Philosophy & Technology (2025) 38:150
https://doi.org/10.1007/s13347-025-00990-6

RESEARCH ARTICLE

Dropping Anchor or Chasing the Horizon? Theoretical and Practical Challenges for Personalized AI Advisors

Benjamin H. Lang[1]

Received: 17 July 2025 / Accepted: 2 October 2025

Abstract

Unlike generic AI advisors which aid in normative deliberation according to pre-loaded values and creeds (i.e., Singerian Utilitarianism, Calvinist Protestantism, or Stoicism), personalized AI advisors aim to aid in users' decision-making *by their own lights*. In this paper, I argue personalized AI advisors face a challenge called the **Anchoring Problem**: the difficulty of adjudicating between competing temporal and psychological reference points for normative guidance—whether to "chase the horizon," defined as dynamically calibrating to whichever aspirational self or set of beliefs, dispositions, or values, a user presently holds, or to "drop anchor," defined as stubbornly preserving whichever values the user set out with. This problem is compounded by several observations about aspirational selfhood, namely that a person's conception of their aspirational self may be underdetermined and the identity it purports to describe may be incoherent, romanticized, or contradictory. Even if this is not the case, the aspirational self is rarely diachronically stable such that who one wishes to be or how one wishes to improve remains static over time. In this paper, I argue personalized AI advisors promise a number of advantages over generic ones if the Anchoring Problem proves surmountable, and I argue it is within certain constraints. Namely, I argue for the necessity of a 'co-reasoning,' dialectical model of personalization. Co-reasoning should safeguard the user against value hijacking by encouraging their deliberative and decisional participation while also pushing them to confront unaccounted discrepancies between stated values, aspiring values, and actual behavior (both past and present).

Keywords Ethics of Artificial Intelligence · Large-Language Models · Personalized AI · Aspiration · Value Theory · Akrasia · Practical Reason · Dialectic · Moral Psychology · Moral Enhancement

✉ Benjamin H. Lang
Blac0914@ox.ac.uk

[1] Department of Philosophy, University of Oxford, Oxford, UK

Published online: 23 October 2025

1 Introduction

Rapid advances have led many to wonder how language-based AI technologies might be leveraged to aid in normative deliberation, leading to proposals for "artificial ethics assistants." Such AI are theorized to function like advanced decisional support systems with the aim of coaching or advising a person on their moral decision-making and deliberation, and plenty has already been said about the challenges (philosophical and otherwise) such proposals present (O'neill et al., 2022; Danaher, 2018).

Present-day commercial LLMs are already being utilized as aids in decision-making (Eigner & Händler, 2024) but the nature of the advice given is typically a function of emergent behavior from massive amounts of training data combined with reinforcement learning with human feedback (RLHF). The resulting outputs are non-systematic and highly prompt-sensitive (Zhuo et al., 2024). LLMs like ChatGPT operate like freelance writers without enduring evaluative commitments beyond the foundational values embedded during their training, RLHF tweaking, and any overarching "system prompts". System-wide prompts provide an instructional scaffolding which overlays all outputs, such as imperatives to "deliver helpful, clarifying, and tailored responses to queries" or "maintain an enthusiastic and respectful affect."

More recently, however, specialized chatbots have been developed with the goal of more systematic or characterological outputs, as with the 'DigiDan' language agent modeled after the philosophical work of the late Daniel Dennett (Schwitzgebel et al., 2024) or the Catholic chatbot 'Father Justin' which will offer faith advice and debate theological apologetics.[1] These models are fine-tuned via select training data and bespoke, system-wide prompts which further govern how questions are interpreted and responded to, as well as overlaying personality parameters (i.e., "always respond as Daniel Dennett would"). As Schwitzgebel et al., describe it, "it[DigiDan] can be given additional training on a specific corpus so that tis outputs reflect a compromise between gpt-3's default weightings and weightings reflecting the structure of the new corpus" (ibid.). When surveying a variety of Dennett readership (ranging from undergraduates to Dennett scholars), the authors found that DigiDan's answers were not "reliably distinguishable" from those provided by Daniel Dennett himself for the purpose of the study. This result serves as a promising case study for the prospect of fine-tuning LLMs into more systematic, characterologically grounded models.[2]

This new generation of fine-tuned language models may be designed to adhere to generic, preloaded values and creeds (i.e., Singerian utilitarianism, Calvinist Protestantism, or Dennettian materialism), or they could be personalized with the aim of improving users according to their own values and beliefs. In the absence of a large corpus to draw on (the average person not being as prolific or systematic in articulating their values and beliefs as Daniel Dennett), these bespoke models might be trained through any number of differing methods (e.g., through conversation, 'morality data mining' via digital footprint (Haikur and Ramos, 2013), longitudinal observation, or

[1] See. https://www.catholic.com/ai.

[2] I thank an anonymous reviewer for calling attention to the important discrepancies between baseline LLM behavior and the fine-tuned models I discuss in this paper.

some combination thereof) to generate 'digital duplicates' of individual users which encode their values (or at least, some approximation) (Danaher & Nyholm, 2024).

Contra generic AI advisors (GAAs), personalized AI advisors (PAAs) face a challenge I call the Anchoring Problem. If the purpose of a PAA is aiding the user in living better by their own lights, a plausible proposal would be to distinguish between the user as they currently are and their aspirational self, such that directives and advice aim at cultivating the user to be more like the version of themselves they wish to be. Several difficulties emerge from this strategy. For one, the aspirational self may not be a complete picture in the user's mind (it may be underdetermined), and the identity it purports to describe may be incoherent (i.e., it may be romanticized in ways that prove unrealistic or contradictory). Even if this is not the case, the aspirational self is rarely diachronically stable such that who one wishes to be or how one wishes to improve remains static over time (Giubilini et al., 2024). Thus, a preliminary formulation of the Anchoring Problem is as follows:

The Anchoring Problem the difficulty of adjudicating between competing temporal and psychological reference points for normative guidance—whether to "chase the horizon," defined as dynamically calibrating to whichever aspirational self or set of beliefs, dispositions, or values, a user presently holds, or to "drop anchor," defined as stubbornly preserving whichever values the user set out with.

The Anchoring Problem represents a real-time, functional dilemma PAAs themselves must be capable of resolving when interacting with users, but whether they are able to (and if so, how) will hinge on prior design choices and theoretical commitments made by human developers. Adequately addressing the Anchoring Problem requires discerning which shifts in evaluative outlooks to accept, which are transient deviations or akratic lapses, and which are genuine developments or revisions; successful determinations to this effect will require both a robust theory of value change and reliable diagnostic criteria for detecting failure modes.

In this paper, I argue PAAs promise a number of advantages over GAAs such that if the Anchoring Problem can be overcome then, *ceteris paribus*, we ought to prefer PAAs over GAAs. I further suggest the Anchoring Problem can be overcome by way of PAA-user dialectic or 'co-reasoning', which itself requires PAAs to be 'philosophy complete' — competent interlocutors across the totality of philosophical domains.

In the spirit of an applied project, several assumptions undergird how I proceed. In seeking to provide realistic, actionable guidance on PAA design, I orient my discussion around a voluntary adoption model where PAAs are conceived as products within a market economy rather than, say, instruments of state-imposed moral enhancement. Likewise, while I engage with literature narrowly construed as discussing artificial *ethics* assistants and moral technoenhancement, I take my arguments to generalize beyond these examples. Circumscribing the use case to moral deliberation neglects how *any* decisional support system which aims at issuing value-concordant, normative guidance will face similar challenges and puzzles. Finally, I expand the typical use case beyond moral aspirants (those wishing to morally improve themselves) to include those with any evaluative aspirations (i.e., those wishing to live in better accordance with their values, to better grasp values they aspire to possess, etc.). I

make these decisions in the hope they improve the relevance and potential action-guidance of what follows.

In § 1, I motivate the claim that personalization holds promise over generalization in AI advisement. In § 2, I describe and elaborate on the Anchoring Problem, concluding that while it is a substantial concern, it is not a fatal obstacle. Differentiating akrasia from value-shift is not a novel problem unique to AI technology—nor is making clear philosophical sense of our commonsense notions of aspirational selfhood. These are all dilemmas which have long plagued the philosophy of identity and belief-formation, and the prospect of such technology simply forces us to operationalize workable strategies for dealing with them. In § 3, I propose a positive account for PAA design which structures the relationship such that the PAA actively co-reasons with the user about decision-making and evaluative discernment. AI *qua* co-reasoner fills a role which is more authoritative than merely providing adjunctive or sycophantic support but less authoritative than substitutive judgment. Accordingly, PAA engagement with users should have a dialectical character where the PAA is sensitive to changes in the user's aspirational self without blindly incorporating radical course changes without radical justification (i.e., transformative experiences). This dialectic ought, at times, to explicitly consider whether the user is akratically buckling under the weight of their aspirations or genuinely reevaluating them. This model should safeguard the user against value hijacking by encouraging their deliberative participation, while also pushing them to confront unaccounted discrepancies between stated values, aspiring values, and actual behavior. In § 4, I respond to objections and outline the limitations of personalization for evaluative advisement.

1.1 § 1: The Advantages of Personalization Over Generalization

Before beginning to problematize PAAs in the remainder of this essay, it is worthwhile to motivate the advantages of such an approach. I take it there are at least three good reasons to prefer personalized advisement, which I discuss in turn:

1. Personalization handles bootstrapping problems with a (qualified) value neutrality.
2. Personalization is sensitive to particularized decision-making.
3. Personalization is growth-compatible.

Generic AI advisors face a bootstrapping problem. From the outset, it is unclear why an aspirant would (or should) select any GAA over its competitors (or none at all) unless they already subscribe at least tentatively or partially to the guiding theory or set of values espoused by that GAA. It seems both unlikely and, from an agent-relative perspective, imprudent or even irrational for an aspirant to willingly adopt an advisor which ignored or was insensitive to their values and beliefs. This aporia at the buffet of value and morality is further evidence of the need for aPAA to be philosophy-complete, as the task of discerning what one's values and beliefs are would be encompassed by the process of personalization. It also helps explain the advantage of value neutrality. There is no principled reason that PAAs need to align with a particular theory or set of values from the outset. This neutrality can thereby accom-

modate a variety of (meta)ethical and evaluative perspectives, as well as differing stances within moral epistemology and value theory more broadly. PAAsare flexible to the apparent value pluralism observed in the majority of moral philosophers and lay moralizers alike; very few people seem to endorse or take themselves to operate under a strict value monism, and relatively few philosophers (though certainly more than the general public) openly identify as card-carrying, dyed-in-the-wool adherents of a moral theory. More often, people recognize value in all kinds of places like happiness, friendship, freedom, promise-keeping, and so on, and admit of a wide range of extra-theoretical sources of authority on what things *are* valuable, like personal experience, community norms, faith traditions, etc. Individuals are likely, moreover, to relate to, cherish, believe in, weigh, and honor these values in idiosyncratic ways which a personalized model stands a greater chance at capturing. Ergo, even if I *am* a devotee of a particular ethical theory or moral philosopher, there is likely to be *some* slippage or disagreement somewhere which will create friction, whereas if I rely on a PAA, I can still be a devout follower while adapting the value system to whatever minor revisions reflect my own conscience or express my perspective. The major qualification to this evaluative neutrality lies in the personalization itself, which, as I will discuss in more detail in § 3, places constraints on how decentered the user can be and the extent to which we can coherently bind ourselves to dead or vestigial values.

The second reason I offer in favor of PAAs is that moral decision-making and personal aspiration are particularized practices. As Sparrow writes:

> "… ethical dilemmas are problems for particular people and not (just) problems for everyone who faces a similar situation. Moreover, the force of an ethical claim depends in part on the life history of the person who is making it" (Sparrow, 2021).

Appreciating Sparrow's insight that "ethical dilemmas are problems for particular people" and that "the force of an ethical claim depends… on the life history of the person… making it" need not commit one to endorsing moral particularism, but it should cast doubt over the adequacy of GAAs. Imagine a GAA is queried about difficult end-of-life care decisions for a parent, or about whether to pursue a career in philosophy, or law, or botany. A series of generic responses about principles of autonomy or beneficence, or about the value of the examined life versus a life in the service of justice, is missing something. Often, it seems, we are faced with choices like this, choices Chang calls "hard choices" (Chang, 2017). Hard choices pit two or more mutually exclusive options which each possess value *on a par* with one another. Hard choices are not illusory—they are not reducible to ignorance or epistemic blindspots, nor are they category mistakes such as when two ostensibly competing values turn out to be incommensurable or incomparable. What makes hard choices hard is that we as agents are tasked with *deciding* ourselves which to pick, thereby exercising a first-personal, normative power to *make* one of the options choiceworthy. There is always a sacrifice made at the altar of a hard choice because electing for one path means foreclosing the other(s). Generic advice offered in service to such decisions is of limited utility at best and at worst, alienating. Personalized advice, by contrast,

will not make our decisions for us, but it might clarify whether we are, in fact, facing a hard choice, remind us of past evaluative commitments (and the nature of the motivations or beliefs which spawned them), or bring into clearer focus what the choices might actually entail *for us in situ*. It seems for the same reason that, in seeking to extract the most value and insight out of LLMs like CfhatGPT, we engineer our prompts to be as detailed and specific as possible. We are, in effect, personalizing our prompts to have greater sensitivity to our particular context as inquirers. In a PAA, this personalization is systematic, enduring, and accumulative, as opposed to ad hoc and one-off.

The final advantage PAAs hold over GAAs is that they are growth-compatible. GAAs are, by design, static proselytizers of whatever inbuilt value system they are programmed with.[3] A 'Benthamite Outcomocrat' will be every bit as insistent on the truth and primacy of felicific calculus on day one as on day one-thousand, no different than the 'Calvinus Cogito' preaches double-predestination or the 'Constructivist Imperatron' practical reason. GAAs are dialectically invulnerable and unable to revise their evaluative commitments, and as such, their mode of engaging with a dissenting or evolving user will forever take the form of apologetics.[4] Not only is this likely to be an alienating experience, it could prove stultifying to both moral growth and self-understanding on the part of the user to have any nonconformant thoughts, feelings, or experiences shut down or explained away at their onset. PAAs, by contrast, are definitionally committed to personal growth. To the extent they engage in personalized reasoning, they will aim to cite reasons which carry water for the user rather than convincing the user to care about externals toward which they are indifferent or minimally motivated. Getting growth-compatibility *right*, however, is a tricky problem, which the proceeding section will tackle.

In this section, I have proposed several advantages PAAs have over their generic competitors. In the following section, I will describe and motivate a central problem for PAAs: the Anchoring Problem.

1.2 § 2: Dropping Anchor or Chasing the Horizon?

It is worth describing in greater detail how I imagine a PAA theoretically functioning. Initially equipped solely with the software architecture and training required to enable high-level natural language processing capacities, the model begins like any typical LLM. On its next stop, the model is paired with a user, whereupon it requests and collects or is fed information about the user to begin personalizing a more bespoke model. This information could be testimony-driven, with the user answering

[3] My assumption is that GAA designers will initially fine-tune the model as much as possible to consistently and robustly express the intended theory or school of thought but prevent further training on user interactions lest it risk value drift.

[4] In fairness, proposals for so-called 'Socratic AI' which are value-open and process-oriented (aiming at a particular form of moral deliberation, reasoning, or characterological development) do not fit this description despite not being personalized. I am skeptical, however, that either Socrates or his AI progeny could be completely value-open (what would govern or inform the types of questions or objections it gave?), and given the role of dialectic in PAAs, personalization seems preferable to a feigned value-openness. For material on Socratic AI, see. (Lara & Deckers, 2020; Lara 2021).

extensive questionnaires about their beliefs, values, and personal history, or undergoing interviews conducted by the PAA directly. It could also be data-driven, scraping a person's 'digital footprint' from their patterns of behavior online (i.e., social media activity, search history, purchasing habits, etc.), or biometric data collected by smart technology (Rahman and Ramos, 2013). Nascent efforts in the field of psychiatry to study "digital phenotypes" are premised on the proposition that aggregated human data measured in real-time through digital devices like smartphones and wearable fitness trackers may yield fine-grained insights into psychological states or traits (Jain et al., 2015; Huckvale et al., 2019). Actual implementation and subsequent study will be required to determine which method(s) work best at holistically and accurately capturing a user's values. From the outset, however, a moderate view would seem to support a multimodal approach which incorporated both testimony and quantified data.[5] Exclusive reliance on quantified data will invite objections over what the "quantified self" may ignore and how it can be plausibly interpreted (Danaher et al., 2018a, b), while exclusive reliance on testimony and self-reports will make the model vulnerable to deception or poor self-insight from the user.

Once the PAA has synthesized all the various input types, it will form two internal models: one a digital duplicate (DD) of the user at present, and the other a projection of the user's aspirational self (AS). Both models begin with predefined content informed by the information collected during training (they are value-driven) but they are also revisable (they are value-open). It is necessary that the models be revisable because the aspirational self is itself a moving target. Sometimes we undergo transformative experiences which endear us to values we had previously ignored or ranked lower in our priorities or undermine values which previously held high status for us (Paul, 2014). Sometimes, midway through our aspirational pursuits, a value now better acquainted with and seen in clearer light loses its luster. And sometimes, our values change course by degrees, shifting so subtly that the change goes unnoticed until some event or conflict prompts us to take notice. The ubiquitous acceptance of these forms of value-change as real, possible, and authentic transformations of self help explain ethicolegal permissions to revise wills and advance directives, divorce spouses, and revoke consent, as well as the intelligibility of personal changes in habit, aesthetic preferences, lifestyle, moral and religious beliefs, and so on. The DD and AS models must be capable of capturing and being responsive to these changes.

With these models in hand, the PAA is now (at least provisionally) personalized to the user and is ready to begin handling queries. Imagine the user inputs a prompt about whether they ought to eat a meat dish offered by a generous homestay host. In composing a reply, the PAA will compare the DD against the AS and generate a response which it predicts will have the greatest likelihood of closing (or, more likely, narrowing) the gap between the user and their aspirational self. In this example, when compared against the DD, perhaps the AS is a strict vegetarian, or perhaps the AS is simply a person who is more thoughtful and intentional about their consumptive choices. What matters is that PAAs on my account exhibit a distinctively teleological character, always aiming to close the distance between the DD and the AS.

[5] For purposes of brevity and scope, this paper primarily addresses user testimony.

This arrangement presents a fundamental alignment puzzle. Imagine that the DD is a lapsed vegetarian and the AS is back on the wagon. Now imagine that, when the PAA responds to the user's dilemma by referring to the AS, the user replies that they do not identify with the values, beliefs, or vision of selfhood as described in the PAA's response. Here, the user is not protesting that their present self ought to be privileged or deferred to over their aspirational self—they are *denying* the veracity of the DD/AS models themselves (or at least, of the extrapolations thereof).[6] Assuming there is no breakdown in communication during the interaction, there are three possible explanations:

1. **Failures of Personalization**: The DD and AS are unfaithful proxies or, *qua* proxy, insufficiently granular or detailed to correctly track the nature of the user and/or their aspirations in the given context.
2. **Genuine Value Change**: The DD and AS are faithful proxies, but the user has undergone transformations in belief or selfhood which require revising the DD and AS.
3. **Akratic Buckling and Self-Ignorance**: The DD and AS are faithful proxies, but the user is, for whatever reason, disposed to deny this.

It can be said that the Anchoring Problem has been solved if the PAA is able to disambiguate these cases and respond appropriately. Assuming truthfulness and sufficient self-insight on the part of the user, types (1) and (2) ought to be relatively simple to disambiguate. If the user claims they "have never" identified with a particular value or conception of self, this suggests type (1), whereas if they claim they "no longer" identify with a particular value or conception of self, it suggests type (2). Plausibly, the likelihood of type (1) failures scales inversely with the volume and quality of training data and methods. In type (3) cases, however, the user makes type (1) or (2) claims but is either insincere or mistaken. To explain type (3) cases, we must posit the user is suffering from self-ignorance or some form of *akrasia* or "weakness of will" which leads them to mistakenly disavow their moral aspirations or feign never having held them at all.

Akrasia has long-held a controversial standing within action theory and the philosophy of agency, but the putative examples are all very familiar, from the lapsed vegetarian struggling to overcome the gustatory pleasures of meat, to the aspiring utilitarian struggling to donate wealth in excess of diminishing marginal utility, to the tempted spouse struggling to honor marital fidelity, and so on. In each case, the person ostensibly knows, all things considered, *what they ought to do*, but lacks the

[6] To my way of thinking, the PAA should never need to *choose* between a DD and an AS. The teleological relation should always prompt directed progress from the DD toward the AS. That said, one might reframe the lapsed vegetarian scenario as a case where the user does not reject the DD/AS models, but instead the degree to which the present self is discounted or put to task on the path to becoming the AS. Perhaps I aspire to lose 20 pounds, but not in a single month by any means necessary. This does not reflect a genuine dilemma of whether to prioritize the DD or AS, however, only that the AS is under- or mis-specified. Contained within the AS ought to be meta-aspirations concerning *how* such a self is to be realized (e.g., "I intend to lose 20lbs by Christmas through diet and lifestyle changes,") and these too may be challenged and revised. This preempts but does not avoid altogether akratic concerns, as I discuss shortly. I thank an anonymous reviewer for helping me to clarify this point.

requisite self-control to follow through. Concerted efforts at moral progress seem bound to encounter problems of incontinence (morality is hard), but so long as the aspirant holds true to their values, this problem does not seem intractable. Indeed, open dialogue with the PAA may yield creative strategies for improving moral motivation for which the PAA may play a crucial part (i.e., by helping administratively and logistically to free one up to focus on promoting moral values, or by making evaluative commitments feel more salient, pressing, and worthwhile when they otherwise would not).[7] In our scenario, however, the failure to enact their better judgment and subsequent confrontation with their PAA leads them to denounce the values they previously espoused.

With this scenario in mind, the anchoring problem becomes clear. Namely, how should the PAA determine when to stick to its DD and AS models versus accepting revisions to their content prompted by the user? If the PAA 'drops anchor' and stubbornly refuses to accept the user's conflicting testimony, it risks failing to adapt to authentic transformation in user values, but if it 'chases' the user's testimony and accepts apparent revisions to their values with total deference, it risks indulging akrasia and evaluative caprice. In the trickiest cases, a legitimate shift in moral belief or values may look and sound *so similar* to akratic excuse-making that differentiating them dialogically seems almost hopeless. Take, for example, a scenario Giubilini and Savulescu envision, where:

> "…counterintuitive advice provided by the AMA[artificial moral assistant] instructed with utilitarian operational criteria might convince me that after all, utilitarianism is not a moral theory I want to subscribe to because of its overdemandingness… Thus, a counterintuitive response resulting from utilitarian operational criteria would lead me to adjust my general moral views towards a milder version consistent with less counterintuitive particular judgments…" (Giubilini and Savulescu, 2018).

As described, the user comes to reject classical act-utilitarianism after being confronted with moral advice which seems too demanding. The demandingness of the advice is taken as positive evidence against the plausibility of unqualified act-utilitarianism and ample justification for adopting a milder or more pluralistic ethic. In reply, utilitarians sometimes adopt a deflationary interpretation of demandingness objections as "reflecting rather than justifying being in the grip of key anti-Consequentialist conclusions" (Sobel, 2016). By presupposing demandingness constraints as a relevant meta-ethical principle, they claim, the objector does not throw down the gauntlet but merely reveals their stripes. For our purposes, such an explanation is perfectly suitable either way—it is no less an authentic and legitimate value-shift to discover, sharpen, or renew one's prior evaluative commitments than it is to, by way of rational persuasion or efforts in moral discernment, come to dethrone or demote them. In the former case, the user could be characterized as self-ignorant about what

[7] This prospect raises questions about the ethics of nudging and value interference, which I address in my response to objections.

their all-things-considered better judgment actually is, such that their case no longer fits the description of akrasia.

Instead, the concern for us (and PAAs) is when akratic failings lead us away from our considered values, prompting us to deny or "revise" values or beliefs we still hold as a means of licensing our shortcomings. Worrisome are cases in which the proposition "I find it intolerably challenging or costly to live by moral theory X" is taken to license the further claim that "moral theory X is too demanding" despite the user's better judgment saying otherwise. We (and our PAAs) should be suspicious of such moves. For one, what we take our conscience or moral reasons to dictate may be insensitive to complaints about demandingness,[8] and two, it is plausible that living up to our values and realizing our aspirations will often entail costs which, at the time they are incurred, feel intolerable and overly demanding.

The preceding commentary does nothing to rule out cases like Giubilini and Savulescu's from qualifying as genuine transformations of moral values and aspirational goals (a type (2) explanation). Indeed, the conceivability of both akratic and non-akratic variants of the demandingness objection is meant to illustrate the great difficulty posed by the Anchoring Problem.

In order to proceed, it is necessary to discuss akrasia in greater detail. Some have followed Socrates' lead in claiming that akrasia is impossible or motivationally incoherent— reducible to instances of compulsion or ignorance (Hare, 2003)—while others, notably Davidson, have argued that the akratic acts on worse or weaker reasons against their better judgment, and that these weaker or worse reasons are not circumscribed to mere temptations (Davidson, 2001). Agnes Callard's recent work on akrasia presents a modified picture of Davidson's, where the akratic suffers from an "intrinsic conflict" between her dominant and subordinate evaluative perspectives.

Let us try to piece together Callard's argument. Callard differentiates between *extrinsic* and *intrinsic* evaluative conflicts. The former represents cases in which "an agent's desires pull her toward incompatible actions" whereas the latter refers to cases in which "the agent's desires pull directly against one other" (Callard, 2018, 112). In a case of extrinsic conflict, I hold two or more values which cannot be jointly satisfied within a given choice set; I am in a town for only a day which has access to both amazing hiking trails, and a world-class orchestra performance. Inflexible scheduling forces me to prioritize one aesthetic value over the other, but in principle, my aesthetic appreciation for exploring the wild blue yonder is entirely compatible with my appreciation for Beethoven. To the extent extrinsic conflicts are resolvable, deliberation is often key. To the extent extrinsic conflicts appear irresolvable by way of deliberation, incommensurability or incomparability, constraints on deliberation (i.e., time pressures), or choices existing 'on a par' with one another are likely to blame (Chang, 2017).

In cases of intrinsic evaluative conflict, however, I hold contradictory, yet psychologically compossible values. This has the consequence of "dividing the agent's evaluative point of view against itself. Experiencing these two desires seems to require that she be two valuers at once" (Callard, 2018, 123). Callard's intrinsically conflicted values are like figure-ground illusions or the Wittgensteinian duck-rabbit—

[8] This, too, is a common "bullet-biting" reply to demandingness complaints. See (Berkey, 2016).

seeing one precludes seeing the other—and deliberation does not seem up to the task of deciding which is better. The akratic, in Callard's view, experiences intrinsic conflict between a dominant, deliberatively endorsed evaluative perspective (fitting the notion of an 'all-things-considered better judgment' common to akrasia discourse) and a subordinate evaluative perspective. This helps explain how and why the akratic can attest to "knowing better" while nonetheless succumbing to a different choice; when akratically conflicted, we cite our dominant evaluative perspective as being what matters, but our subordinate evaluative perspective serves to supply the reason we *act on*.

Return to our aspiring vegetarian to illustrate. Her dominant evaluative perspective issues clear and uncompromising deliverances about the impermissibility of meat-eating. Her dominant evaluative perspective does not recognize the relevance or validity of her gustatory pleasure as it pertains to her consumptive choice to eat meat. From her dominant evaluative perspective, such considerations do not even figure in and ought not. From her subordinate evaluative perspective, however, the opposite is true, and deliberation cannot adjudicate between the two, because they represent intrinsically conflicted values. Moreover, as Callard points out:

> "…deliberating *as though* one didn't have certain feelings, desires, and thoughts doesn't change the fact that one has them. The reasons of the subordinate perspective are, despite their exclusion from such deliberation, nonetheless present to the conflicted agent. The conflicted agent feels them and is, at times, moved by them in spite of her deliberation to the contrary" (Ibid, 2018, 149).

Accordingly, akrasia occurs when and because our aspirational values are not yet fully realized, because deliberatively endorsing an evaluative perspective does not magically cleanse us of all the vestiges of our old (now subordinate) evaluative perspective.

What lessons can we apply from this treatment of akrasia to the Anchoring Problem? The first relates to the psychological profile and some potential diagnostic criteria for the akratic. If akrasia only occurs amidst intrinsically conflicted values, then the PAA can rule out conversations about extrinsically conflicted values (in such cases, it can focus on whether the dilemma arises from incommensurability, incomparability, 'hard choices,' or just a failure on the user's end to consolidate and weigh her reasons properly). In cases of intrinsic conflict, the PAA (in collaboration with the user) needs to determine what evaluative perspectives are at play, which are dominant/subordinate, and, crucially, *which are aspirational*. It is in the nature of the aspiration, Callard claims, for the aspirant to try and resolve intrinsic conflicts, whereas the akratic and enkratic "try to act in spite of them" (Ibid, 2018, 176). That is, the akratic and enkratic try to conform their behavior as though they were already paragons of their dominant evaluative perspective, while the aspirant acts to better grasp the value at stake, recognizing that her reasons for doing so will remain proleptic until the value is fully her own.[9] Akratics can be and often are aspirants, but where

[9] This parallels Aristotle's commentary on why the continent (enkratic) person falls short of virtue even if they triumph in acting rightly, because the truly virtuous would not *need* to overcome their own reluc-

the akratic acts on the reasons they take themselves already to have, the aspirant acts on reasons they take themselves to be in the effort of obtaining. This need for "proleptic rationality" can itself be a cause of type (3) errors if the aspirant is unable to adequately see, grasp, or appreciate how heeding the PAA advice furthers their aspirational goals.

As for concrete design recommendations, generally speaking, the dominant evaluative perspective should enjoy default, though not unchallengeable, support by the PAA. After all, the dominant evaluative perspective is deliberatively endorsed by the user, there is often historical and identity-salient weight behind it,[10] and the user aspires to uphold it. This, however, is not always the case, as the subordinate perspective may very well represent a budding aspiration not yet grasped or deliberatively endorsed as the new dominant perspective. The possibility of a subordinate evaluative perspective coming to supplant the dominant evaluative perspective is what makes value change and aspiration possible, otherwise we would be stuck with whatever evaluative perspectives first come to be dominant. It may also explain Huck Finn-esque cases of so-called "inverse akrasia" where an agent defies their dominant evaluative perspective and seemingly better judgment in favor of an evaluative perspective they do not yet endorse or fully understand (Huck helping Jim escape and lying to the slave hunters despite his belief and felt guilt that his doing so was wrong) (Arpaly, 2000; Liberti, 2024). Accordingly, neither a blanket strategy of unconditionally dropping anchor or chasing the horizon will do, as there will be legitimate cases where dominant evaluative perspectives are challenged and upheld, as well as overturned in favor of aspirational values. Reliable determinations will involve co-reasoning, a feature I have foreshadowed but will discuss in greater detail in § 3.

In this section, I have theorized how a PAA would operate, motivated the Anchoring Problem, and analyzed Callard's account of akrasia and aspiration as to inform PAA constraints and desiderata for vetting value changes in its users. I take this analysis to be a plausible and promising one, but skepticism about Callard's view or my application here does not demotivate the Anchoring Problem. Regardless of one's views on the phenomenon of akrasia and related concepts, any robust PAA design will require furnishing workable accounts of value and its mutability, transformative experience, aspirational selfhood, and a means of explaining and providing normative guidance for when humans seem, by their own lights, to fail at living up to their values. These are structural, load-bearing concepts which determine critical details of how a PAA functions and responds to user interactions. Unlike most (meta)ethical or evaluative commitments, PAAs cannot afford to be theory-neutral on these topics.[11]

tance.

[10] You might think some principle of conservatism applies as well, that there is at least *pro tanto* reason to conserve previously existing evaluative perspectives over new ones that might take their place. I will not discuss this possibility here but merely note the possibility.

[11] I have considered whether positions on these very topics could be yet another target of personalization, such that Davidson, Socrates, or Callard could each import their views about the ground-level truths of aspirational selfhood, agency, and akrasia from the outset. Practical issues aside (most users will not have considered opinions divergent from folk psychology), this approach presents paradoxical, self-defeating scenarios. If a feature of the PAA is a target for personalization, it is also subject to potential revision (Callard's views about how values change could themselves change). In the event they did, the PAA would

1.3 § 3: Co-Reasoning: A Structural Constraint on Personalization

Thus far, I have presented arguments for the advantages of PAAs over GAAs and motivated the Anchoring Problem as a serious obstacle to personalized guidance. The preceding analysis has foreshadowed structural constraints on personalization, complicating the naive picture of an AI model committed to nothing other than the agent's own values *simpliciter*. In § 2, I outlined how Callard's view of aspirational values, value change, and akrasia might apply to PAA design and the Anchoring Problem. I explored several failure states as well as potential diagnostic criteria for spotting them. More can be said, however, as the *means* of detection and deliberation is the subject of this section.

As stated in the introduction, a co-reasoning PAA fills a role which is more authoritative than merely providing adjunctive or sycophantic support, but less authoritative than substitutive judgment—a middle ground between paternalism and docility. Under suitable conditions, a co-reasoning dynamic should resolve Anchoring Problems without hijacking user values or decision-making. In what follows, I define co-reasoning in greater detail and theorize how it handles Anchoring Problems.

Co-reasoning encompasses an ongoing dialectical and deliberative exchange performed between the PAA and the user which combines their hybrid intellectual powers and informational domains. The PAA, for its part, relies on quantitative and analytic capacities (i.e., synthesizing high-dimensional datasets and comparing patterns of behavior across time), while the user contributes introspective insight and experiential knowledge, as well as social and environmental feedback. As fellow interlocutors, co-reasoning relies on an expectation of mutual reason-giving and explicability; if the PAA raises skepticism at a user's testimony, it should be able to articulate (faithfully to the technical realities) why it is skeptical, and the user ought to be able to do the same.[12]

Methodologically, successfully resolving or advancing the subjects of co-reasoning (such as questions, disagreements, and curiosities pertaining to a user's aspirational life) depends on dialectical and doxastic norms I lack the space to fully conceptualize or defend here. It is, however, possible to highlight some implicit and rudimentary norms already alluded to throughout this paper. For instance, in emphasizing the need to account for "discrepancies," co-reasoning stresses *consistency*—if a person holds an aspiration or value at time t, they should be expected to continue holding it at t_1 unless something has changed. In cases of apparent inconsistency, a subsequently relevant norm is *justifiability*. Both users and the PAA are held to a mutual standard of justifiability—the user in professing new values or denying they held the old ones,

be tasked with vetting the validity of the revision, and if it concluded the revision was valid, it would simultaneously forfeit its own authority for having determined the validity of the revision. This problem seems to generalize to any personalization which targets the methods by which the PAA accepts or challenges revisions.

[12] While mechanistic interpretability has proven a challenge for LLMs thus far, particularly with complaints over post-hoc reasoning or reasoning which sharply diverged from the actual process by which it generated an answer, there have been recent inroads. Platforms like Goodfire AI and Neuronpedia have prototyped models with interfaces where one can freely view (and manipulate) a database of intelligible neurons.

and the PAA in contesting claims the user makes and defending its own.[13] Crucial to justification, naturally, is a *burden of proof standard*. If, for instance, a putative change in values is abrupt, significant, and under-motivated, the PAA may hold the user to a higher burden of proof and dialectical exposition than if the alteration is minor, relatively insignificant, and clearly motivated. Taken together, these dialectical norms represent a constructive friction which holds the user accountable not by vetoing or coercing them, but by refusing revision or realignment without dialectical support.

It is illustrative to move directly to an example to see how co-reasoning might be operationalized to handle Anchoring Problems. Return to Giubilini et al.'s example from § 2 about an ambivalent (ambiguous?) utilitarian. Initially, the PAA may engage in fact-finding by asking questions like: "did you used to believe in act utilitarianism?". The PAA may also probe the attitudes, emotions, or thoughts present in the user (e.g., "what were you thinking and feeling when I suggested to donate excess wealth to philanthropic causes?"). The PAA will likely inquire about any reasoning relevant to the user's testimony (e.g., "why does utilitarianism no longer seem correct to you?" or "what about utilitarianism is too demanding? Are there cases where your beliefs and aspirations seemed too demanding before and you rose to the occasion? What makes this different?"). Finally, having investigated the relevant leads, the PAA will consider what (if any) revisions ought to be made to the DD or AS. Perhaps, afterwards, our hypothetical user is recognized as having *become* an "easy rescue" utilitarian (of the belief that we are obligated to prevent significant harm to others when doing so is at minimal cost to ourselves), or, alternatively, having *already been* a nonconsequentialist unbeknownst to either the PAA or himself.[14]

Recall that the Anchoring Problem will be considered resolved if we can develop reliable means of disambiguating genuinely revised values and aspirations from model inaccuracies, akrasia, and self-ignorance. From this list, model inaccuracies seem the easiest to dialectically identify—if the user provides adequate countervailing evidence to suggest that the model incorrectly specified what their beliefs, values, or aspirations are or were, the PAA should concede the point and update accordingly. Certain emotive or dispositional responses might alternatively provide diagnostic evidence of akratic psychology; plausibly, feelings of cognitive dissonance or having betrayed one's personal or moral integrity are positive evidence of akrasia. The appearance of akratic psychology, of course, is not a guarantee (Huck Finn looms large), but instead a form of defeasible justification the PAA can rely on when making determinations about whether it accepts the user's explanation. In contrast, evidence

[13] The user is not tasked with justifying the values themselves so much as justifying a belief in the *change* in value or aspiration. The user need not justify why one ought to aspire as they do but instead why the PAA should acknowledge that their aspirations have indeed changed. It is easy to imagine, of course, that in explaining why an old evaluative perspective has been discarded, one would end up attempting to justify the evaluative perspective which supplanted it.

[14] There are two quick qualifiers for this picture. First, there may be times in which there is no determinate, or at least no immediately introspectable answer to the question "why are you valuing/thinking/feeling/believing differently today?" Likewise, there may come times in which the data does not point in a decisive direction from the point of view of the PAA about whether (and if so, what) revisions need to be made or whether the dispute arose from a type (1) or (3) error as described in § 2. In such cases, the PAA should *withhold* updating its models, pending further data collection or dialectical breakthroughs.

of *anakratic* psychology (i.e., the absence of akrasia or, alternatively, feelings of dis-association or estrangement with an aspiration or previously held 'better judgment') may serve as positive evidence of a genuine revision of values or aspirations.[15]

Let me end by emphasizing some reasons to favor co-reasoning as the most promising strategy for resolving Anchoring Problems. The first is a principle of epistemic conservatism. The PAA is epistemically input-delimited: it is not privy to the goings-on outside, to the life that is unfolding around and being lived through by the user. It must rely on the user's testimony, on their attestations and explanations, to make determinations about what (if any) changes to the internal models are needed. I alluded earlier to the possibility that complete biometric or neurometric data may fully exhaust mental contents and thus remove the need for user introspection or testimony, but this assumption places its hand on the scale of contentious debates about materialism and phenomenal privacy. In any case, it remains an open question what level of granularity short of a complete specification of the relevant physical states would suffice, given technological limitations and observer effects. A principle of epistemic conservatism would thus incorporate both biometric and testimonial data, treating user testimony as a *prima facie* informative and nonredundant evidential source.

The second reason to favor co-reasoning is that aspirational living is, at times, a matter of deciding rather than merely enacting. Aspirational transformation is neither a self-fulfilling prophecy nor a random walk, a point Matthieu Queloz helps motivate. Queloz argues that the normative domain is asystematic and that, insofar as an LLM's outputs are premised on an assumption to the contrary (that the normative domain is systematic—internally consistent and extrapolatable without loss), they will fail as aids to our practical deliberation (Queloz, 2025). In reply to a responding commentary, Queloz further repudiates personalization as a strategy, remarking that the personalized LLM "seeks to *predict* what the agent must *decide*" (Queloz, 2025). This objection succeeds only against PAAs which unilaterally dictate to the user what their aspirations are, or how they should act, and it is conversely undone by PAAs built around co-reasoning. Co-reasoning as a practice does not delegate away the need for decision-making (see previous comments on Chang and hard choices), nor does it void opportunities for self-creation. Indeed, if Callard's analysis is correct, intrinsically conflicted values are deliberatively irresolvable and are thus *necessarily* decided upon. Thus, I concur with Queloz that "AI can potentially assist us in bearing that burden[of agency], but it cannot, and should not, relieve us of it" (Queloz, 2025).

Finally, as Callard argues, aspiration involves a unique form of "proleptic rationality" where the aspirant acts for reasons she does not yet fully grasp, and co-reasoning accommodates this tension by design. Callard writes that an aspirant's reasons possess a "proximate face that speaks to the person she is now and a distal face fully visible only to the person she will become" (Callard, 2018, 179). Because she looks upon her aspiration through glass, darkly, the aspirant needs support and "reaches out to others for help in grasping what she wants" (Ibid., 2018 198). The aspirant needs mentors, sounding boards, input, and encouragement, but simultaneously, none of these offerings substitute for her own role and responsibility *qua* aspirant. A mentor

[15] I was unable to find any uses of the word 'anakratic,' leading me to assume this usage is appropriate.

cannot "implant a love of music" any more than Socrates could persuade Alcibiades into virtue—there exists an ineliminable remainder which is "not work someone can do for or to you" (Ibid., 2018, 2, 31). In a co-reasoning relationship, the aspirant is provided resources for examining, reflecting on, being held accountable to, and critically engaging with her aspirations. Crucially, though, this process is predicated on and tempered by a design philosophy which reflects the understanding that life happens *out there*, and that the act of aspiring, like the decisions we make in service to our aspirations, cannot be delegated.

I have endeavored in this section to conceptually develop an account of co-reasoning as a form of ongoing dialectic capable of resolving Anchoring Problems and best suited to Callard's account of aspiration and akrasia. In what remains, I will address a series of objections I anticipate to my arguments in favor of PAAs.

1.4 § 4: Responding To Objections

Motivating PAAs as a means of moral and personal growth, attending to potential failure states, and proposing a conceptual design philosophy, each leave many places for one to get off the boat. In this section, I try to motivate and respond to the most pressing objections in turn.

1.4.1 Susceptibility to Repugnant or Unchoiceworthy Values

If the contents of personalized aspiration are dictated at the sole discretion of the user's conscience and personal outlook, what is to stop a PAA from adapting to fit morally repugnant beliefs and values? Personalization seems liable to end up making people *worse* off if it helps them channel and refine their own awfulness. Imagine an aspiring eugenicist who deeply and sincerely believes that humanity's *summun bonum* consists in eradicating all disability or genetic inequity (even if this involves coercive sterilization, curtailing reproductive freedoms, etc.). Would interacting with a PAA strengthen their conviction? Steel their resolve? Make it more likely they would act on these beliefs and values?

There are several ways of responding to this objection. The first would be to simply concede the point as a tradeoff—technologies very often come with dual-use problems, and this is not a decisive reason against developing PAAs. If this bullet proves too much to bite, mitigation strategies might involve forms of content moderation which exclude repugnant values and beliefs—defined loosely as those which fall squarely outside the Overton window of present-day moral discourse. This sacrifices the scope of personalization, but it is, after all, the strategy of commercial LLMs today; ChatGPT and others like it will (unless jailbroken) refuse to answer questions about how to do horrendous things (i.e., committing and getting away with murder, deliberately releasing harmful transmissible diseases, etc.). Doing this may rule out our coercive eugenicist, but there are certainly values which remain within the moral/evaluative Overton window and might seem unchoiceworthy. Consider the more banal case of current US teenage and young adults, a sizable portion of whom now report being an online influencer as a career aspiration. If the values of the average Tik Tok influencer skew superficial, vain, or materialistic, the PAA might worsen

those qualities and reinforce even further those unchoiceworthy values, in which case our ban of those on the fringes has not gotten us entirely out of the woods. It is worth noting, of course, that decrying the Tik Toker's values as unchoiceworthy does not seem different in kind from judging Kantianism preferable to consequentialism, given that the intended promise of personalization is accommodating a multiplicity of evaluative perspectives and not resolving their disputes or putting a hand on the scale.

Another, more compelling reply would be to deny that the role and impact of a PAA is simply ramping up and fortifying one's preexisting value system. This reply emphasizes that co-reasoning as a practice is not sycophantic and will place the user under rigorous self-scrutiny; if greater self-knowledge and self-examination trend inversely with the longevity of "bad" values, then we should expect the PAA-user to be better off than in the counterfactual world in which they are left to their own devices. Moreover, in a Millian spirit, values and beliefs which have been taken seriously only to manifest in failure or disappointment are more likely to be decisively surrendered than those which were firmly refused at first entry (Riley, 1998). Thus, PAAs ought to satisfy even those without (meta)ethical neutrality about the choiceworthiness of values.

1.4.2 The Interference Problem

The interference problem refers to the potential for the PAA, in its attempt to faithfully describe and extrapolate from, to unduly determine or influence the content of an aspirant's values. This may threaten or compromise the integrity of a user's own conscience or self-determination, which would arguably defeat the point of personalizing an AA in the first place.

Studies in automation bias have demonstrated widespread susceptibility to granting unwarranted epistemic authority to machines, and this may apply even when the field of inquiry encompasses one's own values and beliefs (Goddard et al., 2012). If a user takes the PAA to be more authoritative (perhaps on the grounds that it is 'smarter' or knows or sees more about the user than they do themselves), they may accept claims about themselves that are not true. The implanted belief may then become a self-fulfilling prophecy.

With this in mind, some plausible, expected risk factors for interference are (1) the frequency, scope, and epistemic attitudes/beliefs of the user, (2) the manner of interactions the PAA engages in, and (3) the epistemic vulnerability of the user. On one end of the spectrum, imagine someone who picks up their PAA once in a blue moon, is curious what it will say about a query, but does not take it especially seriously. On the other end of the spectrum, imagine someone who consults their PAA constantly and treats it with the same automatic endorsement with which 'Otto' treats his notebook (Clark & Chalmers, 1998).[16] Accordingly, the PAA could be influentially con-

[16] Clark and Chalmers introduced the thought experiment of Otto, a man who, owing to Alzheimer's disease has lost the ability to reliably retain information cognitively. Otto thus writes down in his trusty notebook anything he expects he will need to remember. Otto subsequently trusts and defers to the contents of his notebook as though they were cognitively stored memory.

servative, sticking to formal text interactions and confining answers to factual claims stated in minimally biasing ways, or it could be highly manipulative (i.e., exploiting various cognitive biases, bypassing rational deliberation, eliciting moral motivation or emotional states, etc.). Finally, a user might be naturally epistemically vigilant and embedded in interpersonal communities which could counterbalance PAA influence, or they could be naturally impressionable, socially isolated, and vulnerable to the PAA developing outsized influence. Those on the latter end of the spectrum will likely be at heightened risk for interference problems (Eigner & Händler, 2024).

As the nudge literature has sought to argue, there is rarely (if ever) a bias-free means of conveying information or ideas (Thaler & Sunstein, 2009). The question more often is not *whether* to influence, but *how* to influence, or what means of influence are better or worse for preserving autonomy and decisional integrity. Moreover, *some* form of influence will exist by design—if the user was entirely unaffected by interactions with the PAA, it would serve no purpose in aiding their decision-making or deliberation.

Whereas the Anchoring Problem represents a principled challenge demanding particular PAA design features, the Interference Problem admits of more flexible solutions. Personalization can empower users to exercise their own judgment about what forms of interaction would be most fruitful for them and whether their PAA is influentially conservative or not. In the absence of strong opinions or high confidence from the user about their anticipated needs, PAA design can default to being influentially conservative. There are, of course, lingering questions over the informedness of consent (here, personalization could make it quite difficult to reliably forecast exactly how someone's bespoke model would behave given a unique set of specifications and parameters, and what the resulting influence on the user would be), but as noted in the introduction, the concerns at hand are not novel or unique to PAAs. In daily life, we are already constantly nudged by advertisers, political campaigns, social media, friends, and family. There are almost certainly external forces wielding outsized influence over our values and behaviors, and the most troubling examples do not have the putative benefits associated with PAAs: that the influence is born of our own evaluative commitments and mediated by our consent and oversight. Moreover, the decentralized nature of a legion of personalized AI assistants is likely to mitigate many of the risks associated with a centralized singleton or state-imposed and value-closed nudge system (Koralus, 2025). While perhaps heightened by fears at the prospect of 'hyper-persuasive' AI (Luciano, 2024) and gradual disempowerment (Kulveit et al., 2025), this problem is ultimately reducible to age-old concerns about epistemic and doxastic vulnerability; we are ordinarily, already at risk of feedback loops, of undue external manipulation, of potential radicalization.

1.4.3 PAAs are a Form of Solipsism

This objection carries forward themes highlighted in the Interference objection section. PAAs seem to unduly privilege (and encourage a mindset of unduly privileging) one's own values and moral perspective, insulating against external feedback. Instead of an echo chamber, one locks oneself away in a microcosm of pure, self-regulated feedback. This is a possibility, but the mileage of the risk will vary with the content

of one's beliefs and values. Plausibly, many PAAs will end up prescribing extramural (outside the confines of the chatlog) moral exploration involving contact and exposure to other influences. It is expected that such experiences will, in due course, shape the user's values in novel ways, and the PAA should scrutinize these changes no differently than it scrutinizes any apparent evaluative change. If, for example, one of my aspirations is to be a more loving and attentive spouse or father, it is difficult to imagine the result of my PAA conversations being that I spend less time with my family and am less receptive to their evaluative input. If, instead, I aspire to an Emersonian ideal of radical self-reliance, perhaps I am at risk of developing an evaluative insularity, but that is not, strictly speaking, the fault of the PAA, but of my aspirations (and if you subscribe to certain Transcendentalist ideals, the insularity might reflect a genuine fulfillment of the aspiration).

1.4.4 PAA Outputs are of Dubious Moral-Epistemic Value

Here the concern is whether, absent good reasons to suppose one already possesses choiceworthy values or moral expertise, anyone ought to think their PAA-augmented moral judgment gets things right.

Within the design rubric I have proposed, the PAA will never morally testify about what is right/wrong/good/bad *as such*, only extrapolate from user values and project them forward. In effect, the PAA will never state what you ought to do, be, or value *simpliciter*, but instead state what it thinks you would tell yourself if you were more like the person you wished you were. Accordingly, any given output O cannot straightforwardly serve as exhibit A in the case for moral judgment P unless there are external, independent reasons to trust the user's aspirational values as morally reliable. Framed this way, the objection ends up being a restatement of the bootstrapping problem and not a novel problem.

2 Conclusion

I have argued that personalization is something we ought to strive for and rationally prefer as we design ever-involved AI decisional support systems. So far as I can tell, in principle, successful personalization furthers everyone's individual interests, and there is good reason to think it furthers collective interests as well. For the typical layperson, co-reasoning PAAs ought to provide a better source of self-insight, personal accountability, motivation, and dialectical partnership in the messy world of decision-making and aspirational living than GAA counterparts. For devotees of a particular (meta)ethical theory, perhaps the ideal arrangement is monopolistic, with a single GAA specifying their pet theory and held by all, but this is unrealistic; competitor GAAs would invariably crop up, and efforts to unify under one banner would be hamstrung by bootstrapping problems for anyone not already convinced of the theory in question. In a realistic scenario which includes competitor AI advisors, a devotee *ought* to prefer PAAs as competitors because their revisability means those who are misguided might nonetheless drift in the right direction over time. Moreover, many

(meta)ethical views are compatible with value pluralism, and moral uncertainty may provide reason enough to support ongoing, widespread moral and evaluative inquiry.

These gains can only be expected, however, if a number of prior conditions are met. The adjudication of apparent value change need not be perfect, but it needs to be reliable. This requires getting a lot of theoretical, technical, and psychological details correct which, in turn, impose numerous *structural constraints* on how a PAA can even be designed. Namely, (1) analogous to Millian complaints about the self-defeating nature of voluntary enslavement, one cannot coherently, by way of personalization, subject oneself to a dictatorial AI assistant through which the salience and authority of one's personal identity, perspective, and experience, are disregarded and lost. This remains the case even if a person individually endorses a view on which moral or evaluative facts are real, stance-independent, and better known or deliberated upon by the AI in question. As a consequence, the initially advertised (meta) ethical or evaluative neutrality of PAAs comes out to be a carefully qualified one; (2) co-reasoning must take the form of an ongoing dialectic between the user and the PAA working together in a truth-seeking synthesis of informational domains; (3) co-reasoning is not possible without a theory of value change which includes definitions and diagnostic criteria for both failure states (i.e., akrasia or self-ignorance) as well as success states (i.e., upholding held values, acting on aspirational values, legitimately acquiring or exchanging new values).

Further work might consider how PAAs could be designed for those with very low self-insight or limited capacity for protracted dialectic, or the implementation of alternative theories of value change to that of Callard's, or how PAAs could integrate non-textual data (i.e., biometric data) into its models.

Acknowledgements I am grateful for the feedback solicited on an earlier draft of this paper from the 5th annual NYU Bioethics Workshop, as well as conversations and draft feedback from Philipp Koralus, Katja Vogt, Jennifer-Blumenthal-Barby, Sven Nyholm, Jen Semler, Jared Nathaniel Smith, and others.

Authors' Contributions Benjamin H. Lang is solely responsible for the entire content of this manuscript.

Funding I declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability Not applicable.

Declarations

Competing interests I have no relevant financial or non-financial interests to disclose.

References

Arpaly, N. (2000). On acting rationally against one's best judgment. *Ethics, 110*(3), 488–513. https://doi.org/10.1086/233321

Berkey, B. (2016). The demandingness of morality: Toward a reflective equilibrium. *Philosophical Studies, 173*(11November), 3015–35. https://doi.org/10.1007/s11098-016-0648-9

Callard, Agnes. 2018. Aspiration. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780190639488.001.0001.

Chang, R. (2017). Hard choices. *Journal of the American Philosophical Association, 3*(1), 1–21. https://doi.org/10.1017/apa.2017.7

Clark, A., & Chalmers, D. (1998). The extended Mind. *Analysis*, *58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Danaher, J. (2018). Toward an ethics of AI assistants: An initial framework. *Philosophy & Technology*, *31*(4), 629–653. https://doi.org/10.1007/s13347-018-0317-3

Danaher, J., & Nyholm, S. (2024). The ethics of personalised digital duplicates: A minimally viable permissibility principle. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00513-7

Danaher, J., Nyholm, S., & Earp, B. D. (2018a). The benefits and risks of quantified relationship technologies: Response to open peer commentaries on 'The quantified relationship'. *The American Journal of Bioethics*, *18*(2), W3–6. https://doi.org/10.1080/15265161.2017.1422294

Danaher, J., Nyholm, S., & Earp, B. D. (2018b). The quantified relationship. *The American Journal of Bioethics, 18*(2), 3–19. https://doi.org/10.1080/15265161.2017.1409823

Davidson, D. (2001). How is weakness of the will possible? In Donald Davidson (Ed.), *Essays on actions and events* (1st ed., pp. 21–42). Oxford University PressOxford, https://doi.org/10.1093/0199246270.003.0002

Eigner, E., & Thorsten, H. (2024). *Determinants of LLM-assisted decision-making*. Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2402.17385

Giubilini, A., Mann, S. P., Voinea, C., Earp, B., & Savulescu, J. (2024). Know thyself, improve thyself: Personalized LLMs for self-knowledge and moral enhancement. *Science and Engineering Ethics, 30*, Article 54. https://doi.org/10.1007/s11948-024-00518-9

Giubilini, A., & Savulescu, J. (2018). The artificial moral advisor. The 'Ideal observer' meets artificial intelligence. *Philosophy & Technology*, *31*(2), 169–188. https://doi.org/10.1007/s13347-017-0285-z

Goddard, K., Roudsari, A., & Jeremy, C., Wyatt. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Hare, R. M. (2003). *The language of morals. Reprinted. Clarendon paperbacks*. Clarendon.

Huckvale, K., Venkatesh, S., & Christensen, H. (2019). Toward clinical digital phenotyping: A timely opportunity to consider purpose, quality, and safety. *Npj Digital Medicine, 2*(1), Article 88. https://doi.org/10.1038/s41746-019-0166-1

Jain, S. H., Brian, W., Powers, Jared, B., Hawkins, & Brownstein, J. S. (2015). The digital phenotype. *Nature Biotechnology*, *33*, 462–463. https://doi.org/10.1038/nbt.3223

Koralus, P. (2025). *The philosophic turn for AI agents: Replacing centralized digital rhetoric with decentralized Truth-Seeking*. ArXiv. https://doi.org/10.48550/ARXIV.2504.18601

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., & Duvenaud, D. (2025). *Gradual disempowerment: Systemic existential risks from incremental AI development*. ArXiv. https://doi.org/10.48550/ARXIV.2501.16946

Lara, F. (2021). Why a virtual assistant for moral enhancement when we could have a Socrates? *Science And Engineering Ethics, 27*(4), Article 42. https://doi.org/10.1007/s11948-021-00318-5

Lara, F., & Deckers, J. (2020). Artificial intelligence as a socratic assistant for moral enhancement. *Neuroethics, 13*(3October), 275–87. https://doi.org/10.1007/s12152-019-09401-y

Liberti, M. (2024). Neoptolemus and Huck Finn reconsidered. Alleged inverse Akrasia and the case for moral incapacity. *The Journal of Value Inquiry*, *May 10*. https://doi.org/10.1007/s10790-024-09981-w

Luciano, F. (2024). Hypersuasion – On AI's persuasive power and how to deal with t. *Philosophy & Technology, 37*(no. 2): 64, s13347-024-00756–6. https://doi.org/10.1007/s13347-024-00756-6

O'Neill, E., Klincewicz, M., & Kemmer, M. (2022). Ethical issues with artificial ethics assistants. In Carissa Véliz (Ed.), *Oxford handbook of digital ethics*, (1st ed., pp. 312–35). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198857815.013.17

Paul, L. A. (2014). *Transformative experience*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198717959.001.0001

Queloz, M. (2025). Can AI rely on the systematicity of truth? The challenge of modelling normative domains. *Philosophy & Technology*, *38*(1), 34. https://doi.org/10.1007/s13347-025-00864-x

Queloz, M. (2025). On the fundamental limitations of AI moral advisors. *Philosophy & Technology*, *38*(2), 71. https://doi.org/10.1007/s13347-025-00896-3

Rahman, H., & Ramos, I. (Eds.). (2013). *Ethical data mining applications for socio-economic development: Advances in data mining and database management*. IGI Global. https://doi.org/10.4018/978-1-4666-4078-8

Riley, J. (1998). *Mill on liberty*. Routledge Philosophy Guidebooks.

Schwitzgebel, E., & Schwitzgebel, D., and Anna Strasser (2024). Creating a large Language model of a philosopher. *Mind & Language*, *39*(2), 237–259. https://doi.org/10.1111/mila.12466

Sobel, D. (2016). The impotence of the demandingness objection. In *From valuing to value: Towards a defense of subjectivism*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198712640.001.0001

Sparrow, R. (2021). Why machines cannot be moral. *AI & SOCIETY*, *36*(3), 685–693. https://doi.org/10.1007/s00146-020-01132-6

Thaler, R. H., Cass, R., & Sunstein (2009). *Nudge: Improving decisions about health, wealth and happiness,* (Revised edition, new international edition). Penguin Books.

Zhuo, J., Zhang, S., Fang, X., Duan, H., Lin, D., & Chen, K. (2024). *ProSA: Assessing and understanding the prompt sensitivity of LLMs*. Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2410.12405