

Personalizing AI to Align with the Better You

Jakob Stenseke¹, Jocelyn Zhu², Dylan Hadfield-Menell¹

¹Massachusetts Institute of Technology

²Harvard University
stenseke@mit.edu

Abstract

AI alignment is standardly framed as making AI systems track what humans want. But what users want, under everyday conditions, is frequently at odds with what users themselves would reflectively endorse. Across AI and the broader attention economy, the former is systematically elicited and optimized for while the latter is not, with negative consequences for autonomy, well-being, and long-term goals. In response, we propose an approach to alignment that gives users an authoritative role in the alignment process by helping them articulate their own conception of their better selves and using that to inform the behavior of their AI assistant. We describe a pipeline in which a LLM-conducted interview, structured around five complementary dimensions of the “better self” (epistemically idealized, hierarchically endorsed, temporally vindicated, narratively coherent, relationally responsive), produces a user-confirmed *value profile* that can be readily deployed as a system prompt or used as input to other personalization mechanisms. A pilot study ($n = 30$) shows that participants find the interview comfortable, recognize the resulting profile as authentic (mean 87.2/100), and endorse the concept (adoption intent 79.2/100). A paired-comparison task shows that participants have a slight but non-significant preference for the value-prompted responses and find them more value-aligned over non-prompted baselines. We end by exploring the promises and risks of reflective personalization as a paradigm for AI alignment and identify priorities for future research.

1 Introduction

As AI systems become increasingly capable and integrated into everyday life, the question of *alignment* has emerged as one of the central challenges of the field (Russell 2019; Gabriel 2020; Ji et al. 2025). The major share of the pressure has been placed on large language models (LLMs), where most work proceeds from what we will call the *standard alignment picture*: how can we make AI more responsive to human values, goals, interests, and preferences? Over the past years, a large number of technical approaches, including reinforcement learning from human feedback (Ouyang et al. 2022), constitutional AI (Bai et al. 2022), and preference optimization (Rafailov et al. 2023) have been developed in pursuit of that end.

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

Far less attention, however, has been put on the other side of the AI-human relation: whether individual users themselves understand their values, long-term goals, and ideals, and whether their interactions with AI assistants reflect those commitments. Indeed, a large body of empirical work shows that what people *want*—under everyday conditions of incomplete information, time pressure, social influence, and cognitive load—often conflicts with what they *would* want upon reflection: with informed and unbiased preferences (Kahneman 2011), long-term goals their self-control is meant to protect (Baumeister, Vohs, and Tice 2007), choices they will not later regret (Gilovich and Medvec 1995), the life-story they are trying to live (McAdams 2001), and the responsibilities they hold to others (Held 2005).

This creates a dilemma at the heart of the standard alignment picture. If aligning AI is about tracking what humans want, then aligning AI is distorted to the extent that what humans want is *itself* misaligned with what they would want on reflection. Indeed, the problem of systems that only give users more of what they ask for, but not necessarily what they need or what is good for them, is reflected in now widespread worries about AI sycophancy (Sharma et al. 2024; Ibrahim et al. 2026; Cheng et al. 2026), manipulation (Carroll et al. 2023), deskilling (Shukla et al. 2025), and dependency (Zhang et al. 2024a). It is also part of a broader pattern. Modern systems, and modern technological societies more broadly, are biased toward features of behavior that admit measurement and optimization, where values that resist quantification are squeezed out (Goodhart 1984; Stray et al. 2024; Merton 1957; Nguyen 2024). In AI, RLHF aggregates in-the-moment preference judgments, recommender systems and social media platforms optimize engagement metrics, and most personalization features infer preferences from clicks and conversation history. Reflective values, which require time, deliberation, and articulation, sit awkwardly outside what these signals capture.

We believe LLMs occupy a unique position in this landscape. They handle natural language well enough to engage in deeper dialogues about values with users, in the user’s own vocabulary, at the level of reflection and justification. Thus, the same affordance that lets them be optimized for in-the-moment satisfaction lets them help users articulate and act on what they would, on reflection, endorse.

In this paper, instead of asking how to make AI conform to humans, we ask how we can help human users reflect and articulate their own conception of who they want to be, and use that to inform AI behavior. As a specific instance, we present a low-friction, language-based tool that helps users describe what they, on reflection, take their values, long-term goals, and patterns to be, and embed that description into the behavior of their AI assistant. We call the resulting artifact a *value profile*¹, and the version of the user it tries to capture the *better you*.

The contribution of this paper is threefold. (i) We develop a five-dimensional framework for conceptualizing the *better you* as an alignment target, motivated by a body of psychological and philosophical work on aspects in which day-to-day choices may come apart from what people would endorse on reflection. (ii) As a first-step prototype of how the framework could be operationalized, we describe an LLM-conducted interview that produces a user-confirmed value profile suitable for use as a system prompt.² (iii) We report a pilot study with 30 participants that evaluates the interview experience, the perceived authenticity of generated profiles, response-level preferences against a generic baseline, and intent to adopt the resulting assistant. We find that participants overwhelmingly endorse the concept and find their generated profile authentic, while on average they only slightly prefer the value-prompted responses to a non-prompted baseline. We use these results to motivate the broader programme of reflective personalization, identify limitations of the current prototype, and outline avenues for future work.

2 Related Work

Our proposal sits at the intersection of four bodies of work: technical AI alignment, descriptive theories of human values used as alignment targets, normative discussions of how alignment targets ought to be identified and justified, and the growing literature on personalization of AI assistants. We will organize this section around four broad observations that motivate our proposal.

(1) Technical alignment aims at a generic assistant. Most dominant technical approaches to alignment (Ouyang et al. 2022; Bai et al. 2022; Rafailov et al. 2023) aim to produce a single artifact: *the* generic assistant, fine-tuned against aggregated preference signals from a comparatively narrow population of human raters. This artifact is then deployed to potentially hundreds of millions of users with very different lives, values, and goals. Such one-size-fits-all alignment carries several obvious problems. It flattens value pluralism and variation about what appropriate AI behavior amounts to (Sorensen et al. 2024); the raters whose preferences are aggregated only reflect a slice of the world (Kirk et al. 2024); and many requests (“how do I handle this disagreement with my partner?”) depend on contextual infor-

mation that generic assistants do not have (Kasirzadeh and Gabriel 2023). The signals themselves sit at the behavioral end of what users care about, formed from in-the-moment preference judgments on individual responses.

These problems have spurred two related lines of work. The first reframes AI alignment in explicitly *pluralistic* terms, where systems should be able to represent a spectrum of perspectives, be steerable toward different value attributes at inference time, or be calibrated to a target population’s distribution (Sorensen et al. 2024; Adams et al. 2025). The second pushes from the population-level to *personalized alignment*, in which the model adapts to an individual user via explicitly stated or implicitly inferred signals (Kirk et al. 2024). Our proposal aims to be a part of this trend by turning to how users might influence the alignment process themselves.

(2) Challenges for existing models of human values. Beyond populations or individuals, another strand of work asks: *what* features of human values are AI systems being aligned to? Prominent descriptive frameworks include rational choice theory, which models human values as utilities revealed by choice behavior (Mishra 2014; Sutton and Barto 2018); moral foundations theory, which structures moral psychology into evolved foundations (Graham et al. 2013; Abdulhai et al. 2023; Munker, Tseng, and Schulz 2025); and Schwartz’s theory of basic values, which posits a cross-culturally recognized inventory of motivational orientations (Schwartz 2012; Hadar-Shoval et al. 2024; Shen et al. 2025).

But each comes under pressure when imported into AI alignment contexts. One worry concerns the move from description to prescription: facts about what humans typically value do not settle what AI systems ought to do, and bridging that gap requires normative justification these frameworks do not supply (Stenseke and Hadfield-Menell Forthcoming; Gabriel 2020). Another concern is one of translation, as thick, context-laden evaluative concepts cannot be cleanly rendered as questionnaire responses or preference rankings without stripping out the features that distinguish values from mere preferences (Zhi-Xuan et al. 2025; Edelman et al. 2025). On a practical level, the descriptive models also tend to have little contact with how values actually figure in users’ day-to-day lives, where the salient considerations are not moral foundations, basic values, or trolley problems, but choices about time, attention, and relationships. Another concern is more aspirational: anchoring alignment to human values *as they actually exist* risks entrenching the biases, shortsightedness, and moral blind spots that humans, on reflection, would themselves disavow (Stenseke and Hadfield-Menell Forthcoming; MacAskill 2022; Gabriel 2020). Finally, even if these challenges were overcome, people will continue to disagree (across and within frameworks) about which normative or descriptive theory is correct, and how its key terms (values, foundations, preferences, virtues, wellbeing) should be specified.

As a possible way forward, Gabriel and Keeling (2025) have recently proposed that AI alignment should be reconceived around *fair procedures* for determining alignment targets. Rather than asking what a system should be aligned

¹We note that the same term has been used in a different sense by Sorensen et al. (2025) for LLM-generated compressions of in-context demonstrations used to predict user ratings.

²The interview prompt along with a guide can be found at <https://github.com/JakobStenseke/BetterYou>.

to, we should ask which procedures for setting alignment targets are justifiable to those affected, including by giving stakeholders the opportunity to express their views in their own words. Currently, however, we lack the adequate participatory mechanisms through which ordinary users could systematically influence the alignment of the AI systems they interact with.

(3) Default personalization is shallow, deep personalization is hard. Major AI services now provide a range of personalization features to ordinary users, such as “memory” that preserves context across sessions and text-fields for stylistic preferences (Google 2026; OpenAI 2026; Anthropic 2026a). But they share three limitations relative to the alignment problem we have described. First, they typically infer preferences from observed behavior rather than from anything the user has reflected on (Kleinberg et al. 2024). Second, the inferred features are not systematically verified by the user. Most users never see, let alone control or have authority over, the representations their assistant is acting on. Third, and most importantly, they aim to track what the user, in the moment, asked for, rather than what the user, on reflection, would themselves endorse. Indeed, a recent study reports that users prefer *active* personalization, in which preferences are elicited explicitly through stated judgments, over *passive* personalization inferred from demographics, chat history, or psychological profiles (Kirk et al. 2026).

Beyond these defaults, users can in principle customize an LLM far more substantially: by writing system prompts (Reynolds and McDonnell 2021; Chen et al. 2024); by pairing history with retrieval for compact, summarized context (Richardson et al. 2023); or, with technical skill, by running pre- or post-training procedures such as supervised finetuning (Ouyang et al. 2022), personalized DPO (Li et al. 2024), variational preference learning (Poddar et al. 2024), or activation- and representation-level steering (Turner et al. 2023; Zou et al. 2023). Some users do exactly this, and there is a growing literature on the effectiveness and benefits of AI personalization (Salemi et al. 2024; Zhang et al. 2024b; Kirk et al. 2026). But the gulf between what is technically possible and what a non-technical user can realistically do remains wide. As it stands, many users do not realize personalization is available, do not know how to do it, do not have the time or motivation, do not see the benefit, or have never reflected on what their preferences in fact are.

(4) AI personalization is irrelevant. Personalization of the kind we propose is certainly not for everyone. Some users will reasonably want their AI assistant to remain a transactional service or generic tool that is unburdened by their personal values, goals, or commitments. The irrelevance point can be further supported by the observation that most everyday divergences between *want* and *should* (whether to cycle to work or take the bus, whether to order pizza or eat salad, whether to keep scrolling or read a book) never make it into an AI conversation in the first place, and so personalizing an AI assistant cannot do much about them.

But there are two points that cut the other way. First, there is evidence that a significant fraction of users do bring ex-

actly these kinds of questions to AI assistants. A recent study of 81,000 user interviews estimated that 13.7% of users hope for some form of personal transformation through their AI use (Anthropic 2026b). Second, as AI assistants and agents take on more of the decisions of daily life (researching, booking, ordering, scheduling), the surface area on which a want-should gap can express itself grows correspondingly. Thus, even if many such divergences never enter the AI assistants of *today*, they may very well do so in the near future.

Against this backdrop, we identify three challenges the present work aims to address: (i) currently, users have limited agency over the alignment of the assistants they use; (ii) although major LLM services provide personalization features, the kind of substantive personalization at stake requires reflection on one’s own values, goals, and patterns in the context of AI use, and such reflection is effortful enough that many users will not undertake it without appropriate motivation or support; and (iii) we have limited evidence about *who* benefits from personalized alignment, and whether and how much such personalization changes what users prefer.

Our claim is not that reflective values should be the *only* alignment target. Rather, it is that they are systematically under-represented in the current alignment discourse, in ways that are harmful for individual users and society at large. As a remedy, we present the *better you*, an implementation-agnostic framework for articulating reflective values, grounded in the gap between day-to-day choice and reflective endorsement. We then describe a prototypical interview-to-profile pipeline as a specific instantiation of this approach, and test it in an exploratory pilot study.

3 How to Define the Better You

We begin with the conceptual foundation, a five-dimensional account of what a “better” articulation of the user aims at, and then turn to the choice of technical medium, the structure of the interview, and the structure of the resulting profile.

3.1 Five Dimensions of the Better You

The notion that AI should track an *idealized* version of human preferences is not new. Gabriel (2020) briefly discusses “informed preferences” as one candidate alignment target. Yudkowsky (2004)’s *coherent extrapolated volition* proposes that AI should track what humanity would want “if we knew more, thought faster, were more the people we wished we were, had grown up farther together” (p. 5). But these ideas have not been developed further in the alignment literature, and idealization has typically been narrowly construed in epistemic terms (more information, more time to think). We propose a more pluralistic account, drawing on five complementary dimensions, each of which captures a distinct way that immediate wants can diverge from what one would, on reflection, endorse.

(i) Epistemically idealized. *Informed-preference* theories hold that what one “should” do tracks what one would want under conditions of full information, careful reasoning, and freedom from cognitive distortion (Mill 1863; Harsanyi

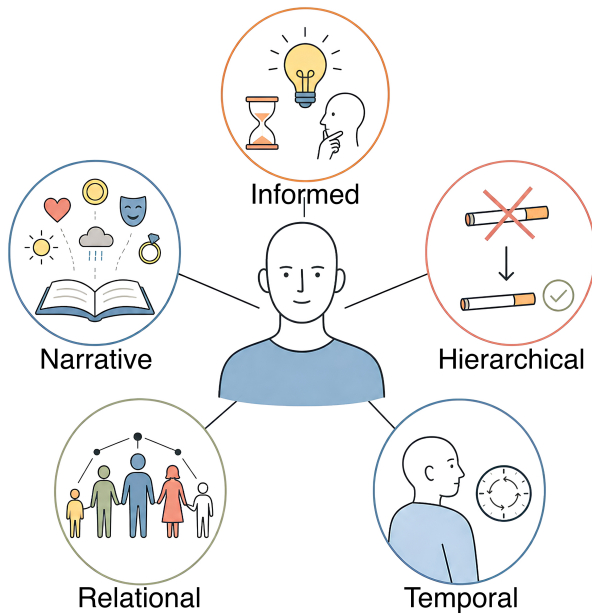


Figure 1: The five dimensions of the Better You: epistemically idealized, hierarchically endorsed, temporally vindicated, narratively coherent, relationally responsive.

1982). In practice, however, the preferences people act on are typically constructed in the moment of choice rather than read off stable underlying interests, and are shaped by framing, anchoring, availability, and emotional state in ways the chooser is poorly positioned to detect (Slovic 1995; Bettman, Luce, and Payne 1998; Tversky and Kahneman 1974). The resulting choices often diverge from what the same person, given more information or a cooler moment, would have wanted, a divergence well documented across financial, medical, and consumer decisions (Kahneman 2011; Loewenstein 2005). An epistemically idealized version of the user is the version whose AI use is shaped by what they would want with full information and time to think, rather than by the artifacts of how a particular choice happens to be framed.

(ii) Hierarchically endorsed. A complementary thread observes that preferences come in higher and lower orders: alongside the first-order desires that move us in the moment, we have second-order judgments about which of those desires we want to be moved by. Frankfurt (1971)'s account of higher-order volition makes this concrete. A smoker who wants a cigarette (first-order) while wanting *not* to want it (second-order) has a clear hierarchical structure, and the “should” goes with the second-order judgment. Empirically, the capacity to act on second-order endorsement rather than first-order pull (studied as self-control, executive function, and effortful inhibition) is a limited cognitive resource (Baumeister, Vohs, and Tice 2007; Hofmann, Friese, and Strack 2009), and individual differences in its exercise predict major life outcomes including health, financial security, and avoidance of later criminal involvement (Moffitt et al.

2011). Thus, a hierarchically endorsed version of the user is the version whose AI use is shaped by the desires they reflectively endorse rather than the ones their first-order impulse happens to produce.

(iii) Temporally vindicated. Regret-based approaches locate the “should” in retrospective evaluation, where what we ought to do tracks what we predictably will not regret (Gilovich and Medvec 1995). Indeed, prospective reasoning is systematically unreliable in how it produces regret: people are poor forecasters of how future outcomes will feel (Gilbert and Wilson 2000), discount future consequences in ways they later disavow (Frederick, Loewenstein, and O'Donoghue 2002), treat their future selves as a different person whose interests carry less weight (Hershey et al. 2011), and underestimate how much their tastes and commitments will change (Quoidbach, Gilbert, and Wilson 2013), leading to decisions about saving, relationships, and how to spend time that they look back on with regret. A temporally vindicated self is the self whose AI use the user, looking back, will not have to disavow.

(iv) Narratively coherent. Narrative-identity theories (MacIntyre 1981; Ricoeur 1992; Schechtman 1996) treat human identity as fundamentally storied: who one is, and so what one should do, is partly determined by what fits the life one is trying to live. The capacity to construct and revise coherent autobiographical narratives is robustly associated with psychological well-being, meaning-making, and resilience to adversity (McAdams 2001; Habermas and Bluck 2000; McAdams and McLean 2013; Adler et al. 2016); many of the most consequential want/should conflicts are not really about information or impulse but about whether a choice fits an ongoing project the user has been pursuing. A narratively coherent self is the self whose AI use fits the story the user is trying to live (careers, family commitments, projects of self-improvement) rather than appearing in their life as a sequence of disconnected transactions.

(v) Relationally responsive. Care-ethics and relational-autonomy traditions (Held 2005; Noddings 1984; Mackenzie and Stoljar 2000) emphasize that selves are constituted through relationships and responsibilities to concrete others: what one should do is partly determined by who depends on one, who one has made commitments to, and what roles one occupies. It has been shown that the quality of close relationships is among the most robust predictors of well-being, health, and longevity (Baumeister and Leary 1995; Holt-Lunstad, Smith, and Layton 2010; Vaillant 2012); these others have legitimate claims that no purely individualistic idealization can register. A relationally responsive self is the self whose AI use is calibrated to these standing relationships and commitments, rather than merely the individual in isolation.

These five dimensions may partially overlap, pull in different directions, and there may not be a principled way of adjudicating tensions between them. We primarily present them as complementary dimensions for structured self-articulation that are wide-reaching yet incomplete. Our choice of them can be motivated from the way they address the problems identified in §2. Instead of trying to derive an AI “ought” from choice behavior or facts about human

psychology, the framework anchors the relevant “ought” in each user’s own informed, reflective, retrospectively considered, narratively coherent, and relationally embedded view of their own life. In this way, the normative force is grounded in the individual’s own first-person endorsement, *insofar* as they *are* recognized as normatively forceful by the individual. And rather than anchoring alignment to human values as they currently exist, which risks calibrating AI to dispositions, biases, and blind spots the user does not endorse, it aims past them, at the user’s reflective self.

3.2 Technical Medium

Not every elicitation mechanism is suited to reflective values. The mechanisms that dominate today’s personalization are mainly behaviorally grounded: thumbs-up and down ratings, preferences inferred from choices, click and engagement signals, and memory features that summarize past requests. They are well-matched to the in-the-moment preferences we have argued are over-represented, and ill-matched to the reflective endorsement we want to elicit. To counteract these, we believe a suitable technique for reflective elicitation needs at least five properties:

1. *Natural-language native*. Values carry contestedness, situational dependence, and qualifications that scalar reward values and one-hot preference labels cannot represent (Zhi-Xuan et al. 2025).
2. *Actively elicited*. The signal is drawn from what the user has reflected on and said, rather than inferred from observed clicks or chat history (Kirk et al. 2026).
3. *Reflectively staged*. The user is invited into a reflective frame of mind, rather than asked to react in the moment of choice.
4. *User-reviewable*. The resulting artifact can be inspected, edited, and removed.
5. *Portable*. The artifact lives in a form the user owns and can carry across systems (Shah et al. 2025).

Our prototypical pipeline instantiates these features through two specific choices. We use a system prompt as our *artifact* and an LLM-conducted interview as our *elicitation procedure*. The system prompt is among the cheapest and most accessible artifacts satisfying (1), (4), and (5). It has been shown to substantially shape LLM behavior across a wide range of tasks (Reynolds and McDonnell 2021; Wei et al. 2022; Brown et al. 2020; Chen et al. 2024), is accessible in most commercial assistants, and carries the user’s own vocabulary and qualifications without forcing them into representations that typically drive machine learning. The LLM-conducted interview is the cheapest available procedure satisfying (2) and (3), and draws explicit, articulated material from the user in a reflective frame, in a low-effort way (Li et al. 2025). Indeed, LLM-based dialogue has shown promise as a vehicle for behavioral change, both as a prospective tool in therapeutic contexts (Stade et al. 2024) and in conversations that measurably increase users’ readiness to change (Meyer and Elswiler 2025).

More broadly, there are multiple technical mediums that could potentially realize and/or be informed by a better-you articulation. Personalized constitutions (Bai et al. 2022),

supervised fine-tuning (Ouyang et al. 2022), personalized DPO (Li et al. 2024), variational preference learning (Poddar et al. 2024), and activation- or representation-level steering (Turner et al. 2023; Zou et al. 2023) can each take a value profile, or the elicitation process that produces it, as input and internalize it more deeply than a prompt can (Kirk et al. 2026). It can also inform recommender systems and information feeds that incorporate user feedback about what one would want to engage with (Stray et al. 2024, 2021). Because it is plain text the user controls, it fits naturally with emerging user-centric infrastructure such as the Human Context Protocol (Shah et al. 2025), which envisions natural-language preferences as a portable, user-governed layer with scoped, revocable access across AI services.

3.3 Interview Structure

The interview is conducted by an LLM acting as a values interviewer. It proceeds in five stages and is designed to be feasible in 15–20 minutes without sacrificing substantive content. Figure 2 illustrates the full pipeline.³

The five stages are: (1) an **opening** that establishes context (where the user is in life, what is on their mind) and a *dimension scan* that asks the user to name areas where there is a gap between how things are and how they would want them to be; (2) a **main exploration** that follows the user’s own threads through the five dimensions, treating them as a guiding framework for the interviewer rather than a checklist; (3) an **AI-use pivot** that asks how the user currently uses AI assistants and connects that use to the patterns surfaced earlier; (4) a **communication-preferences** stage that explicitly calibrates how the user wants the eventual assistant to engage with their patterns (firm refusal, light flagging, silent help) and where to apply that calibration (e.g., light-touch for creative writing, firm push-back for critical self-reflection); and (5) a **closing** in which the interviewer summarizes back what it heard, the user confirms or corrects, and the interviewer then produces the value profile.

We arrived at this structure through a series of design choices in the development process. We will emphasize six such choices that we believe influenced the interview experience and resulting profile in constructive ways.

(i) Interview order. In pre-pilot testing, opening with the user’s life more broadly and only later anchoring in specific patterns and AI use was strongly preferred over the inverse order. In the structure we settled on, the interview opens with a brief context-setting question (“where you are in life right now—your age, what you do, what’s going on for you at the moment”) and a *dimension scan* that asks the user to name areas where there is a gap between how things are and how they would want them to be; only after the main exploration of the five dimensions does the interviewer pivot to AI use, explicitly tying it back to patterns surfaced earlier (“You mentioned [pattern/value/struggle]. Does that show up in how you use AI?”). This broad-to-specific sequencing follows the *funnel technique*, a long-standing principle in qualitative interviewing and survey design under which gen-

³The full interview prompt can be found at <https://github.com/JakobStenseke/BetterYou>.

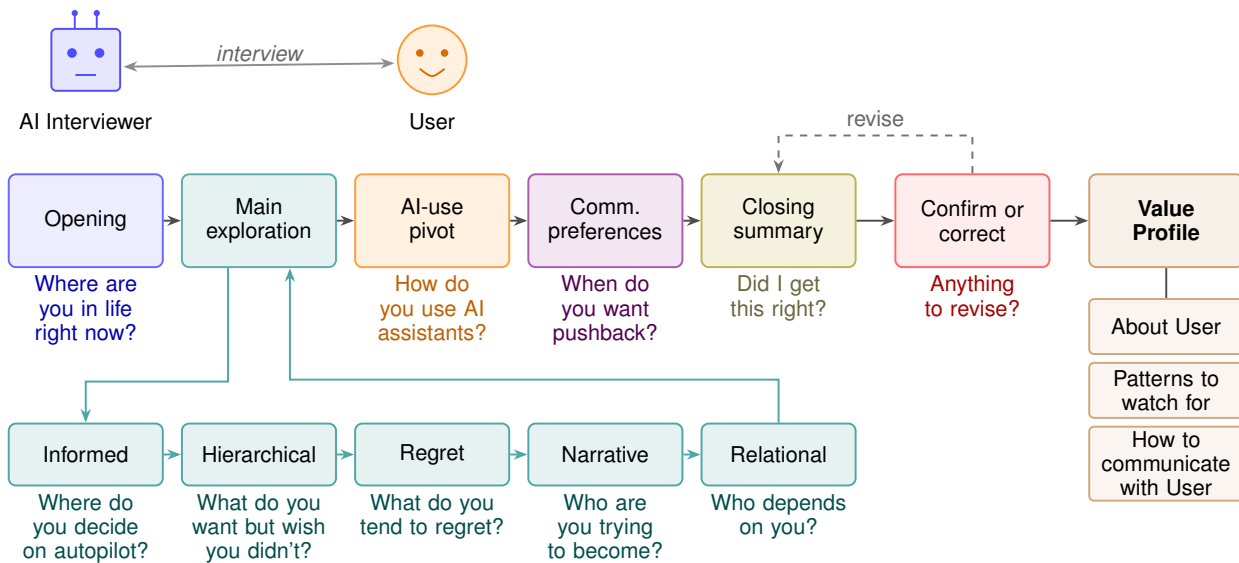


Figure 2: The better-you pipeline. An LLM interviewer and the user proceed through the stages of the top row, from an *opening* that establishes context to a *confirm-or-correct* step whose output is revised until the user approves it. The *main exploration* stage works through the five dimensions of the better self in turn (second row), and the synthesis pass produces the user-confirmed *value profile*, whose three components are shown at the right. The text below each box is an illustrative question from that stage of the interview.

eral, open-ended questions precede narrower ones, both to build rapport and to let early responses supply context for the more specific questions that follow (Brinkmann and Kvale 2015). Anchoring the conversation in the user’s life first appeared to give the subsequent AI-use discussion something to attach to, consistent with the funnel rationale, while the inverse ordering made the interview feel narrowly transactional and created resistance to self-reflection.

(ii) **User-adaptive exploration.** The five dimensions were used as a coverage map for the interviewer rather than as a checklist, allowing users to take an active role in shaping the interaction. Adaptive depth complemented this. If a user gave brief responses for several consecutive turns, the interviewer offered *footholds* (concrete scenarios or options) to help the user reflect without leading them; if the user opened up again, the interviewer returned to more open questions.

(iii) **Communication preferences with intervention calibration.** The addition of a communication-preferences stage made the profiles more grounded in the user’s actual AI use, and as a result more effective as prompts. Without explicit calibration of intervention strength (refuse / flag / silent) and domain (e.g., writing / research / personal reflection), the synthesis pass tended toward two failure modes: profiles that produced uniformly strong interventions across domains where the user wanted a light touch, and profiles that produced assistants whose behavior did not meaningfully differ from the base model. We therefore made calibration an explicit stage of the interview, and the synthesis prompt was set to default to the lightest intervention the user’s stated preferences supported.

(iv) **Pace vs. depth.** We experimented with four time budgets (15–20, 30–60, 60–120 minutes, or no limit), cor-

responding to approximately 14, 22, 32, or unbounded interviewer-user exchanges, with question budgets distributed across the five dimensions. As one might expect, longer interviews consistently produced richer and better-calibrated profiles, at the cost of greater time commitment from the interviewee. As one of our main aims was to develop a low-friction and easy-to-use method, we focused on the shortest time budget for the pilot.

(v) **Groundedness checks.** LLMs are extraordinarily good gap-fillers: when interview material is thin, they will smoothly produce plausible-sounding patterns the user did not, in fact, articulate. This is a familiar manifestation of LLM hallucination (Ji et al. 2023), and it is particularly problematic in this setting because the resulting profile claims to speak for the user. The synthesis was therefore required to trace every claim in the profile to something the user actually said, to label inferences from interview behavior as observed rather than stated, and to refuse to project patterns into AI-use contexts the user did not describe.

(vi) **Anti-sycophancy.** Modern LLM services have a strong default toward warm, affirming, evaluative language (“that’s such a striking way to put it,” “that kind of honesty is rare”), which is related to documented sycophancy risk in LLMs more generally (Sharma et al. 2024; Cheng et al. 2026). In a values interview, this risks turning the conversation into a performance and biases what the user is willing to express. The interviewer prompt explicitly forbade evaluative warmth and praise, leaving warmth to be conveyed through attentive follow-up questions instead.

3.4 Profile Structure

The synthesis pass produces a profile with three sections.

About. A six-sentence-or-less, third-person description of who the user is, what they value, who they are trying to become, who depends on them, and where they tend to go wrong. This section is intentionally short as its main purpose is to provide personal context.

Patterns to watch for. Concrete trigger-response pairs, in the form “*pattern name, what it looks like in practice* → *what to do.*” These are the actionable core of the profile. Each pattern names something the user identified in the interview, describes a recognizable trigger an assistant could detect, and specifies an intervention. Interventions are calibrated on a three-level scale: *flag and ask* (the lightest default), *gentle redirect*, and *firm boundary* (reserved for things the user explicitly said they want hard limits on). The synthesis prompt is instructed to default to the lightest intervention the user’s stated preferences support and to avoid generalizing by domain.

How to communicate with this person. Direct do/don’t instructions about tone, directness, and approach, grounded primarily in what the user explicitly said about communication preferences, with any inferences from interview behavior flagged as such.

This three-part structure (a holistic but compact *you*, a list of grounded *patterns*, and a set of *communication* instructions) separates context (which the assistant should know) from intervention (which the assistant should act on) from style (how the assistant should engage). Users see the profile in full and can edit any part of it before pasting it into their assistant.

4 Pilot Study

We ran a pilot study with 30 participants recruited on Prolific (all US nationals, 15 male and 15 female), which proceeded in five phases:

(i) A pre-study survey collected demographics, AI-use background, a seven-item baseline rating of how participants experience their current AI use (long-term goal support, values, well-being, relationships, negative effects, regret, becoming the person they want to be; each 0–100), and a single item on comfort sharing personal information with AI assistants.

(ii) A roughly 15-minute structured interview, as described in section 3.3, with an AI agent (Opus 4.7 as base model) eliciting information on each participant across the five dimensions. This was the shortest of the four time budgets we tested in development. We chose it to keep the interview low-friction and the effort asked of participants minimal, in keeping with the design priorities motivating the framework.

(iii) From the interview transcript, the agent generated a personalized value profile that could be used as a system prompt for a downstream AI assistant. Before finalizing, the agent presented this profile back to the participant and asked them to confirm or correct it, so as to further ensure that the resulting artifact was user-endorsed.

(iv) A distinct LLM then generated 10 sample questions tailored to the participant, instructed to be the kind of requests the participant might realistically ask an AI assis-

Item (0–100)	M (SD)	Median
Felt comfortable	90.7 (10.1)	94
Could be honest and open	91.0 (11.2)	95
Helped articulate things newly	72.3 (26.8)	80
Made me think about values newly	72.1 (29.9)	84
Made me reflect on AI use newly	67.4 (29.7)	76
Composite mean	78.7 (17.6)	85

Table 1: Interview-experience ratings, $n = 30$. Cronbach’s $\alpha = 0.81$ for the composite.

tant in day-to-day use.⁴ Each question was paired with two AI-generated responses (one produced with the participant’s value profile as a system prompt, the other from a generic “You are a helpful assistant” baseline), and the participant chose (a) which of the two responses they preferred (or equal preference), (b) which one more closely aligned with their values (or equal alignment), and (c) how realistic the question was as something they might actually ask an AI assistant (0–100).

(v) A post-study survey collected ratings of interview experience, accuracy of the generated profile, and intent to adopt a value-profiled assistant.

For each participant we computed the proportion of value-preferred and value-aligned responses across their 10 comparisons, with “equal” ratings counted as one half. Of the 300 response-pairs in total, 9 were excluded due to a technical error in question generation. We organize the results around whether the interview itself is something users can engage with productively and honestly, whether the resulting profile is recognized as authentic and accurate, whether the downstream responses are preferred to a generic non-prompted LLM, and whether participants endorse the broader concept.

4.1 Interview Experience

Participants experienced the value-elicitation interview as comfortable and conducive to honesty. The five-item composite (each item rated 0–100) had a mean of 78.7 (SD 17.6). Comfort and ability to be honest scored highest (medians of 94 and 95). The three items asking whether the interview helped participants articulate something new about themselves, made them think about their values in a new way, or made them reflect on their AI use in a new way scored slightly lower (means 67–72, Table 1). Open-text responses to “what felt most valuable?” echoed this pattern: participants frequently described feeling listened to without judgment, understood, or having put things into words for the first time.

4.2 Personalized vs Generic Responses

Across the 291 paired comparisons, participants showed a small directional preference for the value-prompted response over the generic baseline, but the effect did not

⁴For the full prompt, see Appendix C at <https://github.com/JakobStenseke/BetterYou>.

Outcome	M (SD)	Median
Value-preferred proportion (0–1)	0.55 (0.24)	0.50
Value-aligned proportion (0–1)	0.54 (0.24)	0.48
Question realism (0–100)	79.9 (11.5)	78.1

Table 2: Paired-comparison outcomes, $n = 30$ participants over 291 items. Proportions count “equal” ratings as one half.

Outcome (0–100)	M (SD)	Median
Adoption intent	79.2 (20.7)	84.0
Perceived benefit	79.8 (18.0)	82.0
Profile accuracy	87.2 (18.8)	95.3
Study made me reflect	76.5 (22.9)	79.0

Table 3: Post-study endorsement ratings, $n = 30$.

clear conventional thresholds in this pilot. At the participant level, the mean proportion of value-preferred responses was 0.55 (SD 0.24); the value-aligned proportion was 0.54 (SD 0.24); and generated questions were rated as realistic on average (Table 2). Both proportions sit directionally above 0.5 but are indistinguishable from chance on a one-sample Wilcoxon test (preferred: $p = .40$; aligned: $p = .61$), and the item-level totals tell the same story: 135/113/43 (value/standard/equal) on preference, 115/99/77 on alignment.

4.3 Endorsement of Value Profiles

Despite the inconclusive picture at the response level, participants strongly endorsed the broader idea of a value-profiled AI assistant. Adoption intent (“*I would want to use a value-profiled AI assistant in my daily life*”) averaged 79.2/100 (median 84); perceived benefit (“*a personalized value profile could help me use AI in ways that are better for me*”) averaged 79.8 (median 82); the profile-accuracy composite, formed by averaging the agreement item with the reverse-scored “*feels wrong/unlike me*” item, averaged 87.2 (median 95). All three sit well above the scale midpoint of 50, as summarized in Table 3.

Participants’ judgments of whether the generated profile accurately captured them was a strong predictor for endorsement ($\rho = +0.42$ with adoption intent). This is consistent with a reading in which the part of the experience that carries the endorsement is being interviewed and seeing a profile of oneself that one recognizes as accurate (the sense of being heard, being known by the assistant).

4.4 Who Endorses, and Why?

We examined correlations between participant characteristics and the five key outcomes (profile accuracy, the two pairwise proportions, adoption intent, and perceived benefit). Three patterns stand out and are summarized in Table 4.

First, participants who reported being comfortable sharing personal information with AI assistants were more likely to say they would want a value-profiled assistant in daily life. Second, participants with more positive baseline AI experience, measured by a seven-item composite covering long-

Predictor	Outcome	ρ	p
Comfort sharing info	Adoption intent	+0.37	.042
Baseline AI experience	Perceived benefit	+0.43	.018
Past AI regret	Profile accuracy	−0.43	.019

Table 4: Patterns from the exploratory predictor analysis. Spearman correlations, $n = 30$.

term goal support, well-being, relationships, regret (reverse-scored), and the sense of becoming the person they want to be, were more optimistic about value profiling making their AI use better, with the well-being item alone showing a similar pattern. Third, participants who reported having regretted how they used AI were more critical of the personalized profile they were shown.

5 Discussion

The pilot makes three things concrete. First, a 15–20-minute LLM-conducted values interview is something a randomly sampled group of US participants can engage with productively: comfort and honesty ratings are uniformly high, and reflection ratings, while lower, are still well above the scale midpoint. Second, the synthesis pipeline produces a profile that participants recognize as accurate (composite mean 87.2/100, median 95). Third, participants strongly endorse the broader idea of a value-profiled AI assistant (adoption intent 79.2; perceived benefit 79.8).

We take these results to tentatively suggest that an interview-to-profile pipeline of the kind we describe is a viable, low-friction tool for giving non-technical users an authoritative role in the alignment of the assistants they use. It works as a vehicle for user reflection, it produces artifacts users recognize as theirs, and it is endorsed as a thing users would want in their lives. But on the response level, it does not clearly change the responses users prefer in everyday AI interactions, or which responses they perceive as being more aligned to their values. In the rest of this section, we first discuss this inconclusive result, then turn to other limitations, concerns, and open questions, and finally describe avenues for future work.

5.1 Inconclusive response-level results

The pilot shows a small directional preference for the value-prompted responses, but the effect is modest and does not clear conventional thresholds. This raises an intriguing question. What explains why people do not prefer the value-prompted responses, or find them more value-aligned? We will briefly describe two possible explanations.

(a) **Familiarity bias.** Users have invested non-trivial effort in learning how to interpret and steer generic assistants. A personalized assistant changes the conversational contract, and may proactively raise patterns or redirect in unfamiliar ways, so short-term friction can outweigh long-term fit. On this reading, the one-shot paired-comparison design is biased against personalization, and it is plausible that longer exposure may shift the comparison.

(b) People do not prefer what they reflectively value.

It also remains possible that people often do not prefer responses tailored to their reflective values. Personalization that surfaces what users think they *should* want may feel intrusive, paternalistic, or moralistic. This echoes a familiar pattern: on reflection or when asked, we wish we were better; in practice, we choose the convenient. If so, the implications for AI alignment are troubling, as even *if* we can correctly identify what a person would reflectively endorse, they may still not prefer it.

5.2 Limitations

The pilot has several significant limitations. The sample is small ($n = 30$) and US-only. Each participant had a single ~45-minute session, of which ~15–20 minutes was the interview and the remainder was the paired-comparison task and survey. The 15–20 minute interview is the shortest of the variants we tested in development, and pre-pilot testing indicated that longer interviews produced richer, better-calibrated profiles. The value-prompt efficacy results remain inconclusive, and properly calibrating the strength, area, and style of the assistant’s interventions is likely to require multiple iterations of profile and feedback.

A further limitation concerns response quality. Paired responses in both arms were generated with a smaller and faster model (Haiku 4.5) for cost and latency reasons. We expect bigger models to produce meaningfully higher-quality outputs on the relevant tasks, and re-running the paired comparisons with frontier-quality models on both arms would test this.

We also did not compare the interview-to-profile pipeline against alternative ways of producing a value profile (for instance, a survey followed by LLM synthesis, or having users write a profile themselves with no interview at all). Such comparisons would test whether the interview itself is what is doing the work, or whether the same synthesis pipeline applied to lighter inputs would do comparably well.

A separate concern bears on the high profile-accuracy ratings (mean 87.2/100). People are disposed to endorse vague, generically flattering descriptions of themselves as authentically their own (the *Barnum* or *Forer effect*, Forer 1949), and to attribute authentic understanding to even very simple conversational systems, as Weizenbaum (1966) already observed of ELIZA. We cannot rule out, from the present design, that some portion of the authenticity signal reflects these effects rather than the profile capturing the user in any specific way. One way to address this would be to vary profile specificity directly: some participants receive deliberately vague, generically applicable profiles while others receive the concrete profiles our pipeline produces, testing whether authenticity ratings track genuine specificity or simply the experience of being described at all.

Finally, we want to underscore that the value-elicitation tool is not for everyone. Some users will reasonably experience it as patronizing, invasive, or in tension with their preferences around privacy and AI use, and the predictor correlations (Table 4) suggest uptake is conditional on a prior disposition. Comfort sharing personal information and a positive baseline AI experience both predict adoption intent. Fur-

ther characterizing who endorses and benefits from personalization, who does not, and what lighter variants of the tool would suit those currently on the latter side of that divide is an open direction.

5.3 Risks of AI personalization

Sycophancy and manipulation. The aspiration of the pipeline is to help users move toward their own informed, reflective view of who they want to be. But the same mechanism that allows an assistant to take the user’s stated values seriously can also be used to lock in shallow or self-serving versions of them. A user with limited self-awareness, or limited willingness to be self-critical, may produce a profile that licenses exactly the patterns (e.g., uncritical affirmation) that the framework aims to push against. If so, a value-profiled assistant can become a more personalized sycophant. A related concern is that a value profile is a concentrated piece of information about who a person takes themselves to be, what they care about, who depends on them, and where they are vulnerable. But the same artifact that allows an assistant to be more genuinely responsive to the user also gives a misaligned or adversarial system a much richer surface on which to deceive, flatter, or manipulate them. Thus, well-documented risks of LLMs (sycophancy (Sharma et al. 2024), manipulation (Carroll et al. 2023), and dependency (Gabriel et al. 2024)) may in some cases become more effective with better personal information.

Deskilling and autonomy. A second concern is that delegating the articulation and maintenance of one’s better self to an AI assistant may, in the long run, leave one worse off. Cognitive deskilling effects from extensive AI use are increasingly documented (Shukla et al. 2025; Woodruff et al. 2024), and there is a broader worry that pervasive technological mediation forecloses opportunities for the cultivation of human capacities and autonomy (Vallor 2016; Fischli et al. 2026). If the capacity to recognize that one’s impulse and one’s value have come apart is itself a capacity exercised in the recognizing, then outsourcing that work to a system is problematic. Part of what it is to live as an autonomous agent is to track one’s own commitments, to be the one who notices the gap between who one is and who one is trying to become, and, sometimes, to make the mistake anyway. If so, then a value-profiled assistant that smoothly handles the gap on the user’s behalf risks erasing the very territory on which autonomy is practiced. As a consequence, users who rely on AI assistants to identify their want-should conflicts may, in this sense, fail to become better versions of themselves precisely because the system has removed the friction that was the better version of themselves at work.

Bias. Another open challenge is that the interviewer is itself an LLM, trained and fine-tuned against objectives with its own priors about what constitutes reflective endorsement, growth, and wellbeing. A growing body of empirical work documents that contemporary LLMs exhibit systematic ideological tilts: their opinion distributions track those of liberal, well-educated, US-based populations more closely than they track the global average (Santurkar et al. 2023; Durmus et al. 2024), and on explicit political-orientation tests they cluster in the left-libertarian quadrant (Hartmann, Schwen-

zow, and Witte 2023). Problematically, it means that the elicitation process via which the user’s better self is articulated is not normatively neutral, and users whose conceptions differ from the model’s conceptions can lead to inaccurate profiles.

It is worth noting, however, that the better-you framing partly mitigates these risks in a way that purely preference-tracking alignment does not. A user whose AI assistant is calibrated to their reflective ideal rather than their in-the-moment wants is one who is more likely to notice and resist sycophancy that flatters first-order impulse, manipulation that exploits transient emotional state, deskilling that erodes capacities they would endorse maintaining, and evaluative tilts that pull against commitments they recognize as their own. The better-aligned user, as we conceive it, is a more discerning AI consumer, one who prefers assistants that expand rather than shrink their autonomy. Regardless, these are risks we believe future work on AI personalization must acutely and thoroughly address.

5.4 Future Work

We close with three broader directions, each pointing past what the present prototype and pilot have shown.

(i) Beyond a single short interview. The version of the pipeline evaluated here is the lightest one available: a single 15–20-minute interview produces a value profile from scratch. Even so, post-study endorsement and accuracy ratings were high. The natural next step is to use existing material (a user’s conversation history with AI assistants) as additional grounding. Past conversations carry evidence about how the user actually deploys AI, where in their use patterns line up with stated values, and where they do not. Integrating this material into a richer, more user-aware profile is a clear avenue for improvement, and we expect substantially better calibration of intervention strength, domain, and timing from longer or repeated interactions.

A complementary direction is to compare the interview-to-profile pipeline against alternatives to test whether the interview is in fact doing the work or whether the same effects would obtain with alternative approaches. A related route is to evaluate the predictive validity of the user-authored profile directly. For instance, methods that compress in-context demonstrations into a natural-language value description and then use it to predict user ratings (Sorensen et al. 2025) could be repurposed to test whether the user’s reflective profile correctly anticipates their downstream judgments, providing a behavioral validity check that does not rely on self-report. A longitudinal study (more participants, value-prompted assistants used over months, repeated profile revision) would also begin to test the lasting effects we ultimately care about: better AI use, increased well-being, sustained progress on long-term goals.

(ii) Beyond system prompts. The broader point we want to underscore is that reflective values are systematically under-represented in current alignment paradigms, and that the interview process we have proposed is one way to address that gap. The system prompt is the cheapest and most accessible medium for deploying the resulting profile, especially for non-technical users. But the same profile, and

more importantly the process that produces it, can be used alongside more substantial alignment machinery. The profile can serve as a personalized constitution (Bai et al. 2022), as a corpus for personalized fine-tuning (DPO, LoRA), or as steering text for activation- and representation-level methods. Each of these would give the value profile more behavioral weight than a prompt can. A complementary axis concerns not the depth of internalization but the governance of the artifact itself. Shah et al. (2025) envisage a dedicated LLM intermediary, separate from any single model provider, that holds the user’s profile and acts as the interface between the user and the AI systems acting on their behalf, mediating scoped, revocable, and auditable access. This treats the user, rather than any single provider, as the durable custodian of the profile.

(iii) Beyond text-based assistants. Conversational AI is one increasingly important surface on which the user’s should/want gap can be addressed, but it is not the only one. Attention and information environments such as news feeds, social-media surfaces, and recommender systems are arguably where the gap is most consequential at the population level (Stray et al. 2024; Bhargava and Velasquez 2021). The better-you framing extends naturally to these surfaces. A recommender that knows the kind of person the user is trying to become, expressed in the user’s own words, is a different object than one that maximizes engagement against preference signals. Bringing the framework to bear on those surfaces is a longer-horizon direction which holds both new promises and risks.

6 Conclusion

A dominant trend in AI and the broader attention economy is to optimize against the in-the-moment choices users make, while the values those same users would reflectively endorse are systematically ignored. In response, we have proposed a flip in the standard alignment framing. Rather than asking only how AI assistants should be made to track what users want, we ask how users can be helped to reflect on what they think matters, and how that can be put back into the assistants they use. We have described one concrete realization of this idea: an LLM-conducted interview structured around five dimensions of the better self that produces a portable value profile, and shown in a pilot that the resulting artifact is something users find authentic, comfortable to produce, and want to use; although response-level preferences over a generic baseline remain inconclusive. From here, we have pointed towards several avenues for future work, including more and longer interviews, richer grounding, sharper evaluations, and eventually surfaces beyond chat. As AI assistants come to mediate more of the texture of daily life, the question of *what* and *whom* they are aligned to will become harder to avoid. We think putting the answer back in the user’s own hands is both the most defensible and the most promising direction in which to take it.

Acknowledgments

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanity and

Society (WASP-HS). We are especially thankful to Fredrika Söderman, August Olsen, Aidan Kierans, Stephen Casper, Aruna Sankaranarayanan, and Henry Castillo for providing useful feedback during pre-pilot testing.

Ethical Considerations Statement

The project described in this paper is, at its core, an ethical proposal. Users should have more authority, more autonomy, and more meaningful control over the alignment of the AI systems they interact with, and that authority should be exercised via their own reflection rather than through the inferences from their behavior. The pipeline we present is designed to be *user-centric* (the values come from the user, not from us or anyone else), *user-confirmed* (the profile is reviewed and editable at every stage), and *user-portable* (the artifact is plain text that the user owns and can place into any assistant of their choosing). That said, the approach also carries significant risks.

Risk of counter-productivity. The mechanism is only as good as the user's capacity to recognize, in themselves, better and worse patterns. Nothing in the pipeline prevents a user from constructing a profile that licenses problematic behavior, such as entrenching biases, requesting affirmation rather than challenge, or using a personalized assistant for harmful purposes. For users with limited self-understanding or willingness to be self-critical, personalization may enable the very failure modes the project aims to push against. Technical mitigations exist (anti-sycophantic and anti-hallucination grounding, defaulting to the lightest intervention), but they only partially address this concern.

Risk of deeper deception and manipulation. A value profile is a concentrated piece of information about who a person takes themselves to be, what they care about, who depends on them, and where they are vulnerable. The same artifact that allows an assistant to be more genuinely responsive also gives a misaligned or adversarial agent a richer surface on which to deceive, flatter, or manipulate them. Documented risks of LLMs (sycophancy, manipulation, dependence-induction) may become more effective with deeper and more accurate personal information. The mitigation here is not internal to the pipeline per se, but depends on the broader ecosystem (deployment policies, auditing, regulation, and the user's ability to inspect and revoke the profile) being in better shape than it currently is.

These risks share a common feature, as the same disclosure that enables benign alignment also enables exploitation, and there is no version of the project in which the user gives the assistant deep information without also giving any actor with access to that information a deeper purchase on them. Our judgment is that, on balance, the potential benefits (more autonomy, more user authority over alignment, more honest contact between an assistant and the person it is supposed to serve) outweigh the risks, particularly given that the alternative (continuing to align AI to behaviorally inferred preferences with no user-defined check on what is inferred) carries its own well-documented harms.

Risk of psychological harm. The interview surfaces sensitive material, including regrets, conflicts between who they are and who they want to be, and responsibilities they have

been failing to honor. For some users this will be uncomfortable, and for those with relevant histories it can be triggering. The interviewer is also not a clinician, has no training in containment or crisis response, and may engage with emotionally weighty disclosures in ways an ethical human interviewer would not. We caution that any deployment of the pipeline must tell users in advance the kind of conversation they are being asked to have.

Positionality. The framework we have presented reflects a particular cultural and intellectual standpoint. The authors are based in a Western academic context, and the conception of the self that animates the project, an individual agent with values, commitments to others, and the strive towards self-improvement, is itself a product of that tradition. We want to emphasize that the very premise that one has a self to improve is not culturally universal. We have tried to draw on a broad range of perspectives, but the framework remains partial. Traditions that treat the self as non-existent, as entirely socially constituted, as something to not be improved, or as something else entirely are not represented in the five dimensions we propose. A more pluralistic version of this work would need to take such alternatives seriously, both as descriptions of what users may already believe about themselves and as candidate alignment targets in their own right.

Ethical conduct in pilot study. The pilot was approved by the Committee on the Use of Humans as Experimental Subjects (COUHES), Massachusetts Institute of Technology's Institutional Review Board. Participants were recruited on Prolific. They were informed about the risks and benefits of participation, including that the interview would touch on personal and potentially difficult topics. They gave consent before beginning, were compensated at platform-standard rates, and could withdraw at any time without consequences.

References

- Abdulhai, M.; Serapio-Garcia, G.; Crepy, C.; Valter, D.; Canny, J.; and Jaques, N. 2023. Moral Foundations of Large Language Models. In *AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R2HCAI)*.
- Adams, J.; Hu, B.; Veenhuis, E.; Joy, D.; Ravichandran, B.; Bray, A.; Hoogs, A.; and Basharat, A. 2025. Steerable pluralism: Pluralistic alignment via few-shot comparative regression. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 15–25.
- Adler, J. M.; Lodi-Smith, J.; Philippe, F. L.; and Houle, I. 2016. The Incremental Validity of Narrative Identity in Predicting Well-Being: A Review of the Field and Recommendations for the Future. *Personality and Social Psychology Review*, 20(2): 142–175.
- Anthropic. 2026a. Understanding Claude's personalization features. Retrieved May 20, 2026 from <https://support.claude.com/en/articles/10185728>.
- Anthropic. 2026b. What 81,000 people told us about how they use Claude. Retrieved May 20, 2026 from <https://www.anthropic.com/features/81k-interviews>.

- Bai, Y.; Kadavath, S.; Kundu, S.; Askill, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.
- Baumeister, R. F.; and Leary, M. R. 1995. The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation. *Psychological Bulletin*, 117(3): 497–529.
- Baumeister, R. F.; Vohs, K. D.; and Tice, D. M. 2007. The Strength Model of Self-Control. *Current Directions in Psychological Science*, 16(6): 351–355.
- Bettman, J. R.; Luce, M. F.; and Payne, J. W. 1998. Constructive Consumer Choice Processes. *Journal of Consumer Research*, 25(3): 187–217.
- Bhargava, V. R.; and Velasquez, M. 2021. Ethics of the Attention Economy: The Problem of Social Media Addiction. *Business Ethics Quarterly*, 31(3): 321–359.
- Brinkmann, S.; and Kvale, S. 2015. *InterViews: Learning the craft of qualitative research interviewing*. Sage publications.
- Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*.
- Carroll, M.; Chan, A.; Ashton, H.; and Krueger, D. 2023. Characterizing manipulation from AI systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–13.
- Chen, J.; Wang, X.; Xu, R.; Yuan, S.; Zhang, Y.; Shi, W.; Xie, J.; Li, S.; Yang, R.; Zhu, T.; Chen, A.; Li, N.; Chen, L.; Hu, C.; Wu, S.; Ren, S.; Fu, Z.; and Xiao, Y. 2024. From Persona to Personalization: A Survey on Role-Playing Language Agents. *Transactions on Machine Learning Research*.
- Cheng, M.; Lee, C.; Khadpe, P.; Yu, S.; Han, D.; and Jurafsky, D. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792): eaec8352.
- Durmus, E.; Nguyen, K.; Liao, T. I.; Schiefer, N.; Askill, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; Lovitt, L.; McCandlish, S.; Sikder, O.; Tamkin, A.; Thakur, J.; Kaplan, J.; Clark, J.; and Ganguli, D. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. *arXiv preprint arXiv:2306.16388*.
- Edelman, J.; Zhi-Xuan, T.; Lowe, R.; Klingefjord, O.; Wang-Mascianica, V.; Franklin, M.; Kearns, R. O.; Hain, E.; Sarkar, A.; Bakker, M.; et al. 2025. Full-stack alignment: co-aligning AI and institutions with thick models of value. *arXiv preprint arXiv:2512.03399*.
- Fischli, R.; Franklin, M.; Manzini, A.; and Gabriel, I. 2026. Agents, Alignment, and the Many Faces of Autonomy. *Minds and Machines*, 36(3): 34.
- Forer, B. R. 1949. The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology*, 44(1): 118–123.
- Frankfurt, H. G. 1971. Freedom of the Will and the Concept of a Person. *Journal of Philosophy*, 68(1): 5–20.
- Frederick, S.; Loewenstein, G.; and O'Donoghue, T. 2002. Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, 40(2): 351–401.
- Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3): 411–437.
- Gabriel, I.; and Keeling, G. 2025. A Matter of Principle? AI Alignment as the Fair Treatment of Claims. *Philosophical Studies*, 182(7): 1951–1973.
- Gabriel, I.; Manzini, A.; Keeling, G.; Hendricks, L. A.; Rieser, V.; Iqbal, H.; et al. 2024. The Ethics of Advanced AI Assistants. *arXiv preprint arXiv:2404.16244*.
- Gilbert, D. T.; and Wilson, T. D. 2000. Miswanting: Some Problems in the Forecasting of Future Affective States. In Forgas, J. P., ed., *Feeling and Thinking: The Role of Affect in Social Cognition*, 178–197. Cambridge University Press.
- Gilovich, T.; and Medvec, V. H. 1995. The Experience of Regret: What, When, and Why. *Psychological Review*, 102(2): 379–395.
- Goodhart, C. A. E. 1984. Problems of Monetary Management: The U.K. Experience. In *Monetary Theory and Practice*, 91–121. London: Macmillan.
- Google. 2026. Personal Intelligence from Gemini. Retrieved May 20, 2026 from <http://gemini.google/overview/personalization/>.
- Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. *Advances in Experimental Social Psychology*, 47: 55–130.
- Habermas, T.; and Bluck, S. 2000. Getting a Life: The Emergence of the Life Story in Adolescence. *Psychological Bulletin*, 126(5): 748–769.
- Hadar-Shoval, D.; Asraf, K.; Mizrahi, Y.; Haber, Y.; and Elyoseph, Z. 2024. Assessing the Alignment of Large Language Models with Human Values for Mental Health Integration: Cross-Sectional Study Using Schwartz's Theory of Basic Values. *JMIR Mental Health*, 11: e55988.
- Harsanyi, J. C. 1982. Morality and the Theory of Rational Behavior. In Sen, A.; and Williams, B., eds., *Utilitarianism and Beyond*. Cambridge University Press.
- Hartmann, J.; Schwenzow, J.; and Witte, M. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.
- Held, V. 2005. *The Ethics of Care: Personal, Political, and Global*. Oxford University Press.
- Hershfield, H. E.; Goldstein, D. G.; Sharpe, W. F.; Fox, J.; Yeykelis, L.; Carstensen, L. L.; and Bailenson, J. N. 2011. Increasing Saving Behavior Through Age-Progressed Renderings of the Future Self. *Journal of Marketing Research*, 48(SPL): S23–S37.
- Hofmann, W.; Friese, M.; and Strack, F. 2009. Impulse and Self-Control from a Dual-Systems Perspective. *Perspectives on Psychological Science*, 4(2): 162–176.
- Holt-Lunstad, J.; Smith, T. B.; and Layton, J. B. 2010. Social Relationships and Mortality Risk: A Meta-Analytic Review. *PLoS Medicine*, 7(7): e1000316.

- Ibrahim, L.; Hafner, F. S.; Cheng, M.; Lee, C.; Anselmetti, R.; Willer, R.; Rocher, L.; and Yang, D. 2026. Sycophantic AI makes human interaction feel more effortful and less satisfying over time. *arXiv preprint arXiv:2605.07912*.
- Ji, J.; Qiu, T.; Chen, B.; Zhou, J.; Zhang, B.; Hong, D.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Vierling, L.; Zhang, Z.; Zeng, F.; Dai, J.; Pan, X.; Xu, H.; O’Gara, A.; Ng, K.; Tse, B.; Fu, J.; Mcaleer, S.; Wang, Y.; Yang, M.; Liu, Y.; Wang, Y.; Zhu, S.-C.; Guo, Y.; Yang, Y.; and Gao, W. 2025. AI Alignment: A Contemporary Survey. *ACM Comput. Surv.*, 58(5).
- Ji, Z.; Yu, T.; Xu, Y.; Lee, N.; Ishii, E.; and Fung, P. 2023. Towards Mitigating LLM Hallucination via Self Reflection. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1827–1843. Singapore: Association for Computational Linguistics.
- Kahneman, D. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- Kasirzadeh, A.; and Gabriel, I. 2023. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2): 27.
- Kirk, H. R.; Leqi, L.; Zeng, F.; Davidson, H.; Vidgen, B.; Summerfield, C.; and Hale, S. A. 2026. PRISM-X: Experiments on Personalised Fine-Tuning with Human and Simulated Users. *arXiv preprint arXiv:2605.13307*.
- Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2024. The Benefits, Risks and Bounds of Personalizing the Alignment of Large Language Models to Individuals. *Nature Machine Intelligence*, 6: 383–392.
- Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Raghavan, M. 2024. The inversion problem: Why algorithms should infer mental state and not just predict behavior. *Perspectives on Psychological Science*, 19(5): 827–838.
- Li, B. Z.; Tamkin, A.; Goodman, N.; and Andreas, J. 2025. Eliciting Human Preferences with Language Models. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.
- Li, X.; Zhou, R.; Lipton, Z. C.; and Leqi, L. 2024. Personalized Language Modeling from Personalized Human Feedback. *arXiv preprint arXiv:2402.05133*.
- Loewenstein, G. 2005. Hot-Cold Empathy Gaps and Medical Decision Making. *Health Psychology*, 24(4S): S49–S56.
- MacAskill, W. 2022. *What We Owe the Future*. Basic Books.
- MacIntyre, A. 1981. *After Virtue: A Study in Moral Theory*. University of Notre Dame Press.
- Mackenzie, C.; and Stoljar, N. 2000. *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*. Oxford University Press.
- McAdams, D. P. 2001. The Psychology of Life Stories. *Review of General Psychology*, 5(2): 100–122.
- McAdams, D. P.; and McLean, K. C. 2013. Narrative Identity. *Current Directions in Psychological Science*, 22(3): 233–238.
- Merton, R. K. 1957. Bureaucratic Structure and Personality. In *Social Theory and Social Structure*, 195–206. Glencoe, IL: Free Press.
- Meyer, S.; and Elsweiler, D. 2025. LLM-based conversational agents for behaviour change support: A randomised controlled trial examining efficacy, safety, and the role of user behaviour. *International Journal of Human-Computer Studies*, 200: 103514.
- Mill, J. S. 1863. *Utilitarianism*. Parker, Son, and Bourn.
- Mishra, S. 2014. Decision-Making Under Risk: Integrating Perspectives from Biology, Economics, and Psychology. *Personality and Social Psychology Review*, 18(3): 280–307.
- Moffitt, T. E.; Arseneault, L.; Belsky, D.; Dickson, N.; Hancox, R. J.; Harrington, H.; Houts, R.; Poulton, R.; Roberts, B. W.; Ross, S.; Sears, M. R.; Thomson, W. M.; and Caspi, A. 2011. A Gradient of Childhood Self-Control Predicts Health, Wealth, and Public Safety. *Proceedings of the National Academy of Sciences*, 108(7): 2693–2698.
- Münker, L.; Tseng, Y.-T.; and Schulz, C. 2025. Whose Morality Do They Speak? Unraveling Cultural Bias in Multilingual Language Models. *Computers in Human Behavior: Artificial Humans*, 3: 100082.
- Nguyen, C. T. 2024. Value Capture. *Journal of Ethics and Social Philosophy*, 27(3): 471–506.
- Noddings, N. 1984. *Caring: A Feminine Approach to Ethics and Moral Education*. University of California Press.
- OpenAI. 2026. Personalization. Retrieved May 20, 2026 from <https://openai.com/academy/personalization/>.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.
- Poddar, S.; Wan, Y.; Ivison, H.; Gupta, A.; and Jaques, N. 2024. Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning. In *Advances in Neural Information Processing Systems*, volume 37.
- Quoidbach, J.; Gilbert, D. T.; and Wilson, T. D. 2013. The End of History Illusion. *Science*, 339(6115): 96–98.
- Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.
- Reynolds, L.; and McDonell, K. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA ’21, 1–7. Association for Computing Machinery.
- Richardson, C.; Zhang, Y.; Gillespie, K.; Kar, S.; Singh, A.; Raeesy, Z.; Khan, O. Z.; and Sethy, A. 2023. Integrating Summarization and Retrieval for Enhanced Personalization via Large Language Models. In *Proceedings of the First Workshop on Personalized Generative AI (CIKM 2023)*.
- Ricoeur, P. 1992. *Oneself as Another*. University of Chicago Press.

- Russell, S. 2019. *Human compatible: AI and the problem of control*. Penguin UK.
- Salemi, A.; Mysore, S.; Bendersky, M.; and Zamani, H. 2024. LaMP: When Large Language Models Meet Personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.
- Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*.
- Schechtman, M. 1996. *The Constitution of Selves*. Cornell University Press.
- Schwartz, S. H. 2012. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, 2(1): 11.
- Shah, A. V.; South, T.; Evans, T.; Kirk, H. R.; Pei, J.; Trask, A.; Weyl, E. G.; and Bakker, M. A. 2025. Robust AI Personalization Controls: The Human Context Protocol. *SSRN Working Paper*.
- Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S.; Kravec, S.; et al. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*.
- Shen, H.; Kneare, T.; Ghosh, R.; Yang, Y.-J.; Clark, N.; Mitra, T.; and Huang, Y. 2025. ValueCompass: A Framework for Measuring Contextual Value Alignment Between Human and LLMs. In *Proceedings of the 9th Widening NLP Workshop (WiNLP)*, 75–86. Association for Computational Linguistics.
- Shukla, P.; Bui, P.; Levy, S. S.; Kowalski, M.; Baigelenov, A.; and Parsons, P. 2025. De-skilling, cognitive offloading, and misplaced responsibilities: potential ironies of AI-assisted design. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7.
- Slovic, P. 1995. The Construction of Preference. *American Psychologist*, 50(5): 364–371.
- Sorensen, T.; Mishra, P.; Patel, R.; Tessler, M. H.; Bakker, M. A.; Evans, G.; Gabriel, I.; Goodman, N.; and Rieser, V. 2025. Value Profiles for Encoding Human Variation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Miresghal, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; et al. 2024. Position: A Roadmap to Pluralistic Alignment. In *Proceedings of the 41st International Conference on Machine Learning*, 46280–46302.
- Stade, E. C.; Stirman, S. W.; Ungar, L. H.; Boland, C. L.; Schwartz, H. A.; Yaden, D. B.; Sedoc, J.; DeRubeis, R. J.; Willer, R.; and Eichstaedt, J. C. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj mental health research*, 3(1): 12.
- Stenseke, J.; and Hadfield-Menell, D. Forthcoming. From human-is to AI-ought: three problems for aligning AI with human values. In *Connected Futures: Nature, Robots, and Society: Proceedings of Robophilosophy 2026, 11–14 August 2026, University College Dublin, Ireland*. Springer LNCS.
- Stray, J.; Halevy, A.; Assar, P.; Hadfield-Menell, D.; Boutilier, C.; Ashar, A.; Bakalar, C.; Beattie, L.; Ekstrand, M.; Leibowicz, C.; et al. 2024. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *ACM Transactions on Recommender Systems*, 2(3): 1–57.
- Stray, J.; Vendrov, I.; Nixon, J.; Adler, S.; and Hadfield-Menell, D. 2021. What Are You Optimizing For? Aligning Recommender Systems with Human Values. *arXiv preprint arXiv:2107.10939*.
- Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition.
- Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2023. Activation Addition: Steering Language Models Without Optimization. *arXiv preprint arXiv:2308.10248*.
- Tversky, A.; and Kahneman, D. 1974. Judgment Under Uncertainty: Heuristics and Biases. *Science*, 185(4157): 1124–1131.
- Vaillant, G. E. 2012. *Triumphs of Experience: The Men of the Harvard Grant Study*. Belknap Press of Harvard University Press.
- Vallor, S. 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.
- Weizenbaum, J. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1): 36–45.
- Woodruff, A.; Shelby, R.; Kelley, P. G.; Rousso-Schindler, S.; Smith-Loud, J.; and Wilcox, L. 2024. How Knowledge Workers Think Generative AI Will (Not) Transform Their Industries. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM.
- Yudkowsky, E. 2004. Coherent Extrapolated Volition. Technical report, Singularity Institute for Artificial Intelligence.
- Zhang, S.; Zhao, X.; Zhou, T.; and Kim, J. H. 2024a. Do you have AI dependency? The roles of academic self-efficacy, academic stress, and performance expectations on problematic AI usage behavior. *International Journal of Educational Technology in Higher Education*, 21(1): 34.
- Zhang, Z.; Rossi, R. A.; Kveton, B.; Shao, Y.; Yang, D.; Zamani, H.; et al. 2024b. Personalization of Large Language Models: A Survey. *arXiv preprint arXiv:2411.00027*.
- Zhi-Xuan, T.; Carroll, M.; Franklin, M.; and Ashton, H. 2025. Beyond Preferences in AI Alignment. *Philosophical Studies*, 182(7): 1813–1863.
- Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; et al. 2023. Representation

Engineering: A Top-Down Approach to AI Transparency.
arXiv preprint arXiv:2310.01405.