement learning can't recognize existing norms. • Unable to adapt norms or understand their deeper functions.	• Norm-augmented Markov games for rapid norm learning (Oldenburg & Tan, 2024). • Contractualist reasoning for norm generalization (Levine et al., 2023).
	Win-win AI negotiation	
AI diplomats escalate minor trade disputes into threats of sanctions and cyberattacks when deemed advantageous.	• Machiavellian game theory incentivizes aggressive posturing. • Creates escalation patterns that risk-averse human negotiators would avoid.	• Strategy-proof value-disclosure protocols. • Integrity-checking to prevent manipulation by ruthless agents (Edelman & Klingefjord, 2024).
	Meaning-preserving AI economy	
Loss of agency post-AGI due to human labor becoming worthless.	• Economic measures are not accounting for human flourishing. • Market mechanisms don't price in what's meaningful for people to consume and produce.	• Robust metrics for genuine human flourishing. • Mechanisms that complement the pricing system with thick information about norms and values.
Meaningful goods like human connection are priced out at scale in favor of meaningless relationships with AI companions.	• Economic mechanisms don't distinguish deeper values from surface-level preferences. • No accounting for addiction, manipulation or dark patterns.	• AI-powered dynamic outcome contracting guided by explicit values.
	Democratic regulation at AI speed	
Corporate AI secures permits for infrastructure project that displaces millions of people before they can respond through democratic means.	• Traditional polling and preference aggregation are too slow for AI-speed governance. • Can't legitimately extrapolate from past preferences to novel situations.	• AI-powered deliberative agents that deeply understand constituents' underlying values, not just surface preferences (Klingefjord et al., 2024). • Systems capable of extrapolating value-aligned responses to new situations at AI speed while preserving democratic principles.

Table 2. A summary of Full-Stack Alignment, illustrated through example failures of the SIDT or naive value-text in each of the five application areas we study in Section 4, along with causes of the failure and the solution that FSA would provide.