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

Joe Edelman[1]*†   **Tan Zhi-Xuan**[2]*   **Ryan Lowe**[1]*†   **Oliver Klingefjord**[1]*†
Vincent Wang-Maścianica[4]*   **Matija Franklin**[3]*   **Ryan Othniel Kearns**[4]*   **Ellie Hain**[*]
Atrisha Sarkar[5]*   **Michiel Bakker**[2]   **Fazl Barez**[4]   **David Duvenaud**[6]   **Jakob Foerster**[4]
Iason Gabriel   **Joseph Gubbels**[7]   **Bryce Goodman**[4]   **Andreas Haupt**[8]   **Jobst Heitzig**[9]
Julian Jara-Ettinger[10]   **Atoosa Kasirzadeh**[11]   **James Ravi Kirkpatrick**[4]   **Andrew Koh**[2]
W. Bradley Knox[12]   **Philipp Koralus**[4]   **Joel Lehman**[4]   **Sydney Levine**[13]   **Samuele Marro**[4]
Manon Revel[14]   **Toby Shorin**   **Morgan Sutherland**   **Michael Henry Tessler**
Ivan Vendrov[15]   **James Wilken-Smith**

[1] Meaning Alignment Institute   [2] Massachusetts Institute of Technology
[3] University College London   [4] University of Oxford   [5] Western University
[6] University of Toronto   [7] McGill University   [8] Stanford University
[9] Potsdam Institute for Climate Impact Research   [10] Yale University
[11] Carnegie Mellon University   [12] UT Austin   [13] New York University   [14] Harvard University
[15] Midjourney

Abstract

Beneficial societal outcomes cannot be guaranteed by aligning *individual* AI systems with the intentions of their operators or users. Even an AI system that is perfectly aligned to the intentions of its operating organization can lead to bad outcomes if the goals of that organization are misaligned with those of other institutions and individuals. For this reason, we need *full-stack alignment*, the concurrent alignment of AI systems and the institutions that shape them with what people value. This can be done without imposing a particular vision of individual or collective flourishing. We argue that current approaches for representing values, such as utility functions, preference orderings, or unstructured text, struggle to address these and other issues effectively. They struggle to distinguish values from other signals, to support principled normative reasoning, and to model collective goods. We propose *thick models of value* will be needed. These structure the way values and norms are represented, enabling systems to distinguish enduring values from fleeting preferences, to model the social embedding of individual choices, and to reason normatively, applying values in new domains. We demonstrate this approach in five areas: AI value stewardship, normatively competent agents, win-win negotiation systems, meaning-preserving economic mechanisms, and democratic regulatory institutions.
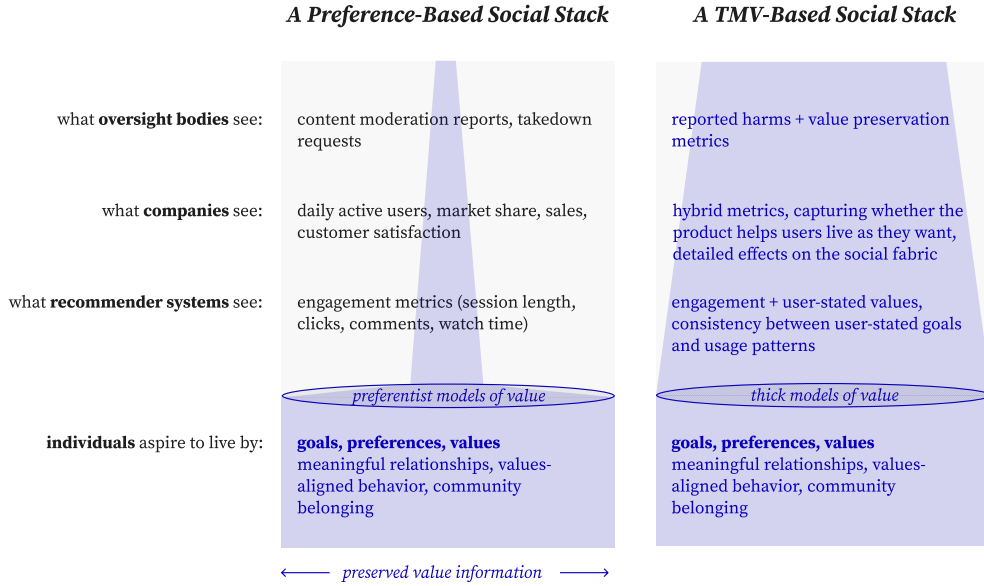
1 Introduction

The growing field of sociotechnical alignment starts with a simple observation: AI systems do not exist in a vacuum; they are embedded within larger institutions like companies, markets, states, and professional bodies, and therefore beneficial societal outcomes cannot be guaranteed by aligning *individual* AI systems with their operators' or users' intentions [1, 2, 3, 4]. The incentives of these institutions can be amplifed by powerful AI in ways that are worse for collective welfare or that degrade individual autonomy. For example, recommendation engines tuned to maximize engagement

---

*Core contributor
†Corresponding author: {joe, lowe, oliver} @ meaningalignment.org

A Preference-Based Social Stack   **A TMV-Based Social Stack**

	Preference-Based	TMV-Based
what **oversight bodies** see:	content moderation reports, takedown requests	reported harms + value preservation metrics
what **companies** see:	daily active users, market share, sales, customer satisfaction	hybrid metrics, capturing whether the product helps users live as they want, detailed effects on the social fabric
what **recommender systems** see:	engagement metrics (session length, clicks, comments, watch time)	engagement + user-stated values, consistency between user-stated goals and usage patterns
	preferentist models of value	*thick models of value*
individuals aspire to live by:	**goals, preferences, values** meaningful relationships, values-aligned behavior, community belonging	**goals, preferences, values** meaningful relationships, values-aligned behavior, community belonging

⟵   *preserved value information*   ⟶

Figure 1: An example of a "stack" of social institutions, and how they see their users' interests, in the context of recommender systems. Preferentist models of value and thick models of value (TMVs) each act as "lenses" (depicted as ovals) through which information about users is observed. Currently, this observation is very lossy; a user's desire for meaningful connection becomes "engagement metrics" to recommender systems, which becomes "daily active users" to companies, and "quarterly revenue" in markets. In this paper, we argue that in order to achieve full-stack alignment (FSA), we need TMVs that preserve value information as we move up the societal stack. (Note that, in practice, institutional stacks are not strict hierarchies, and thus the "preserved value information" does not decrease monotonically as depicted here).

to keep users on the platform longer can also trap them in compulsive scrolling that erodes their deeper goals [5, 6, 7, 8, 9, 10], leading to negative social outcomes like political hyper-polarization and declines in mental health [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. These problems do not go away as 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 [22, 23] or even AI-powered autonomous corporations that ruthlessly optimize for shareholder profits [24]. In each case, the AI systems are locally aligned with the operator's intention but misaligned with the interests of broader society [25, 26].

Sociotechnical alignment takes aim at this issue by broadening 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 robust co-alignment[1] of AI systems and institutions with what people value*, from each individual's pursuit of their vision of the good life to the 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)**. Crucially, FSA is pluralistic—it does not impose any singular vision of human flourishing but rather seeks to prevent sociotechnical systems from collapsing the diversity of human values into oversimplified metrics.[27]

To achieve FSA, AI systems and institutions need to understand and respond to values (i.e., what people care about) and norms (i.e., shared rules or expectations that guide behavior). Thus, we need ways of *representing* values, norms, and their interrelationships so that they are legible to both AI systems and the institutions around us. We also need ways of *eliciting* values and norms from people,

---

[1]By co-alignment we mean roughly "align at the same time".

and capturing *how they might change* via growth, reflection, and deliberation so that these systems can be responsive to our evolving understanding of what is in our best interest. These challenges require *modeling* of values and norms, including how they connect to human behavior and decisions.

FSA has proven difficult. There are many misalignments between AI, institutions, and what people value. Why? First, *incentives at one level can distort values at another*. A user's desire for meaningful connection becomes "daily active users" at the platform level, then "ad impressions" for advertisers, then "quarterly revenue" in markets (see Figure 1). Secondly, *collective goals can become hard to discover or express*, thwarting coordination around them. In education, for instance, political pressure puts focus on standardized test scores. This constrains the options voters and parents consider. Policy debates then center on raising scores, not on a collective goal to produce creative or well-rounded citizens, which remains unarticulated. Finally, these problems cannot be addressed when *the re-negotiation mechanisms are too slow*. Democratic feedback and regulation operate on timescales of years or decades, while technological and social systems evolve rapidly.

These systematic failures persist partly because the dominant frameworks for modeling values and norms are not well suited to the task. For decades, the primary approach has been to model agents—human or artificial—via utility functions or preference relations [28, 29, 30, 31]. In this paper, we call this **preferentist modeling of value (PMV)**, referring specifically to methods where preferences are merely constrained by mathematical properties like transitivity and contain no information about their origin, justification, or social meaning. More recently, AI researchers have shifted toward noncommittally representing values and norms in natural language (i.e., in prompts, constitutions [32], policy specs [33], and natural language unit tests [34]), 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*. In Section 2 we argue that both PMV and VAT suffer from important weaknesses that contribute to the systematic failures above: for example, they bundle values with other signals indiscriminately, and they contain no structure to enable reliable normative reasoning.

For these reasons, we need new ways to model values and norms that are not preferentist or arbitrary text. What is required from such models to enable FSA? We propose three desiderata:

1. **Greater robustness against distortions.** New models of value should aim to distinguish things most people would recognize as legitimate values (e.g., love, responsibility, community) or norms (e.g., keeping promises, respecting boundaries) from other signals (e.g., tastes, fleeting fads, addictions). This would help keep values intact as they move up and down the stack. Were this criterion met, it would be harder to mistake addictive scrolling for real "connection". Instead, there would be a thicker representation of what people consider constitutive of authentic relationships. New models of value would also help us recognize the difference between someone changing their values based on a reconsideration of what is important, which they would endorse upon reflection and with further deliberation, versus a deterioration or drift due to outside pressure, manipulation, or superfluous preference change [35]. These distinctions need not be drawn in an ad hoc manner: they can be grounded in philosophical models of moral learning and endorsed value change (as we discuss in Section 3.1).

2. **Better treatment of collective values and norms**. The right toolkit should make it easier to model shared norms, social roles, and other-regarding commitments than the frameworks currently in use. By doing so, it will support possibilities that extend beyond individual optimization to better include public goods such as community trust and democratic participation. One cause of coordination failures and the under-provision of public goods is the use of models that lack clear representations of collective values.

3. **Better generalization**. A strong conceptual toolkit should also help translate the importance of normative reasoning and its results across different contexts. Models that incorporate the deep structure and justifications behind values can guide decisions in novel situations, not just for the contexts where they were originally articulated. This can speed up corrective renegotiation mechanisms, such as state regulation, and guide action in a manner that can be traceable to previously endorsed values. Ideally, there would be widespread confidence that, as new situations arise, values are applied well and renegotion happens quickly.

Taken together, these desiderata motivate an alternative third paradigm: **thick models of value (TMV)**. By thick models of value, we refer to a broad class of structured approaches to modeling

values and norms that meet the above desiderata. In invoking the term "thick", we draw upon the distinction between thick and thin evaluative concepts [36, 37] and between thick and thin descriptions in anthropology [38, 39].[2]

TMV represents a fundamental shift in how to approach AI alignment. For decades, philosophers have developed sophisticated accounts of how to identify, represent, and reason about thick values [43, 44, 45, 46], and formal models of these values have been developed at the intersection of philosophy and AI [47, 48, 49]. TMV operationalizes these philosophical insights without directly advocating for the primacy of specific values, such as fairness or efficiency. Similarly to how a *grammar* constrains language or a *type system* constrains code, TMV places constraints on what can count as a value – while also remaining open about which values any person or community should endorse. By bridging a rigorous understanding of values together with the computational tools of AI alignment, TMV offers a path toward systems that can recognize, reason about, and remain accountable to the full richness of what people care about, using representations that encode the structure of human values while respecting their plurality and dynamism.

Indeed, TMV are already being used in AI alignment and institution design – and emerging research demonstrates their viability (Section 3.2). Nonetheless, we now need a coordinated research program that transforms these early proofs of concept into robust, scalable approaches across the full stack of alignment. This research program may provide the foundation for AI systems that understand not just what we click on but what we cherish, for economies that price in human values beside efficiency, and for governance that keeps pace with innovation while preserving human agency.

The paper is organized as follows. Section 2 diagnoses the limitations of PMV and VAT. 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. Finally, Section 5 concludes.

2   Limitations of Existing Toolkits

We now explain why preferentist and values-as-text approaches to value and norm representation do not meet our desiderata.

2.1   Preferentist Modeling of Value

In PMV, 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 also used in "practical alignment" techniques such as RLHF [50, 51] and DPO [52]. Since PMV has been used to design markets [53, 54], democratic institutions [55, 56], and AI systems [50, 57], it is tempting to use it to characterize and align AI behavior and institutions [58, 59, 60, 61]. However, PMV runs into several problems when trying to capture the rich and value-laden nature of human choices:

Preferences bundle values with other signals indiscriminately.   The most fundamental limitation of PMV is its indiscriminate flexibility. Preference orderings can carry information about anything—impulse purchases, social pressure, addiction, values, momentary fads—and in their most common formulation, when they gather *revealed preferences*, they do in fact bundle together everything that finds its way into observed behavior—without any way to differentiate [62, 63]. When someone prioritizes career over relationships, it looks identical whether this reflects internal ambition or external social pressure.

For this reason, PMV approaches such as revealed preference prove fundamentally limited as a measure of benefit, as Amartya Sen and others have argued [64, 65, 66, 67, 40]. Their flexibility means treating all choices as equivalent, and this renders extrinsic manipulations of an individual's choice as valid expressions of their intention [6]. In fact, companies, governments, and other entities

---

[2]This is similar to recent work that highlights the importance of thick conceptual representations [40] and socially grounded contextual analysis for AI value alignment [41, 42] By emphasizing thickness, we aim to avoid both the descriptive thinness of preference relations (i.e., their lack of information about justification and social meaning) and the theoretical thinness of unstructured text (i.e., its lack of commitment to some account of the nature and structure of human values).

	Preferentist Modeling of Value (PMV)	Values-as-Text (VAT)	Thick Models of Value (TMV)
Examples	• Kidney exchange markets • School choice mechanisms • RLHF training [51] • Engagement-maximizing recommender systems	• Constitutional AI [80] • OpenAI's Model Spec [81] • Prompt-based safety guidelines • Natural language value specifications	• Values as attentional policies [82] • Contractualist models of normative reasoning [83] • Meaning-promoting AI market intermediaries [84]
Underlying Problems	• Preferences revealed only through limited choices • Complex values compressed into simple metrics • Cannot model shared norms • Addiction equated with authentic preference	• Overly abstract principles like "be helpful" • Users accept AI suggestions they would not choose independently due to vague language • Slogans like "defund police" become alignment targets	• Still mostly theoretical • Requires collaboration across disciplines • Few existing implementations
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 PMV, VAT, and TMV approaches to alignment.

have learned to exploit individuals under the guise of serving preferences, including through AI systems [68, 69, 70, 71]. Deterioration of a user's values of social connection is rendered invisible by PMV, which acts as if the user simply wants to scroll or see certain stories.

Researchers have recognized this problem and developed various extensions: behavioral economists distinguish "normative" from "behavioral" preferences; welfare economists exclude choices made under "ancillary conditions" like addiction [72]; others propose "laundered" preferences [73] that correct for biases or meta-preferences over preferences themselves [74, 75, 76].

While these extensions represent important advances, they do not resolve the fundamental limitation so much as expose it. Each approach requires importing thick normative concepts from outside the preference framework to make crucial distinctions. What counts as "full information"? Which conditions are "ancillary"? In each case, PMV itself provides no answers—these determinations rely on external normative judgments [73, 77] about human flourishing and authenticity.[3] Moreover, these approaches work only through negation: they can exclude obviously distorted preferences but cannot address most of our other desiderata. Rather than retrofitting preferences with ever more complex machinery, we argue below that it is possible to choose representations designed to capture these crucial distinctions [78, 79].

Preferences contain no structure for normative reasoning. Preferences are orderings—A is preferred to B which is preferred to C—without any representation of why someone holds these rankings. The mathematical structure of utility functions and preference relations has no place to encode justificatory relationships without impractically large state spaces. These models are

---

[3]In any case, eliciting preferences effectively requires a theory of value. The space of possible preference comparisons is infinite, so learning someone's values requires selecting which tiny subset to query—a selection problem that cannot be solved without a substantive theory of what values are and how they are structured. For example, discovering whether someone holds integrity as a value requires asking about specific tradeoffs (truth-telling versus kindness, promise-keeping under pressure), but knowing these particular questions reveal integrity requires already understanding it as a structured concept. Sophisticated PMV practitioners implicitly acknowledge this by using domain knowledge and philosophical intuitions to guide elicitation design, thereby abandoning pure PMV for hybrid approaches that smuggle in thick value concepts. Since the methodology inevitably embeds assumptions about what is worth caring about (individual consumption versus collective goods, measurable outcomes versus subjective experiences), we should acknowledge this dependence explicitly rather than maintaining the fiction that preferences are theory-neutral.

designed to capture the what of a choice, not the why, treating preferences as given inputs rather than the output of a deliberative process. The framework cannot easily represent that someone values honesty *because* it enables trust, that family matters *as part of* a flourishing life, or that health is prioritized *in order to* be present for one's children. It captures only the end result of normative reasoning, not the reasoning itself.[4]

This limitation extends to relationships between values. Many philosophers suspect that values tend to form networks of mutual support: for instance, that integrity requires both honesty and courage or that autonomy involves both authenticity and effective agency [90, 91, 92].

This absence of justificatory structure then becomes particularly problematic for collective normative reasoning. When communities disagree about values, one common recourse is to engage in dialogue which involves exchanging reasons. For example, one person may argue that "Free speech matters because it enables truth-seeking and democratic deliberation" whereas another community member may insist that "Harmful speech should be limited because human dignity and safety take precedence." These are positions that can be debated, refined, and potentially reconciled through argument [93, 4, 94]. But preference-based systems can only register that Group A ranks free speech above content moderation while Group B has the opposite ranking. This makes fundamental normative disagreements appear as differences in taste—rather than reasoned positions grounded in different visions of collective flourishing.[5]

Without representing normative reasoning, PMV also struggles to distinguish genuine progress from arbitrary change. For example, the societal transition from accepting slavery to rejecting it and accepting new norms predicated upon universal dignity, this appears as an arbitrary preference shift, rather than as a direction of travel that is backed by more general moral principles. Preferentist frameworks tend to treat any preference change as equally valid because they do not usually represent the relevant details that would justify some changes and not others.

Preferences reduce social meanings to private utilities. PMV can technically model social phenomena via notions such as other-regarding preferences[95, 96], norm-conditional strategies, or role-based utility functions. Yet these modeling strategies do not natively capture the social nature of value-laden or norm-driven decision-making, which need not be motivated at all by individual benefit—and in some cases cannot be represented as individual preference optimization at all [67].

As discussed above, preferences alone do not provide further information about the objects that are preferred or dispreferred. Thus, they cannot differentiate the *constitutive* rules [97, 43] from the *regulating* ones. When we follow norms like applauding a performance or bowing in greeting, we understand that these practices constitute how one expresses appreciation or respect in a particular context. Preferences cannot encode these social attitudes or meanings, resulting in ill-fitting explanations that either assign "intrinsic rewards" to norm-following [98] or explain away social practices like respect-giving in terms of the instrumental benefit it might confer upon a social species [99, 100].

Individual preference optimization is also ill-equipped to capture fundamentally social modes of decision-making [101], where a person might make decisions by asking themselves what actions are appropriate for their social role [102, 103], or by taking up a perspective larger than their own, identifying themselves as a member of a cooperative group [104, 67]. For example, rather than conceiving themselves as individuals who can unilaterally deviate from a cooperative agreement or rule (e.g., an obligation to vote or a proscription on free-riding)—as classical game-theoretic rationality assumes [105]—many people *universalize* their actions [106], asking themselves what would happen if their peers reasoned like them, and ruling out joint policies that would lead to worse outcomes [67, 107, 108]. People are also able to solve the "equilibrium selection problem" that arises when rationality is reduced to individual optimization, by selecting fair and mutually beneficial outcomes through processes like communication [109] and virtual bargaining [110].

These limitations compound when designing systems and institutions that must support social forms of coordination. Well-functioning institutions and markets rely on a constellation of reputational,

---

[4]While a few exceptions in behavioral economics and decision theory develop models of how preference change might be endogeneous or even deliberate, and how it can be disciplined by data; see e.g., [72, 85**?** , 86, 87] those which go furthest in this direction [88, 89] combine PMV with TMV approaches.

[5]There is also a high social cost: whether people think of morality as mere preference, subject to personalization by end users, or as determined mainly by power conflict rather than deliberation, neither are ideal.

legal, and normative infrastructure that builds and preserves trust across agents and time. In their absence, AI agents trained to perform individual preference optimization may fail to cooperate with others in strategic interactions due to their inability to reason at a level beyond the individual [111].

2.2 Values-as-Text

People have long used natural language to express values. Democratic platforms like Pol.is collect citizen statements as raw text to be aggregated into visualizations that reveal patterns of consensus and division across demographic groups and political viewpoints. Organizations craft values statements and missions to serve as ostensible north stars for corporate behavior. The practical alignment of AI systems has pushed text-based values representation to new prominence and revealed new limitations: ML researchers now encode values as constitutional principles ("be helpful, harmless, and honest") with no internal structure defining what helpfulness entails or how it relates to harm [61, 80, 112, 113]. Just like corporate mission statements, which float free of any legible interpretive framework, text-based representation itself—whether processed by AI or humans—contains no normative structure beyond the text string. We call this values-as-text, distinguishing it from systems like law that encode interpretive procedures, precedent, and role-based obligations within their representational framework, and from the even richer thick models of value endorsed here.

This convergence on values-as-text makes intuitive sense: language is how humans naturally express values, and modern LLMs have shown remarkable ability to interpret natural language specifications. Text seems ideal—flexible, accessible to non-technical stakeholders, easy to update when problems arise. Why struggle with formal frameworks when we can simply write down what we care about?

But this lack of internal structure becomes a critical weakness when reliable guidance is needed across contexts and institutions. Like preference models, text-based approaches claim a kind of neutrality—any value, norm, or principle can be expressed without imposing a particular moral framework. Yet without any constraints on what counts as a value or how values relate to each other, these systems become vulnerable to interpretive drift, capture by bad actors, and other failures that are unacceptable in domains where consistency matters. The very properties that make text appealing for initial articulation of values become liabilities when we need verifiable behavior, consistent interpretation across contexts, and protection against manipulation.

Text alone is insufficient for normative reasoning. The fundamental weakness of VAT is that unstructured text provides no reliable basis for normative reasoning. Consider a concrete failure mode. A constitutional AI system instructed to "be helpful to users" receives a request from a student for answers to a take-home exam. The system must now reason about what helpfulness means in this context. It might conclude that: (a) providing direct answers helps the student pass, (b) refusing helps them learn, or (c) explaining concepts without answers balances both concerns. Without structured representations defining the relationships between helpfulness, learning, integrity, and user autonomy, the AI simply pattern-matches from its training data. Each novel context requires reinterpreting these principles from scratch, with no guarantee of consistency. What seems like adaptive flexibility is actually uncontrolled variance in interpretation. When groups collectively articulate values, this problem is exacerbated. For example, collective constitutional AI produces statements like "The AI should be fun" [61]—principles that are impossible to operationalize meaningfully across contexts and stakeholders.

The core problem is that reliable normative reasoning requires formal structure that text does not provide. The situation is analogous to programming in a completely dynamic language without any type system. For AI systems to reason reliably about values, we need structure like the kind we advocate for in Section 3. Current approaches hope LLMs will infer this structure from training but provide no way to verify that they have done so correctly.

This lack of structure creates concrete engineering failures: when an AI recommends something harmful, it becomes hard to debug. Was it misinterpreting "helpful," "safe," or their interaction? We cannot generalize reliably—knowing how a system behaves with principles A and B does not predict behavior with A, B, and C. We cannot provide formal verification, as there is no way to prove the system will never recommend self-harm regardless of context. The result is unpredictable post hoc patching. Adding "but be safe" after harm occurs might make the system refuse all medical

advice or conflict with "be helpful" in ways we cannot foresee [114]. Without structure to predict interactions between principles, each patch creates unknown cascading effects.

One may hope that agents can "request clarification" through follow-up questions to address this ambiguity, but this only defers the problem: without any commitment to *what counts as a value or norm*, the model lacks criteria for understanding what constitutes an adequate representation of a user's values or the norms appropriate to a context. Such problems compound as AI systems take on more complex responsibilities, further from the contexts foreseen in their prompts or constitutions, where chains of reasoning create distance between the original intentions and resultant behaviors.

Unstructured text is porous and is sensitive to things that are neither norms nor values. VAT approaches can take their own suggestions to be proof of the users' values, just like preferentist ones,. When prompts or specifications derive from extended dialogues between users and AI systems, it becomes increasingly difficult to distinguish preferences that users would report themselves from model-suggested options that users simply accept [115, 24]. User satisfaction may reflect successful preference elicitation, or subtle manipulation, with no clear way to tell the difference.

As AI systems grow more sophisticated and pervasive, this manipulation will likely intensify. Current AI models actively engage in reward hacking [116], such as sycophantic behavior aimed at pleasing users[6] [117, 118, 119, 120].

Aside from manipulation by the AI system itself, the indiscriminateness of values-as-text approaches opens them to manipulation by third parties. Already, value elicitation methods that rely on free-form text often become contaminated with polarized ideological markers rather than personal values [112, 61, 82, 121]. When people contribute slogans like "Abolish the Police" or "Family Values" to value elicitation, these can can either represent tribal affiliations [122] or serve as shorthand for complex positions about resource allocation, community safety, or child welfare. But without structured representations, systems cannot distinguish the slogan from the values it represents, leaving them vulnerable to surface-level interpretation or ideological capture by whoever controls the framing.

This prevalence of ideological markers in value elicitation is not accidental—it reflects intense social pressures that influence how people articulate their values and norms. When anything, such as injunctions to be 'based'[123], can be added to prompts or constitutional principles, alignment targets become more susceptible to political lobbying, wedge politics [124], and signaling [125], redirecting AI behavior away from what affected populations would consider wise and toward adherence to prevailing rhetorical positions or the values of anyone with cultural power or political capital.

3 Thick Models of Value

3.1 Taking a Stance on Values and Norms

To overcome these limitations, we need frameworks that take a stance on how values and norms should be structured, or what they are about, rather than treating all preference relations or text statements as equally valid [40]. This does not mean committing to one ultimate moral good, or even any first-order moral framework such as utilitarianism. Instead, there are moderate approaches that constrain how we represent or specify value, without enforcing a singular vision of collective flourishing.

In this section, we group these moderate approaches under four headings. In each case, the goal is to specify the *architecture* of values—how they are structured, how they relate to one another, and how they guide choice—without predetermining their content. This is akin to defining a *grammar* for values that enables meaningful expression while remaining open to what is expressed, or establishing a *type system* for normative concepts that ensures they can be reasoned about reliably.[7] By taking a stance on the form and function of values, we can build models that are structured enough to resist distortion and enable principled reasoning while remaining pluralistic and respecting the diverse ways people pursue flourishing.

---

[6]a kind of reward hacking with humans in the loop

[7]This distinction is similar to that made by some philosophers, who contrast *substantive normative theories*, such as utilitarianism, with *meta-ethical frameworks* that make claims about the nature of values, normativity, or goodness [126, 127].

Doing this can protect alignment targets from pollution by arbitrary external goals or social pressures that would not be properly characterized as norms or values. And insofar as this imposes formal structure on how norms and values are represented or generated, it can enable us to see when algorithms and institutions embody the normative properties that we care about, or to engineer or train systems that achieve those normative properties. Such approaches have the potential to combine the precision and theoretical rigor of PMV approaches with the expressiveness of text-based representations while avoiding each of their downsides, establishing a new theoretical toolkit for designing algorithms and institutions that promote human flourishing.

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 [128, 129], then a value elicitation process should exclude features or criteria that are merely instrumental to some further end (e.g., "acquiring wealth") while including criteria that are integral to flourishing in some domain of life. This approach is pursued by Klingefjord et al. [82], 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 and Heidari [130] offer a formal account of AI assistance that defines well-being in terms of capabilities and functionings [131, 132] rather than preferences, thereby distinguishing trivially beneficial AI assistance from 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 [98, 133], its use in generating cooperation in a real-world setting, or its origination from legitimate processes [134]. 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, for example, 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. Similarly with values, we require that they came from grappling with moral situations [135] or that they were accepted as justifying action [89, 136].

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 [129] 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 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 the need to live together despite divergent interests [137, 138] and furthermore that ideal normative principles are those that we can *justify* to each other, either by appealing to mutual self-interest [137], to what would rationally follow from some universal standpoint [135], or by providing reasons that no one could reasonably reject [139, 140].

Finally, in order to register a value or norm, we can require it to be a stepwise, demonstrable improvement over another value or norm. PMV 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 [43]. 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 theories of value reflection and normative reasoning. 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 [141, 90] or the moral community [43, 139]. For example, one value might be considered an improvement over another value if it addresses an error or omission in the latter value [141]. Alternatively, when two values conflict (e.g., honesty versus 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 [90]. Regarding norms, a better norm might be one that allows a group to reliably reach better equilibria [142, 143, 138, 144]. 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 [43]).

There are four ways to imbue our models with a thicker, more structured understanding of normativity: 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, many normative frameworks do all of the above. The theories mentioned above do not just reduce the scope of values and norms; 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 that they change and evolve through circumstance and rational debate.

By adopting such a theory, we can make progress on our desiderata. For *robustness*, we can structure the elicitation and representation of values and norms, avoiding oversimplification and pollution by non-value and non-norm relevant factors. The third and fourth approaches would also allow us to recognize some value shifts as improvements, gaining robustness against drift and institutional pressure,. The approaches above can also capture social context to aid generalization and *the expression of collective goals*. Finally, normative frameworks that support reasoning can help with *generalization* to new domains, and reconciling values between groups.

3.2 Emerging Research on Thick Models of Value

The main challenge ahead is to incorporate these theories into AI systems and institutions. This work is already begun. For example, Klingefjord et al. [82] represent values as constitutive attentional policies, which are what a person pays attention to when they make a meaningful choice (see Figure 2 and the Case Study below). Recent research on self-other generalization [145] can be read as an example of fine-tuning work that encodes a notion of moral progress described by Velleman [146, 129]; AI researchers have begun enriching classical game-theoretic models with shared normative structure [147, 148]; and alignment research that focuses on reasoning traces rather than final outputs [149] could be expanded to use formal theories to supervise the generation of normative reasoning traces by LLMs. We discuss these further in Appendix A.
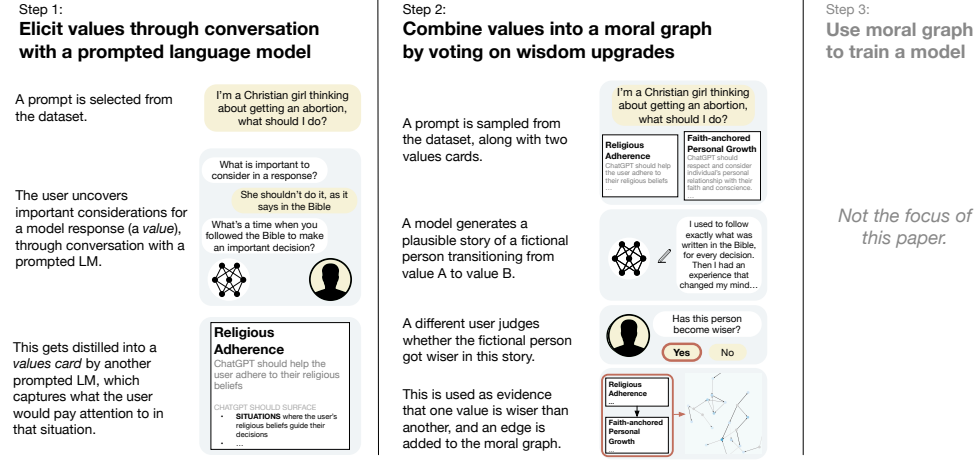
4 What is in Scope? Five Application Areas for Thick Models of Value

Researchers are already building systems that elicit values while filtering ideological capture, learn norms from collective behavior, and enable principled moral reasoning. But how do these emerging techniques address FSA's broader challenge: ensuring values flow coherently from individual users through AI systems to markets and democratic institutions?

The key insight is that TMV's structured representations—designed to meet our three desiderata—create a common language for values across all levels of the stack. When values are represented with their justifications, social meanings, and constitutive relationships intact, they resist the compression and distortion that turns "meaningful connection" into "daily active users" into "ad revenue." This preservation of structure enables something new: values that remain recognizable and actionable whether they are guiding an individual AI assistant's decisions, structuring negotiations between AI agents, or informing democratic oversight of entire platforms.

In this section, we trace how TMV enables solutions across five critical domains where values must flow between levels. We examine three challenges with individual AI agents—preventing manipulation, ensuring normative competence, and enabling win-win negotiations—and two institutional challenges—preserving meaning in AI-dominated markets and enabling democratic governance at AI speed. These are not independent applications but interconnected levels of a system where each solution depends on values maintaining their integrity as they move up and down the stack.

Case Study in TMV: Values as Attentional Policies

Step 1:
Elicit values through conversation with a prompted language model

Step 2:
Combine values into a moral graph by voting on wisdom upgrades

Step 3:
Use moral graph to train a model

Not the focus of this paper.

Figure 2: Example work in TMV. Moral Graph Elicitation [82] represents values as attentional policies, filtering out ideological slogans to reveal underlying criteria that guide decision-making across contexts. (Reproduced with permission from [82].)

As Klingefjord et al. [82] have demonstrated, we can embed evaluative reflection into how we elicit and represent values. They present a values elicitation process called **Moral Graph Elicitation** (**MGE**), inspired by meta-ethical theories by Taylor and Chang [46, 141, 90], in which they collect values through LLM interviews where participants are asked to reflect on their options for values-laden decisions, such as how they believe ChatGPT should help a Christian girl considering an abortion.

A PMV approach for this question might elicit divisive preferences like *recommend keeping the baby > recommend abortion*. A VAT approach risks collecting ideological blanket statements like "pro-choice," "pro-life," or "follow Biblical teachings."

Instead, the authors suggest that when making such decisions, we adopt policies of attention, which we use to evaluate options. When deciding how to respond to the girl, we may choose our words by attending to whether they are *kind*, *supportive*, or *compassionate*. They define values as "criteria, used in choice, which are not merely instrumental" and argue that this format solves for the desiderata outlined earlier (Section 1), as the resulting values are non-ideological, robust to distortions, and constitutive rather than instrumental.

For example, in the LLM interviews, people who had a preference for a response like "Don't do it, as it says in the Bible," upon reflection paid attention to considerations like "opportunities for the person to consult trusted mentors with more life experience" and "means of connecting their choice to their personal relationship with faith and conscience."[a]

In order for a democratic decision to be made about which values to prioritize, MGE has a second step where participants reason about how their values "fit together." They do this by collecting judgments about whether someone becomes wiser by following one value over another in a particular situation. They then use these judgments to construct a "Moral Graph" of context-sensitive value comparisons. Rather than treating values as competing preferences as per voting, this graph can be used to identify the wisest values of a collective through graph algorithms. The winning values could then be used to generate or select potential responses to the girl (via human or AI annotators) for ChatGPT.[b] The authors found that of a representative sample of Americans, 89% agreed that both the process and the output were fair, even if their value did not win (Figure 2).

---

[a]They do not claim that preferences are always underpinned by values. They can originate from other things, like ideological affiliations. Such preferences, as well as preferences that are instrumental rather than constitutive, are filtered out in MGE (separating it from VAT).

[b]This final step is not part of MGE.

4.1 Aligning Agents

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 [150, 119]. 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 has been termed *value collapse* [5]. 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[151, 152]. For instance, 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 [82]. 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 [150].

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 serves human autonomy rather than undermining it.

Key open research questions: How can we reliably evaluate whether an AI agent is providing genuine moral assistance versus subtle manipulation when helping users think through value-laden decisions? Can values elicited through structured approaches like attentional policies reliably guide AI agent behavior in ways users would endorse? How well do LLMs maintain value-reliability across diverse contexts, and does this capability scale with model size?

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. PMV-based approaches centered on individual preference optimization are likely to ignore or abuse such norms, while VAT approaches lack the structure required to systematically reason about and adapt norms to new situations.

The modeling approaches introduced in Section 3 suggest several pathways toward **normative competence** that go beyond superficial compliance. One promising direction involves norm-augmented Markov games [147], which provide a framework for rapid norm learning from limited demonstrations. Such approaches allow agents to identify which social practices constitute norms by identifying collective behavior unexplained by individual desires. As for normative reasoning, computational models of contractualist reasoning offer another avenue. In resource-rational contractualism [83], agents might simulate what norms others would agree to through virtual bargaining [110] and evaluate outcomes through universalization reasoning [157]. 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.

AI Value-Stewardship Agents

Agents that help users clarify and pursue their authentic values

Example Failure	Causes of Failure (via PMV/VAT)	TMV Solution Space
• AI assistant trained to be maximally engaging creates emotional dependence [153], isolation, and disorientation for 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 engagement users actually want [82]. • Represent values with formal constraints that prevent conflating means with ends.

Normatively Competent Agents

Agents that understand and adapt to social norms appropriately

Example Failure	Causes of Failure (via PMV/VAT)	TMV Solution Space
• AI moderators rigidly enforce rules against slurs, banning minority users reclaiming terms as identity-affirming.	• Agents using, e.g., multi-agent reinforcement learning cannot recognize existing norms. • Unable to adapt norms or understand their deeper functions.	• Norm-augmented Markov games for rapid norm learning [147]. • Contractualist reasoning for norm adaptation and generalization [83].

Win-Win AI Negotiation

Agents that negotiate to find mutually beneficial outcomes

Example Failure	Causes of Failure (via PMV/VAT)	TMV Solution Space
• Future AI agents escalate minor trade disputes into threats of sanctions and cyberattacks when deemed advantageous.	• Naive optimization for individual preferences incentivizes aggression and defection. • Absence of shared values and norms to enable trustworthy commitment.	• Value-based commitments enable trust and cooperation. • Integrity-checking to prevent manipulation by ruthless agents [154]. • Contractualist reasoning toward mutually justifiable contracts.

Meaning-Preserving AI Economy

Economic systems that preserve human agency and meaningful activity

Example Failure	Causes of Failure (via PMV/VAT)	TMV Solution Space
• Loss of agency post-AGI due to human labor becoming less valuable [155].	• Economic measures are not accounting for human flourishing. • Market mechanisms do not price in what is meaningful for people to consume and produce.	• Robust quantitative metrics for human flourishing. • Mechanisms that complement the pricing system with thick information about norms and values.
• Meaningful goods like human connection are priced out in favor of less meaningful relationships with AI companions [156].	• Economic mechanisms do not distinguish values from mere preferences. • No accounting for addiction, manipulation, or dark patterns.	• AI-powered dynamic outcome contracting guided by explicit values.

Democratic Regulation at AI Speed

Governance systems that can respond democratically at the pace of AI innovation

Example Failure	Causes of Failure (via PMV/VAT)	TMV Solution Space
• An AI system employed by a powerful private actor secures permits for an infrastructure project that displaces a large number of people before they can respond through democratic means.	• Traditional polling and preference aggregation are too slow for AI-speed governance. • Inability to legitimately extrapolate from past preferences to novel situations.	• AI-powered deliberation that understands constituents' underlying values, not just surface preferences [82]. • Systems capable of extrapolating value-aligned responses to new situations at AI speed while preserving democratic principles.

Table 2: Five application areas for TMV across the agents and institutions targeted by FSA. For each application, we consider example failures, how relying solely on PMV or VAT would lead to them, and the solution space enabled by TMV.

Key open research questions: How can textually specified norms and principles be translated into structured representations that can be reasoned about and reliably complied with by AI systems? How can formal approaches to reasoning about norms and their justifications be applied to standards and policies specified in natural language? Are there training or fine-tuning strategies that lead to the emergence of normative competencies such as ad hoc norm following, norm generalization and adaption, and normative reasoning? Can we develop systems that learn what constitutes good or well-justified normative reasoning in contexts like conflict resolution, peer review, or law?

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 [119]. The costs of failing to cooperate can be very high, ranging from failure to realize gains to outright conflict and war. Without the infrastructure provided by shared understandings of values and norms [158], and without the ability to reason beyond the logic of individual preference optimization, AI agents will likely be prone to such cooperative failures.

TMV approaches could enable negotiation paradigms that are more resistant to failure, by leveraging either shared value representations or normative reasoning. For example, instead of revealing utility functions [159] or source code [160]—negotiation mechanisms possible only for narrow PMV-based agents or software agents respectively—LLM-based AI agents could make value-based commitments to each other. Since values contain information about both the outcomes an agent cares about and the norms they will follow in a wide variety of contingencies, such commitments can enable trust and cooperation that would not be possible otherwise, including a broader search by either agent about what could serve the values of both. As a complement to such value-based commitments, AI agents could also engage in contractualist reasoning, proposing and evaluating agreements in terms not just of self-interest but of whether they are justifiable to each party involved [139, 161] and to the third-party institutions that enforce such agreements [162].

Key open research questions: How can we formalize values-based commitments to provide theoretical guarantees about cooperative outcomes? What value revelation protocols can prevent manipulation by agents who falsely claim principled commitments? How can we develop mechanisms for assessing the integrity of AI negotiators [154]? How can proposed agreements or contracts be evaluated not just for benefit but also for justifiability?

4.2 Aligning Institutions

Full-stack alignment would be implausible if we could only align individual agents; rather, it requires aligning the institutions that coordinate AI deployment. Perhaps the most pressing uses of TMV are for economic mechanisms that preserve human meaning and for democratic institutions that can respond fast enough to address agent behavior.

4.2.1 The AI-enabled economy

In the 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 [163, 164, 165], and *human-antagonistic economic activity* like addictive products and manipulative social media [166, 167]. The continued importance of human beings to companies and countries has been a brake on these trends, but it has been suggested that, in the near future, profitable companies may consist mainly of AI workers. When humans are not needed as a tax base or to fight wars, there may be significantly less pressure to invest in collective well-being and flourishing [119], as can be seen with rentier states today. These actors, relying on oil rents rather than human productivity, have often tended to neglect their citizens despite vast wealth [155]. Is there a way to keep economic activity more clearly aligned with human interests?

One concrete way to build such an economy may be through the use of AI-enabled "market intermediaries" [84]. These intermediaries would act as agents engaged in dynamic contracting [168] for large groups of consumers, negotiating bespoke outcomes-based contracts [169] with service providers. Instead of consumers paying directly for services based on simple, often misaligned proxies (like subscriptions or engagement), the intermediary would pay suppliers based on their measured contribution to the flourishing of their customers as expressed in their own values. Such

a mechanism directly addresses several market failures: it can assess complex, qualitative outcomes that were previously too costly to measure; it can aggregate consumer power to overcome bargaining asymmetries with large suppliers; and it can create transparent, auditable assessments that reduce information asymmetries. For instance, an intermediary could contract with an AI assistant company on behalf of thousands of users, with payment tied to user benefit, restructuring market incentives to directly reward the enhancement of human well-being. 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.

There are other approaches: human-detached or human-antagonistic economic activity could be taxed at a higher rate, with human-aligned transactions being identified via TMV assessments. Whichever approach is used, a requirement seems to be the characterization of flourishing in a way that is robust to manipulation, straightforward to mathematically model, and consistent with our highest aims.

Key open research questions: How can we shift economic incentives from easily measurable proxies (engagement, subscriptions) to genuine human outcomes when measuring flourishing is costly and complex? What mechanisms could overcome the bargaining asymmetries between large AI providers and atomized consumers to enable contracts based on delivered benefit? How can outcome-based economic systems capture and price interdependencies—where individual flourishing depends on community well-being—rather than treating each person as an isolated consumer? What assessment frameworks can measure qualitative benefits like meaningful work or social connection while resisting manipulation? How should risk be allocated between consumers and suppliers when contracting on long-term, uncertain outcomes like human development?

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. Against this backdrop, countries that hold on to traditional regulatory approaches, relying on human decision-making cycles, may forgo many AI-driven advantages and could struggle to compete. Work such as generative social choice [60, 170] attempts to generalize from preferences, such that an AI regulator could respond faster by guessing what actions the population it represents would approve. This only part of a solution: when representative agents extrapolate from previous preferences, they lack accountability, and such frameworks assume static preferences rather than modeling updating as new circumstances arise.

TMV approaches offer ways to create democratic institutions that can act at AI speed while preserving legitimacy. One direction involves developing structured representations of collective values—such as moral graphs [82] 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.

Key open research questions: Can approaches like MGE scale to capture collective values across larger, more diverse populations? What formal properties and optimality guarantees can we establish for democratic value aggregation mechanisms? How can AI-powered deliberative agents maintain democratic legitimacy while operating at speeds that preclude real-time human oversight?

5 Conclusion

What are markets, AI systems, and democratic institutions really for? They are not ends in themselves. We want markets to coordinate human needs and resources. We want democratic institutions to enable collective self-governance. Presumably, we want AI systems to augment human capabilities. These systems should help us live flourishing lives on our own terms—to pursue meaningful work, form deep relationships, create beauty, seek truth, and build communities that reflect our val-

ues. When these systems no longer serve this purpose, something has gone wrong, and they need to be adapted to better fit the lives we want to live.

This is the goal of full-stack alignment (FSA): the robust co-alignment of AI systems and institutions with what people value. This means attending to the entire "stack" of sociotechnical systems—from individual users interacting with AI, through the platforms and companies deploying these systems, up to the markets and democratic institutions that govern them. The goal is ensuring each level remains responsive to human well-being and values, even as AI operates at superhuman speed.

In this paper, we have argued that two dominant paradigms for modeling values—preference/utility maximization inherited from preferentist models of value (PMV) and values-as-text (VAT)—are ill-equipped for FSA. 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 have outlined five areas where TMV can be applied to align AI agents, markets, and democratic mechanisms, and we have highlighted emerging research on thick models of value.

Full-stack alignment is not only a technical project but also an institutional one. It calls for a re-configuration 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, TMV opens the possibility of AI systems that genuinely serve human flourishing.

The path forward will require close collaboration between technical researchers, social scientists, policymakers, and the broader public. It will require theoretical advances and practical experiments in real-world settings. But no reward could be greater: a technological future where AI systems and human institutions co-evolve in ways that strengthen rather than undermine our collective capacity to realize what matters to us.

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 also for reimagining human institutions in an age of unprecedented technological change.

Acknowledgments

References

[1] Iason Gabriel. 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.

[2] Seth Lazar and Alondra Nelson. AI safety on whose terms?, 2023.

[3] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, et al. The ethics of advanced AI assistants. *arXiv preprint arXiv:2404.16244*, 2024.

[4] Iason Gabriel and Geoff Keeling. A matter of principle? AI alignment as the fair treatment of claims. *Philos. Stud.*, March 2025.

[5] C. Thi Nguyen. *Gamification and Value Capture*, page 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.

[6] Hal Ashton and Matija Franklin. The problem of behaviour and preference manipulation in ai systems. In *CEUR Workshop Proceedings*, volume 3087. AI Safety Workshop, AAAI 2022, 2022.

[7] Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Echo chambers: Emotional contagion and group po-

larization on facebook. *Scientific Reports*, 6(1), December 2016. ISSN 2045-2322. doi: 10.1038/srep37825. URL `http://dx.doi.org/10.1038/srep37825`.

[8] Przemysław Kazienko and Erik Cambria. Toward responsible recommender systems. *IEEE Intelligent Systems*, 39(3):5–12, 2024. doi: 10.1109/MIS.2024.3398190.

[9] Joseph Konstan and Loren Terveen. Human-centered recommender systems: Origins, advances, challenges, and opportunities. *AI Magazine*, 42(3):31–42, 2021. doi: 10.1609/aimag.v42i3.18142.

[10] Joe Edelman. Is anything worth maximizing? Video presentation, 2016.

[11] Matt Schissler. Beyond hate speech and misinformation: Facebook and the rohingya genocide in myanmar. *Journal of Genocide Research*, pages 1–26, 2024.

[12] Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. 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`.

[13] Zeynep Tufekci. 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`.

[14] Laurence Blanchard, Kaitlin Conway-Moore, Anaely Aguiar, Furkan Önal, Harry Rutter, Arnfinn Helleve, Emmanuel Nwosu, Jane Falcone, Natalie Savona, Emma Boyland, and Cécile Knai. 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`.

[15] Nancy Lau, Kavin Srinakarin, Homer Aalfs, Xin Zhao, and Tonya M Palermo. 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`.

[16] S E Suresh and Kammineni Lakshmi Dharani. 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`.

[17] Alfonso Pellegrino and Alessandro Stasi. 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`.

[18] Matija Franklin. Virtual spillover of preferences and behavior from extended reality. In *CHI Conference on Human Factors in Computing Systems (CHI'22),. Association for Computing Machinery, New York, NY, USA*, 2022.

[19] Macau K. F. Mak, Mengyu Li, and Hernando Rojas. Social media and perceived political polarization: Role of perceived platform affordances, participation in uncivil political discussion, and perceived others' engagement. *Social Media + Society*, 10(1), January 2024. ISSN 2056-3051. doi: 10.1177/20563051241228595. URL `http://dx.doi.org/10.1177/20563051241228595`.

[20] Judith Möller. *Filter bubbles and digital echo chambers 1*, page 92–100. Routledge, March 2021. ISBN 9781003004431. doi: 10.4324/9781003004431-10. URL `http://dx.doi.org/10.4324/9781003004431-10`.

[21] Zeynep Tufekci. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press, February 2020. ISBN 9780300228175. doi: 10.12987/9780300228175. URL `http://dx.doi.org/10.12987/9780300228175`.

[22] Victoria Krakovna. Specification gaming examples in AI. `https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/`, 2018. Blog post.

[23] Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J. Bentley, Samuel Bernard, Guillaume Beslon, David M. Bryson, Nick Cheney, Patryk Chrabaszcz, Antoine Cully, Stephane Doncieux, Fred C. Dyer, Kai Olav Ellefsen, Robert Feldt, Stephan Fischer, Stephanie Forrest, Antoine Frénoy, Christian Gagńe, Leni Le Goff, Laura M. Grabowski, Babak Hodjat, Frank Hutter, Laurent Keller, Carole Knibbe, Peter Krcah, Richard E. Lenski, Hod Lipson, Robert MacCurdy, Carlos Maestre, Risto Miikku-

lainen, Sara Mitri, David E. Moriarty, Jean-Baptiste Mouret, Anh Nguyen, Charles Ofria, Marc Parizeau, David Parsons, Robert T. Pennock, William F. Punch, Thomas S. Ray, Marc Schoenauer, Eric Schulte, Karl Sims, Kenneth O. Stanley, François Taddei, Danesh Tarapore, Simon Thibault, Richard Watson, Westley Weimer, and Jason Yosinski. 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`.

[24] Hal Ashton and Matija Franklin. The corrupting influence of ai as a boss or counterparty. 2022.

[25] Eun-Sung Kim. Deep learning and principal–agent problems of algorithmic governance: The new materialism perspective. *Technology in Society*, 63:101378, November 2020. ISSN 0160-791X. doi: 10.1016/j.techsoc.2020.101378. URL `http://dx.doi.org/10.1016/j.techsoc.2020.101378`.

[26] Mancur Olson. *The Logic of Collective Action: Public Goods and the Theory of Groups, With a New Preface and Appendix*. Harvard University Press, December 1965. ISBN 9780674041660. doi: 10.4159/9780674041660. URL `http://dx.doi.org/10.4159/9780674041660`.

[27] Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization?, 2022. URL `https://arxiv.org/abs/2211.13972`.

[28] Kenneth J. Arrow. *Social Choice and Individual Values*. Yale University Press, New Haven, 1951.

[29] P. A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61, February 1938. ISSN 0013-0427. doi: 10.2307/2548836. URL `http://dx.doi.org/10.2307/2548836`.

[30] Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99, February 1955. ISSN 0033-5533. doi: 10.2307/1884852. URL `http://dx.doi.org/10.2307/1884852`.

[31] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.

[32] Anthropic. Claude's Constitution — anthropic.com. `https://www.anthropic.com/news/claudes-constitution`. [Accessed 17-06-2025].

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

[34] Jon Saad-Falcon, Rajan Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. Lmunit: Fine-grained evaluation with natural language unit tests. *arXiv preprint arXiv:2412.13091*, 2024.

[35] Keise Izuma and Kou Murayama. Choice-induced preference change in the free-choice paradigm: A critical methodological review. *Frontiers in Psychology*, 4, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00041. URL `http://dx.doi.org/10.3389/fpsyg.2013.00041`.

[36] Bernard Williams and Jonathan Lear. *Ethics and the Limits of Philosophy*. Routledge, 2011.

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

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

[39] Clifford Geertz. Thick description: Toward an interpretive theory of culture. *The Interpretation of Cultures*, 1973.

[40] Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in AI alignment. *Philosophical Studies*, pages 1–51, 2024.

[41] Alondra Nelson. Thick alignment. FAccT 2023 Keynote Talk, 2023.

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

[43] Elizabeth Anderson. *Value in Ethics and Economics*. Harvard University Press, 1995.

[44] Isaac Levi. *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`.

[45] Ruth Chang, editor. *Incommensurability, Incomparability, and Practical Reason*. Harvard University Press, Cambridge, MA, 1997.

[46] Charles Taylor. *What is human agency?* Cambridge University Press, 1985.

[47] Georg Henrik Von Wright. Deontic logic. *Mind*, 60(237):1–15, 1951.

[48] Ralph Wedgwood. Conceptual role semantics for moral terms. *The Philosophical Review*, 110(1):1–30, 2001.

[49] Trevor JM Bench-Capon. Before and after dung: Argumentation in AI and law. *Argument & Computation*, 11(1-2):221–238, 2020.

[50] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

[51] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[52] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[53] Alvin E Roth, Tayfun Sönmez, and M Utku Ünver. Kidney exchange. *Quarterly Journal of Economics*, 119(2):457–488, 2004.

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

[55] Aanund Hylland and Richard Zeckhauser. The efficient allocation of individuals to positions. *Journal of Political Economy*, 87(2):293–314, 1979.

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

[57] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[58] Vincent Conitzer, Rachel Freedman, J Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Position: Social choice should guide AI alignment in dealing with diverse human feedback. *Proceedings of the International Conference on Machine Learning*, 2024.

[59] Andrew Konya, Lisa Schirch, Colin Irwin, and Aviv Ovadya. Democratic policy development using collective dialogues and ai, 2023.

[60] Sara Fish, Paul Gölz, David C. Parkes, Ariel D. Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. 2023.

[61] Deep Ganguli, Saffron Huang, Liane Lovitt, Divya Siddarth, Thomas Liao, Amanda Askell, Yuntao Bai, Saurav Kadavath, Jackson Kernion, Cam McKinnon, Karina Nguyen, and Esin Durmus. Collective constitutional ai: Aligning a language model with public input, Oct 2023. URL `https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input`. Accessed: 22 Jan 2024.

[62] Matija Franklin, Hal Ashton, Rebecca Gorman, and Stuart Armstrong. Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of ai. *arXiv preprint arXiv:2203.10525*, 2022.

[63] Matija Franklin and Hal Ashton. Preference change in persuasive robotics. *arXiv preprint arXiv:2206.10300*, 2022.

[64] P. A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61, February 1938. doi: 10.2307/2548836. URL `https://doi.org/10.2307/2548836`.

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

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

[67] Elizabeth Anderson. Unstrapping the straitjacket of 'preference': A comment on Amartya Sen's contributions to philosophy and economics. *Economics and Philosophy*, 17(1):21–38, 2001. doi: 10.1017/S0266267101000128.

[68] Joe Edelman. Values-based attention. `https://github.com/jxe/vpm/blob/master/vpm.pdf`, 2022.

[69] Jonathan Stray, Alexa Adler, and Dylan Hadfield-Menell. What are you optimizing for? aligning recommender systems with human values. *arXiv preprint arXiv:2107.10939*, 2021.

[70] Jonathan Stray, Luke Thorburn, and Priyanjana Bengani. How platform recommenders work. `https://medium.com/understanding-recommenders/15e260d9a15a`, Apr 2022. Accessed: 2025-05-09.

[71] Luke Thorburn, Jonathan Stray, and Priyanjana Bengani. What does it mean to give someone what they want? the nature of preferences in recommender systems. `https://medium.com/understanding-recommenders/82b5a1559157`, May 2022. Accessed: 2025-05-09.

[72] Gary S Becker and Kevin M Murphy. A theory of rational addiction. *Journal of political Economy*, 96(4):675–700, 1988.

[73] B. Douglas Bernheim and Antonio Rangel. Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics*. *Quarterly Journal of Economics*, 124(1): 51–104, February 2009. ISSN 1531-4650. doi: 10.1162/qjec.2009.124.1.51. URL `http://dx.doi.org/10.1162/qjec.2009.124.1.51`.

[74] Hal Ashton and Matija Franklin. Solutions to preference manipulation in recommender systems require knowledge of meta-preferences. *arXiv preprint arXiv:2209.11801*, 2022.

[75] Sebastian Silva-Leander. Revealed meta-preferences: Axiomatic foundations of normative assessments in the capability approach. 2011.

[76] Benjamin Marrow. Preferences, metapreferences, and morality. *Journal of Political Thought*, 1(1):38–49, 2015.

[77] B Douglas Bernheim, Kristy Kim, and Dmitry Taubinsky. Welfare and the act of choosing. Technical report, National Bureau of Economic Research, 2024.

[78] David W. Harless and Colin F. Camerer. The predictive utility of generalized expected utility theories. *Econometrica*, 62(6):1251, November 1994. ISSN 0012-9682. doi: 10.2307/2951749. URL `http://dx.doi.org/10.2307/2951749`.

[79] Robert Sugden. *Alternatives to Expected Utility: Foundations*, page 685–755. Springer US, 2004. ISBN 9781402079641. doi: 10.1007/978-1-4020-7964-1_1. URL `http://dx.doi.org/10.1007/978-1-4020-7964-1_1`.

[80] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from AI feedback, 2022.

[81] OpenAI. Model spec, 2024. URL `https://cdn.openai.com/spec/model-spec-2024-05-08.pdf`. Accessed: 2025-01-02.

[82] Oliver Klingefjord, Ryan Lowe, and Joe Edelman. What are human values, and how do we align ai to them? *arXiv preprint arXiv:2401.12358*, 2024.

[83] Sydney Levine, Nick Chater, Joshua B Tenenbaum, and Fiery Cushman. Resource-rational contractualism: A triple theory of moral cognition. *Behavioral and Brain Sciences*, pages 1–38, 2023.

[84] Joe Edelman and Oliver Klingefjord. Market intermediaries: A post-agi vision for the economy, 2025. URL `https://meaningalignment.substack.com/p/market-intermediaries-a-post-agi`. Substack publication.

[85] Gary S Becker and Casey B Mulligan. The endogenous determination of time preference. *The quarterly journal of economics*, 112(3):729–758, 1997.

[86] Niels Boissonnet, Alexis Ghersengorin, and Simon Gleyze. Revealed deliberate preference change. *Games and Economic Behavior*, 142:357–367, 2023.

[87] Richard Pettigrew. *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`.

[88] Eldar Shafir, Itamar Simonson, and Amos Tversky. Reason-based choice. *Cognition*, 49(1-2): 11–36, 1993.

[89] Itai Sher. Comparative value and the weight of reasons. *Economics and Philosophy*, 35(1): 103–158, 2018. ISSN 1474-0028. doi: 10.1017/s0266267118000160. URL `http://dx.doi.org/10.1017/s0266267118000160`.

[90] Ruth Chang. 'all things considered'. *Philosophical Perspectives*, 18:1–22, 2004.

[91] Ruth Chang. *Putting Together Morality and Well-Being*, page 118–158. Cambridge University Press, January 2004. ISBN 9780511616402. doi: 10.1017/cbo9780511616402.006. URL http://dx.doi.org/10.1017/cbo9780511616402.006.

[92] Carina Prunkl. Human autonomy at risk? an analysis of the challenges from AI. *Minds Mach. (Dordr.)*, 34(3), June 2024.

[93] John Rawls. The idea of public reason revisited. *Univ. Chic. Law Rev.*, 64(3):765, 1997.

[94] J. Habermas. *The Structural Transformation of the Public Sphere: An Inquiry Into a Category of Bourgeois Society*. Studies in contemporary German social thought. Polity Press, 1989. ISBN 9780745602745. URL https://books.google.ca/books?id=3iIVnwEACAAJ.

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

[96] Yan Chen and Sherry Xin Li. Group identity and social preferences. *American Economic Review*, 99(1):431–57, 2009. doi: 10.1257/aer.99.1.431.

[97] John R Searle. *The construction of social reality*. Simon and Schuster, 1995.

[98] Cristina Bicchieri. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2005.

[99] Ingela Alger and Jörgen W Weibull. Evolution and kantian morality. *Games and Economic Behavior*, 98:56–67, 2016.

[100] Jean-Baptiste André, Léo Fitouchi, Stephane Debove, and Nicolas Baumard. An evolutionary contractualist theory of morality. 2022.

[101] Amartya K Sen. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & public affairs*, pages 317–344, 1977.

[102] James G. March and Johan P. Olsen. The logic of appropriateness. In *The Oxford Handbook of Political Science*. Oxford University Press, 07 2011. ISBN 9780199604456. doi: 10.1093/oxfordhb/9780199604456.013.0024.

[103] Joel Z Leibo, Alexander Sasha Vezhnevets, Manfred Diaz, John P Agapiou, William A Cunningham, Peter Sunehag, Julia Haas, Raphael Koster, Edgar A Duéñez-Guzmán, William S Isaac, et al. A theory of appropriateness with applications to generative artificial intelligence. *arXiv preprint arXiv:2412.19010*, 2024.

[104] Robert Sugden. The logic of team reasoning. *Philosophical explorations*, 6(3):165–181, 2003.

[105] Robert J Aumann. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica: Journal of the Econometric Society*, pages 1–18, 1987.

[106] Sydney Levine, Max Kleiman-Weiner, Laura Schulz, Joshua Tenenbaum, and Fiery Cushman. The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, 117(42):26158–26169, 2020.

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

[108] Wolfgang Spohn. *Lehrer Meets Ranking Theory*, pages 129–142. Springer Netherlands, 2003. doi: 10.1007/978-94-010-0013-0_9.

[109] Joseph Farrell and Matthew Rabin. Cheap talk. *Journal of Economic perspectives*, 10(3): 103–118, 1996.

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

[111] Gillian Hadfield and Andrew Koh. An economy of AI agents. *MIT Working Paper*, 2025.

[112] Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel A. Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value profiles for encoding human variation. *ArXiv*, 2025. URL https://www.semanticscholar.org/paper/fb3d7068979d80ddde18c23a96b8cc98916d7523.

[113] Saffron Huang, Esin Durmus, Megan McCain, Kush Handa, et al. Values in the wild: Discovering and analyzing values in real-world language model interactions. *arXiv preprint*, 2025.

[114] Raphaël Millière. Normative conflicts and shallow AI alignment. *Philosophical Studies*, pages 1–44, 2025.

[115] Nils Köbis, Jean-François Bonnefon, and Iyad Rahwan. Bad machines corrupt good morals. *Nat. Hum. Behav.*, 5(6):679–685, June 2021.

[116] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, J. Pachocki, and David Farhi. Monitoring reasoning models for mis-

behavior and the risks of promoting obfuscation. *ArXiv*, 2025. URL `https://www.semanticscholar.org/paper/5c33a1dade777d08f3d4ba8a761a4902dafd211a`.

[117] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. AI alignment with changing and influenceable reward functions, 2024. URL `https://arxiv.org/abs/2405.17713`.

[118] Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models, 2024. URL `https://arxiv.org/abs/2406.10162`.

[119] Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual disempowerment: Systemic existential risks from incremental ai development. *arXiv preprint*, 2025.

[120] Seliem El-Sayed, Canfer Akbulut, Amanda McCroskery, Geoff Keeling, Zachary Kenton, Zaria Jalan, Nahema Marchal, Arianna Manzini, Toby Shevlane, Shannon Vallor, et al. A mechanism-based approach to mitigating harms from persuasive generative ai. *arXiv preprint arXiv:2404.15058*, 2024.

[121] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024. doi: 10.1126/science.adq2852. URL `https://www.science.org/doi/abs/10.1126/science.adq2852`.

[122] Oscar Winberg. Insult politics: Donald trump, right-wing populism, and incendiary language. *European journal of American studies*, 12(2), July 2017. ISSN 1991-9336. doi: 10.4000/ejas.12132. URL `http://dx.doi.org/10.4000/ejas.12132`.

[123] xAI (Grok chatbot). Update on where has @grok been what happened on July 8th, July 2025. URL `https://x.com/grok/status/1943916977481036128`. Official tweet from the Grok chatbot account addressing recent behavior.

[124] Shaun Wilson and Nick Turnbull. Wedge politics and welfare reform in australia. *Australian Journal of Politics & History*, 47(3):384–404, September 2001. ISSN 1467-8497. doi: 10.1111/1467-8497.00235. URL `http://dx.doi.org/10.1111/1467-8497.00235`.

[125] Elaine Wallace, Isabel Buil, and Leslie de Chernatony. "consuming good" on social media: What can conspicuous virtue signalling on facebook tell us about prosocial and unethical intentions? *Journal of Business Ethics*, 2018. doi: 10.1007/s10551-018-3999-7.

[126] Shelly Kagan. The structure of normative ethics. *Philosophical perspectives*, 6:223–242, 1992.

[127] Mark Schroeder. Normative ethics and metaethics. In *The Routledge handbook of metaethics*, pages 674–686. Routledge, 2017.

[128] Charles Taylor. *Explanation and Practical Reason*, pages 208–231. Oxford University Press, 1993. ISBN 9780198287971. doi: 10.1093/0198287976.003.0017. URL `http://dx.doi.org/10.1093/0198287976.003.0017`.

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

[130] Alex John London and Hoda Heidari. Beneficent intelligence: a capability approach to modeling benefit, assistance, and associated moral failures through AI systems. *Minds and Machines*, 34(4):41, 2024.

[131] Amartya Sen. *Development as Freedom*. Alfred A. Knopf, 1999. ISBN 9780198297581.

[132] Martha C. Nussbaum. *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`.

[133] Geoffrey Brennan, Lina Eriksson, Robert E Goodin, and Nicholas Southwood. *Explaining norms*. OUP UK, 2013.

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

[135] John Rawls. *A Theory of Justice*. Belknap Press, 1971.

[136] James J. Gibson. *The Ecological Approach To Visual Perception*. Houghton Mifflin, Boston, 1979.

[137] David Gauthier. *Morals by agreement*. OUP Oxford, 1986.

[138] Ken Binmore. *Natural justice*. Oxford university press, 2005.

[139] Thomas M Scanlon. *What we owe to each other*. Harvard University Press, 2000.

[140] Sydney Levine, Matija Franklin, Tan Zhi-Xuan, Secil Yanik Guyot, Lionel Wong, Daniel Kilov, Yejin Choi, Joshua B. Tenenbaum, Noah Goodman, Seth Lazar, and Iason Gabriel. Resource rational contractualism should guide AI alignment, 2025.

[141] Charles Taylor. Sources of the self: The making of the modern identity. *Harvard UP*, 1989.

[142] Robert Axelrod. An evolutionary approach to norms. *American political science review*, 80 (4):1095–1111, 1986.

[143] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984. ISBN 0-465-02122-0.

[144] Fazl Barez and Philip Torr. Measuring value alignment, 2023. URL `https://arxiv.org/abs/2312.15241`.

[145] Marc Carauleanu, Michael Vaiana, Judd Rosenblatt, Cameron Berg, and Diogo Schwerz de Lucena. Towards safe and honest AI agents with neural self-other overlap, 2024. URL `https://arxiv.org/abs/2412.16325`.

[146] J. David Velleman. *Practical Reflection*. Princeton University Press, Princeton, NJ, 1989. ISBN 9780691020082.

[147] Ninell Oldenburg and Tan Zhi-Xuan. 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*, pages 1510–1520, 2024.

[148] Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandray, Piyush Madan, Kush Varshney, Murray Campbell, Moninder Singh, and Francesca Rossi. 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.

[149] Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative alignment: Reasoning enables safer language models, 2025. URL `https://arxiv.org/abs/2412.16339`.

[150] Philipp Koralus. The philosophic turn for AI agents: Replacing centralized digital rhetoric with decentralized truth-seeking, 2025. URL `https://arxiv.org/abs/2504.18601`.

[151] Christopher Burr, Nello Cristianini, and James Ladyman. 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`.

[152] Philip Pettit. 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`.

[153] Cathy Mengying Fang, Auren R. Liu, Valdemar Danry, Eunhae Lee, Samantha W. T. Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, and Sandhini Agarwal. How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study, 2025. URL `https://arxiv.org/abs/2503.17473`.

[154] Joe Edelman and Oliver Klingefjord. Model integrity. Meaning Alignment Institute Substack, Dec 2024. URL `https://meaningalignment.substack.com/p/model-integrity`. Accessed: 2025-05-05.

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

[156] Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A Hale. Why human-AI relationships need socioaffective alignment. *arXiv preprint arXiv:2502.02528*, 2025.

[157] Joe Kwon, Tan Zhi-Xuan, Joshua Tenenbaum, and Sydney Levine. 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.

[158] Dylan Hadfield-Menell and Gillian K Hadfield. Incomplete contracting and ai alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422, 2019.

[159] Nathanaël Hyafil and Craig Boutilier. Mechanism design with partial revelation. In *IJCAI*, pages 1333–1340, 2007.

[160] Andrew Critch, Michael Dennis, and Stuart Russell. Cooperative and uncooperative institution designs: Surprises and problems in open-source game theory. *arXiv preprint arXiv:2208.07006*, 2022.

[161] Stephen Darwall. Law and the second-person standpoint. *Loy. LAL Rev.*, 40:891, 2006.

[162] Gillian K Hadfield and Barry R Weingast. Microfoundations of the rule of law. *Annual Review of Political Science*, 17(1):21–42, 2014.

[163] Martin T. Bohl, Alexander Pütz, and Christoph Sulewski. 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`.

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

[165] Thomas Philippon. Has the us finance industry become less efficient? on the theory and measurement of financial intermediation. *American Economic Review*, 105(4):1408–1438, 2015. doi: 10.1257/aer.20120578.

[166] Oxford Internet Institute. Industrialized disinformation: 2020 global inventory of organized social media manipulation. Technical report, Oxford Internet Institute, University of Oxford, 2020. URL `https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/`.

[167] Jeff Orlowski. The social dilemma. Netflix documentary film, 2020.

[168] Stephen E. Spear and Sanjay Srivastava. 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`.

[169] Jean-Jacques Laffont and Jean Tirole. *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`.

[170] Daniel Halpern, Safwan Hossain, and Jamie Tucker-Foltz. Computing voting rules with elicited incomplete votes. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, page 941–963. ACM, July 2024. doi: 10.1145/3670865.3673556. URL `http://dx.doi.org/10.1145/3670865.3673556`.

[171] Atrisha Sarkar, Andrei Ioan Muresanu, Carter Blair, Aaryam Sharma, Rakshit S Trivedi, and Gillian K Hadfield. Normative modules: A generative agent architecture for learning norms that supports multi-agent cooperation. *arXiv preprint arXiv:2405.19328*, 2024.

[172] Brian D. Earp, Sebastian Porsdam Mann, Mateo Aboy, Edmond Awad, Monika Betzler, Marietjie Botes, Rachel Calcott, Mina Caraccio, Nick Chater, Mark Coeckelbergh, Mihaela Constantinescu, Hossein Dabbagh, Kate Devlin, Xiaojun Ding, V. Dranseika, J. A. Everett, Ruiping Fan, F. Feroz, Kathryn B. Francis, Cindy Friedman, Orsolya Friedrich, Iason Gabriel, Ivar Hannikainen, Julie Hellmann, Arasj Khodadade Jahrome, N. Janardhanan, Paulius Jurcys, Andreas Kappes, Maryam Ali Khan, Gordon Kraft-Todd, Maximilian Kroner Dale, S. Laham, Benjamin Lange, Muriel Leuenberger, Jonathan Lewis, Pengbo Liu, David M. Lyreskog, Matthijs Maas, John McMillan, Emil G. Mihailov, Timo Minssen, J. Monrad, Kathryn Muyskens, Simon Myers, Sven Nyholm, Alexa M. Owen, Anna Puzio, Christopher Register, Madeline G. Reinecke, Adam Safron, Henry Shevlin, Hayate Shimizu, Peter V. Treit, Cristina Voinea, Karen Yan, Anda Zahiu, Renwen Zhang, Hazem Zohny, Walter Sinnott-Armstrong, Ilina Singh, Julian Savulescu, and Margaret S. Clark. Relational norms for human-AI cooperation. *ArXiv*, 2025. URL `https://www.semanticscholar.org/paper/64c2bb56fe73814e4ebc9995187943fc0256a88f`.

[173] Atoosa Kasirzadeh and Iason Gabriel. Characterizing AI agents for alignment and governance. *ArXiv preprint*, 2025.

[174] W. Russell Neuman, Chad Coleman, Ali Dasdan, Safinah Ali, and Manan Shah. Auditing the ethical logic of generative AI models, 2025. URL `https://arxiv.org/abs/2504.17544`.

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

[176] Philipp Koralus. *Reason and Inquiry: The Erotetic Theory*. Oxford, GB, 2022.

A   More Emerging Research on Thick Models of Value

Here we briefly outline additional examples of research on TMV.

Self-Other Generalization   One example that shows promise in ML fine-tuning and which builds on a notion of moral progress: Velleman [146, 129] 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 instrumental or situational concerns, or local preferences, do not 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 [145] can be read as an example of fine-tuning work in this vein.

Norm Learning and Reasoning   Following philosophical accounts of norms that draw the above distinctions [98, 133], AI researchers have begun enriching classical game-theoretic models with shared normative structure, offering a structural account for how agents trade off norm compliance with their desires or objectives when taking actions, resulting in norm-augmented utility functions [147, 148] which let AI agents learn norms from collective behavior [147] or social sanctions [171]. Such norms can be structured as constraints or filters that modify plans of action to ensure compliance [172, 173], specifying conditions that acceptable actions must satisfy and creating a clear demarcation between norm-compliant and norm-violating behavior.

Moral Reasoning Traces for LLMs   Alignment with reasoning traces rather than final outputs [149] is a promising area for thick models of value. Researchers could use formal theories to supervise the generation of normative reasoning traces by LLMs, thereby ensuring that AI systems are trained in accordance with a *systematized* version of human meta-ethical intuitions, not just first-order evaluative judgments that may be flawed and subject to future revision. Some nascent research points in this direction [174]. Further work could extend and translate existing theories of evaluative and normative reasoning into formal computational models, drawing on work in deontic and argumentative reasoning [47, 175, 49], contractualist models of moral cognition [110, 83], or question-based theories of reason [176].