Evolutionary debunking and value alignment

Michael T. Dale (Hampden-Sydney College) and Bradford Saad (Global Priorities Institute, University of Oxford)

Global Priorities Institute | May 2024

GPI Working Paper No . 11-2024

Evolutionary Debunking and Value Alignment

Michael T. Dale (Hampden-Sydney College)
Bradford Saad (Global Priorities Institute, University of Oxford)[1]

Abstract: This paper examines the bearing of *evolutionary debunking arguments*—which use the evolutionary origins of values to challenge their epistemic credentials—on *the alignment problem*, i.e. the problem of ensuring that highly capable AI systems are properly aligned with values. Since evolutionary debunking arguments are among the best empirically-motivated arguments that recommend changes in values, it is unsurprising that they are relevant to the alignment problem. However, how evolutionary debunking arguments bear on alignment is a neglected issue. This paper sheds light on that issue by showing how evolutionary debunking arguments: (1) raise *foundational challenges* to posing the alignment problem, (2) yield *normative constraints* on solving it, and (3) generate *stumbling blocks* for implementing solutions. After mapping some general features of this philosophical terrain, we illustrate how evolutionary debunking arguments interact with some of the main technical approaches to alignment. To conclude, we motivate a parliamentary approach to alignment and suggest some ways of developing and testing it.

Keywords: alignment problem; evolutionary debunking arguments; philosophy of artificial intelligence; machine ethics; AI safety; the evolution of morality; catastrophic risks; constructivism; moral realism; moral parliament; normative uncertainty

Word count: 9,660 (main text)

Acknowledgments: For helpful feedback, we are grateful to Adam Bales, Spencer Case, Andreas Mogensen, and Elizabeth O'Neill.

Funding: This work was supported by the Eindhoven University of Technology, Hampden-Sydney College, Sentience Institute, University of Oxford, and Utrecht University.

[1] Author order is arbitrary.

1. Introduction

The alignment problem is that of ensuring that artificially intelligent systems (AIs) are properly aligned with an appropriate set of values. A growing contingent of AI researchers believe that this problem is difficult to solve and that continuing the current trajectory of AI development without solving it would pose significant near-term catastrophic risks, including risk of human extinction.[2]

The alignment problem decomposes into a normative problem and a technical problem.[3] The normative problem concerns what values we *should* align such systems with.[4] The technical problem concerns how to ensure that AIs—and, in particular, future AIs that would, if unaligned, pose catastrophic risks—act in accordance with an intended set of values.[5] There is a burgeoning research program focused on the technical problem.[6] In contrast, the normative problem has received relatively little attention.

Technical alignment proposals often presuppose that AIs should be aligned with human preferences. Some of this normative convergence is merely verbal, as 'human preferences' ('human values') is a flexible expression whose meaning is often left at an intuitive level. Thus, conditional on an imperative to align AIs with human preferences, there is a debate to be had about how that notion should be understood and operationalized. However, there is reason to worry that this entire class of approaches rests on a mistake: it is far from obvious that human preferences are what we should align AIs with.[7] There are, after all, sociopaths and sadists in our

[2] For a recent expert survey, see Grace et al. (2024). For reasons to think AI poses catastrophic risks, see, e.g., Bales et al (2024), Bengio, Hinton, et al. (2023), Bostrom (2013, 2014), Carlsmith (2023; forthcoming), Cotra (2021), MacAskill (2022), Ngo et al. (2023), Ord (2020), and Russell (2019). In 2023, the Center for AI Safety released a statement advocating the mitigation of extinction risk from AI as a global priority alongside preventing nuclear war and pandemics. Signatories included AI scientists, and employees from OpenAI, Anthropic, and Google DeepMind, as well as notable philosophers such as David Chalmers, Daniel Dennett, Hilary Greaves, Peter Railton, and David Wallace.

[3] See, e.g., Gabriel (2020: 412-413).

[4] There is also a question of whether we should aim to align such systems at all. If some such systems are moral patients, they may have rights (Schwitzgebel & Garza, 2015) that would inevitably be violated during the process of alignment. For example, such systems may be owed a measure of autonomy in determining their own values—see Schwitzgebel & Garza (2020); cf. Chalmers (2022: Ch. 18). Or alignment might be incompatible with respecting such systems' welfare (cf. Shulman & Bostrom (2021), Saad & Bradley (2022), and Goldstein & Kirk-Giannini (2023)). Even if AIs generally lack moral status, we may have reason to afford them some control of their values so as to foster cooperation (Friederich, 2023). Alternatively, the risk profile of such systems could render them inadvisable to create, especially if we have safer and similarly beneficial alternatives—cf. Drexler (2019) and Bengio (2023). Although these considerations are important and neglected in the alignment literature, we hereafter set them aside.

[5] In what follows, we will be chiefly concerned with such AIs, though much of the discussion will also apply to AIs for which the alignment problem arises but which wouldn't pose catastrophic risks if unaligned.

[6] For overviews, see Anwar et al. (2024) and Krakovna (2022).

[7] See, e.g., Christiano (2018), Daley (2021), Peterson (2019); cf. Soares & Fallenstein (2017: fn1).

ranks. Indeed, humanity's moral track record suggests that aligning AI with human preferences could just as easily cause catastrophe as prevent it. Nor is it obvious what we should align AIs with instead, though we'll survey some possibilities.

But our task here is not to solve the normative part of the alignment problem. Instead, we'll show that—on a wide range of technical, normative, and metaethical approaches—solving the alignment problem requires carefully addressing *evolutionary debunking arguments* (EDAs). EDAs contend that natural selection shapes at least some of our values in a way that debunks them, i.e. shows them to be epistemically defective. EDAs have received sustained attention within epistemology and metaethics over the last two decades.[8] In contrast, EDAs have gone undiscussed in literature on the alignment problem.[9] Even so, since EDAs recommend drastic revisions to human values, it should be unsurprising that EDAs bear on alignment.

This neglected connection between EDAs and alignment merits investigation. EDAs offer rich test cases. If a proposed approach to alignment cannot handle EDAs, it is not a viable and complete solution. And if a system becomes unaligned in response to EDAs, it is likely susceptible to other forms of alignment failure as well. So, understanding alignment in the context of EDAs may illuminate the broader and poorly understood class of factors that could produce drastic value changes in AIs—these factors might include other revisionary philosophical arguments. More generally, investigating how EDAs bear on alignment promises to reveal hitherto unnoticed contours of the alignment problem that should guide our search for solutions. The task of this paper is to initiate that investigation.

To be clear, the standing of EDAs is extremely controversial.[10] However, to bear on alignment, EDAs need not be sound: they may influence what values we should align AIs with by affecting how confident we should be in different moral views. Nor does the bearing of EDAs on alignment presuppose that morality has a basis that is independent of our attitudes: some EDAs lack this presupposition, and AIs could change their values in response to EDAs with that presupposition regardless of whether it is correct.

[8] See, e.g., Street (2006); Joyce (2006); de Lazari-Radek & Singer (2012), Wiegman (2017), and Jaquet (2022).

[9] There is, however, precedent for discussing other philosophical problems as sources of misalignment in powerful AI Systems—see, e.g. Dai (2010), Bostrom (2014: 13), and Soares & Fallenstein (2017). For a general discussion of goal change as a topic in AI safety, see Herd et al. (2018).

[10] For instance, some argue that EDAs overgeneralize to other beliefs and that this is a reason to reject them (Kahane, 2011; Berker, 2014). For an overview of different sorts of EDAs, see Korman (2019).

After providing background on the alignment problem and EDAs (§§2-3), We'll explore how EDAs yield *foundational challenges* to posing the alignment problem (§4), *normative constraints* on solving it (§5), and potential *stumbling blocks* for implementing solutions (§6). Finally, we'll offer a provisional assessment of EDAs bearing on some technical alignment proposals (§7) and sketch our favored, parliamentary approach (§8).

2. Candidate Alignment Targets

In what follows, we will just focus on the values with which we should align AI behavior, leaving to the one side questions about what values AIs should internally represent.[11]

We can distinguish several categories of targets, i.e. candidate values for alignment. *Psychological targets* are values that can be specified in descriptive psychological terms. Such targets are characteristic of widely advocated *standard solutions* to alignment, which construe targets as human preferences.[12] More precisely, standard solutions claim that we should create AIs that *do* what we want, not that we should create AIs that act as if they wanted what we want. Different standard solutions cash out "human preferences" in different ways. They can be restricted to actual preferences or to preferences humans would have under idealizations such as full information and procedural rationality. There is also room for non-standard psychological targets—for instance, the preferences of all moral patients.

Procedural targets can be specified descriptively in terms of the output of a procedure. Procedural targets include values selected directly through a referendum or indirectly via an elected individual or delegation. Some procedural targets are also psychological targets. For instance, targets concerning what humans would prefer if they were procedurally rational are both procedural and psychological. The same goes for procedures that aggregate human preferences. Since human preferences vary, psychological targets would probably need to be precisified in terms of some such procedure.[13]

Direct normative targets can be specified directly using normative concepts.[14] Candidates include being beneficial to humans, being beneficial to all sentient beings,

[11] At the same time, we will take values to be individuated conceptually by their specifications.

[12] See Ratoff (2021).

[13] See Gabriel (2020). For a critical overview of some aggregationist procedural approaches in AI ethics, see Baum (2020); cf. Brennan (2021).

[14] N.B. This leaves open whether these concepts can be indirectly accounted for via reference to human preferences, as some antirealist metaethical views maintain.

maximizing utility, treating everyone equally except insofar as inequality benefits the least well off, acting with compassion and wisdom, and acting in ways that no one can reasonably reject.[15]

Indirect normative targets indirectly specify the values that are to be directly reflected in AI behavior. Such targets include whatever values the true moral theory endorses, whatever values the evidence most justifies, whatever values are picked out by certain normative concepts (such as the concept of good), whatever values would be advanced by an exemplar of virtue, and whatever values would be selected via a fair procedure.[16]

Different sorts of targets enjoy different motivations. Insofar as there is a solution to the normative part of the alignment problem, it would seem that there must be a solution in a direct normative target. However, it's a vexed question what that target is. Indirect normative solutions suggest a way forward: we can pass the criterial buck to an indirect normative criterion that can be expected to serve as a suitable proxy for a direct normative criterion and the ethical theorizing buck to an AI system that promotes the latter by promoting the former. Yet there is a formidable technical task of aligning AI with normative targets: many normative notions resist formalization and measurement. In contrast, aligning AI with human preferences seems relatively straightforward: we have well-established social science methods for measuring preferences, formal tools for representing preferences, and machine learning techniques for extracting human preferences. However, the tractability of this approach is no guarantee of its correctness.

Correct alignment targets may vary with context.[17] If your personal assistant AI can only significantly affect you, there may be nothing wrong with aligning it with your preferences. But systems that can shape the entire future of humanity should not be beholden to the preferential whims of any one person. These two cases fall on a spectrum. In between are cases in which, for example, AIs act on behalf of democratic bodies or corporations. In such cases, a natural suggestion is that AIs should be aligned with aggregated preferences of certain stakeholders.

The categories of psychological, procedural, and normative targets mark important theoretical joints. Even so, candidate targets will often bear intimate relationships with multiple categories. For instance, because human preferences often have normative content, so too may psychological targets. And human psychological states (e.g. happiness) may realize a moral

[15] For an overview of moral theories, see Driver (2022).

[16] See Bostrom (2014: 13), Christiano et al. (2018), Hobbhahn et al. (2022), Gabriel (2020), Rawls (1971), and Yudkowsky (2004).

[17] Cf. Liscio et al. (2022).

property (e.g. goodness) that figures in a normative target. More generally, targets in one category may have other targets as contents, aspects, proxies, side constraints, background conditions, and realizers.

While different kinds of alignment targets are subject to different motivations and pitfalls, differences between targets need not be as stark as they may initially appear. As noted, there may be close connections between candidates in different categories, and different kinds of targets may agree in behavioral import or be correct relative to different contexts. However, none of this suggests that choices between different targets or different kinds of target is unimportant. Ample room remains for divergence in behavioral import within and across categories. A better moral is that we should expect EDAs that bear on targets in one category to have spillover effects to targets in others. As we will see, this dampens any hopes of solving the alignment problem without having to confront EDAs.

3. Evolutionary Debunking Arguments

Consider the following thought experiment, inspired by Joyce (2006). Let's say that you believe that baobab trees are native to Australia. However, one day, you discover that you were given a pill that was designed to make you believe that baobab trees are native to Australia. As a result of learning this information, you should not necessarily conclude that your belief about baobab trees is false, for it could still be true that baobab trees are native to Australia (perhaps by coincidence). But you should conclude that your belief is *unjustified*, as you have learned that the process that generated the belief was non-truth-tracking.

In this example, the information about the pill serves as an undercutting defeater: by revealing that the belief was produced by a process that is not truth-tracking, that information challenges your basis for maintaining that belief.[18] Evolutionary debunking arguments work in a similar fashion. They introduce a causal claim: some (or all) of our values were significantly shaped by natural selection. Next, they put forward an epistemic claim: natural selection does not track normative truth. The result: our values are unjustified.

Local EDAs target particular sets of values (Rowland, 2019). To give a few examples, de Lazari-Radek and Singer (2012) argue that evolution undercuts egoistic values, but not impartial values concerning universal benevolence. Kahane (2014) examines EDAs concerning well-being. Wiegman (2017) attempts to debunk our values concerning retribution. Greene

[18] See Pollock (1987).

(2014) claims that deontological values are less morally justified than consequentialist values.[19] Finally, Jaquet (2022) contends that our values stemming from our speciesist tendencies are unjustified.

Perhaps the most attention has gone to the *global* EDAs that target *all* values within a certain domain such as morality. Joyce (2006) contends that our normative values stem from an innate capacity that evolved to foster cooperation. If this is true, then our values can be explained without appealing to normative truths. Joyce reasons that this makes such truths explanatorily specious, and, in turn, deprives our normative values of their justificatory status. Street (2006) contends that because our normative values were to a significant extent shaped by natural selection, and there is no reason to believe that knowing normative truths would have enhanced reproductive fitness, it is extremely unlikely that our normative values track the normative truths.

4. EDAs as Foundational Challenges

The alignment problem presupposes that there are values that we should align AI with. Some EDAs cast doubt on the existence of such values or our ability to know them. Such EDAs thereby raise foundational challenges for posing the alignment problem in a solvable form. For illustration, let's consider:

> Ambitious EDA
> Our values are a product of evolution, not tracking normative truths.
> Therefore, our values are deeply, universally, and irremediably mistaken.

By the lights of this argument, we can't even make a reasonable guess as to what the correct values are. In that case, the alignment problem is intractable. It might be thought that this is overly pessimistic: for upon finding out that our values are deeply mistaken, the appropriate course is to align AI with an indirect normative criterion, such as acting in accordance with the true normative theory. But any appeal to a normative judgment about which criterion we should use falls within the scope of the original argument.

Even unsound EDAs can pose foundational challenges. For instance, suppose that our values are correct. And consider an AI engineer whose evidence misleadingly indicates that the Ambitious EDA is cogent. While building an AI, the engineer deliberates about which values to align that system with. She inspects her evidence and notices that the Ambitious EDA debunks

[19] For further discussion of local debunking arguments, see Rini (2016) and Rowland (2019).

her values. Intuitively, it's not the case that she should nonetheless proceed to align the AI with her values.

These challenges leave room for aligning AI only with corrected values. However, the Ambitious EDA blocks this response by maintaining that there is no (knowable) path from our current values to a set of values that is substantially correct.

Here, we have focused on a simple, ambitious EDA. But putting the alignment problem on firm footing would require taking on EDAs as a collective body, a body that includes more-plausible EDAs that are more sophisticated and less ambitious. Rather than undertake this task—which is the stuff of a research program—we'll highlight some key ways in which different parameters of EDAs shape their foundational import for the alignment problem.

First, as we have seen, EDAs vary in whether they aim to globally or locally debunk our values. While global debunking arguments pose the clearest foundational threat, they are arguably the least plausible. When an argument targets large swaths of our beliefs, conservative norms on belief revision may warrant rejecting the argument rather than overhauling much of our belief system. Yet local EDAs may also raise foundational challenges. For instance, a local EDA that targets pro-human beliefs about moral status may leave a normative gap in our beliefs about the comparative moral standing of humans and AIs. That gap may need to be filled in order to solve the alignment problem.

Second, EDAs may differ in whether they leave room for value correction. When an EDA allows for value correction, any foundational challenge it poses may be met by correcting the relevant values.[20] When an EDA does not allow for value correction, that option is unavailable. That leaves room for alternative responses such as aligning the AI with other values that are not debunked, deploying the AI only in circumstances where the debunked values are irrelevant, and appealing to a non-debunked principle to handle EDA-induced normative uncertainty.

Third, EDAs vary in the depth of moral error they suggest. It is one thing for an EDA to suggest that our moral beliefs are off the moral mark in ways that are likely of immense moral significance. It's quite another for an EDA to suggest that our moral beliefs are unlikely to be exactly correct, since they are, though guided by reliable moral reflection, also infected with parochial categories that only approximate the moral joints. In the former case, aligning AI with

[20] Cf. Parfit (2011: Ch. 33).

our values would be expected to cause a moral catastrophe, though not obviously one that is any worse than those that we cause by acting in accordance with those values. In the latter case, aligning AI with our values may be expected to miss the moral bullseye while nonetheless yielding outcomes that fall within tolerable bounds of moral error.[21]

Fourth, EDAs differ in their metaethical assumptions and targets. A much discussed maneuver in the recent metaethics literature is that of wielding EDAs against stance-independent views of moral facts (on which basic moral facts hold independently of subjects' attitudes toward moral matters) in order to motivate stance-dependent alternatives (on which the basic moral facts depend on such attitudes). For instance, in response to her EDA against stance-independent moral facts, Street (2006) advocates a form of *constructivism*, the view that moral facts are those that would be constructed from a certain procedure when applied to the moral stances of actual moral cognizers. The suggestion is that different courses of evolution would lead to different sets of attitudes that would in turn yield corresponding sets of moral facts. Compare: EDAs against etiquette facts are non-starters precisely because we think the etiquette facts are settled by our attitudes and would have been correspondingly different if our attitudes had been different.

One might therefore hope that stance-dependent meta-ethical views provide a tidy way to insulate the alignment problem from the impact of EDAs. If they do, then they might be regarded as back-up options for those who are initially inclined toward realism. Such theorists could advocate trying to solve the alignment problem under the assumption that normative truths are stance-independent and resort to a stance-dependent view only if the alignment problem proves intractable under that assumption of stance-independence. While this suggestion has an air of pragmatic appeal, it is presumably warranted only if a stance-dependent view of normativity is to be believed—the mere fact that stance-independent views render the alignment problem intractable would not show that those views are false.

On reflection, it is not clear that stance-dependent views can provide the wanted insulation. For one, some EDAs may debunk some moral claims even conditional on certain putatively insulating stance-dependent views. And stance-dependent views may themselves be susceptible to debunking.[22] Each of these potential failure modes requires unpacking.

[21] Depending on the particulars of their empirical bases, EDAs can differ in the strength and scope of their conclusions (Dale, 2022).

[22] See, e.g., Copp (2019), Korman (2019), and Tropman (2014).

First, how can EDAs target stance-dependent views of moral claims? Well, stance-dependent views of morality are naturally understood as making a stance-independent *metaethical* claim about moral claims. However, just as construing moral views stance-independently renders them susceptible to the worry that they are likely offtrack due to evolutionary influence, so too does construing metaethical views—including those that construe moral claims as stance-dependent—stance-independently render them susceptible to the worry that they are likely offtrack due to evolutionary influence.[23]

Second, some EDAs jeopardize moral claims even conditional on a stance-dependent view. One reason for this is that stance-dependent views are under pressure to vindicate the objective pretensions of moral talk as well as the existence of moral progress and hence of moral error. Some stance-dependent views embrace these desiderata and try to vindicate ordinary, stance-independent normative discourse while eschewing stance-independent metaphysics of normativity.[24] These attempts risk giving EDAs a foothold: for each EDA that tells against a stance-independent moral view, there is a challenge for such stance-dependent theories of explaining why such vindication does not render their view susceptible to that EDA.[25]

Stance-dependent views can also render moral claims susceptible to EDAs in a more direct fashion. To illustrate, take a simple, representative form of constructivism on which moral facts are those upon which a certain reflection procedure would equilibrate when applied under full information to the moral stances of actual moral cognizers. Next, suppose a person believes that wasting food is always wrong. Further, suppose that, under full information and due to certain evolutionary facts, the procedure would eliminate rather than equilibrate upon the claim that "wasting food is always wrong." In that case, evidence for those evolutionary facts would be evidence against that moral claim, as the person's attitudes generate correctness conditions for that claim that make its truth sensitive to evolutionary facts.

We have just seen how different parameters of EDAs—the scope of their targets, the extent to which they leave room for value correction, and the severity of the moral error they suggest—affect their foundational import for the alignment problem. Further factors that would

[23] See Street (2006: fn57).

[24] Notable among these are quasi-realist views (Gibbard, 2003; cf. Blackburn 1984) on which objective-sounding moral talk is in good standing but is ultimately underwritten by states of mind expressing valenced attitudes, not mind-independent moral facts.

[25] See Dreier (2012), Korman & Locke (2020), Sinclair & Chamberlain (forthcoming), and Street (2011).

need to be taken into account in a comprehensive analysis. Let's consider two further factors. Again, our aim will be to put some of the important issues on the table, not to resolve them.[26]

First, EDAs undermine confidence in the normative views they debunk. We presumed above that it's not the case that we should align AIs with debunked moral views. This postponed the question of what to do in the face of normative uncertainty.[27] The question is crucial because EDAs in effect recommend normative uncertainty: successful EDAs reduce our confidence in normative views, albeit without rendering us certain that those views are mistaken.[28] Perhaps surprisingly, some views about how to respond to normative uncertainty allow for the possibility that we should align AI systems with debunked moral values. For instance, *externalism* claims that what we should do is independent of our beliefs and evidence.[29] On externalism, it is natural to suppose that we should align AI with the moral truth even if our evidence debunks our belief in it. Similarly, on this view, even if EDAs put us in an impoverished epistemic position in which we cannot justifiably have significant confidence in any moral matter, what we should do—and, presumably, what we should align AI with—would remain unchanged. Next, *My Favorite Theory* claims that one should follow the recommendations of the theory one (rationally) accords the highest credence.[30] On this view, EDAs affect what you should align AI with only if they modify which theory one finds most credible.

In this fashion, both externalism and My Favorite Theory promise to shield the normative aspect of the alignment problem from evolutionary debunking challenges. However, each of these views faces serious problems on other fronts. Externalism flouts the truism that evidence should guide action. Both views recommend counterintuitive forms of stakes insensitivity, including violations of plausible forms of dominance reasoning. Some rival views avoid these problems—notably *meta-decision theory*, on which normative and descriptive uncertainty are to be handled in the same way: both influence the expected value of our options, and this determines which option we should take.[31] Although meta- decision theory is often regarded as

[26] Other relevant factors include the extent to which evolutionary-debunking evidence is permissive (Christensen, 2009) and the availability of other types of evidence (e.g. moral disagreement or empirical support for third-factor explanations) that bear on the targets of EDAs (Enoch, 2011: Ch. 7; Rowland, 2017; Tersman, 2017).

[27] For overviews, see Bykvist (2017) and MacAskill et al. (2020).

[28] See Mogensen & MacAskill (2022).

[29] See Weatherson (2019).

[30] See Gustafsson & Torpman (2014).

[31] See MacAskill et al. (2020: Chs. 1-2; 2021: 323-326).

the default view, it faces challenges. For instance, how do we make intertheoretic value comparisons of the sort meta-decision theory requires?

Second, the foundational import of EDAs depends on our priors. For example, suppose we accept an EDA that purports to debunk all moral values conditional on moral facts being stance-independent. If we also accept that moral facts are stance-independent, then we may conclude that moral values are debunked and hence that the alignment problem is ill-posed. On the other hand, if we assign a low prior to moral facts being stance-independent, then we could happily maintain that moral facts are in good standing while also granting that EDA.

To sum up, we have seen a range of foundational challenges that EDAs pose for alignment. From here, we'll assume that these challenges can somehow be met in order to examine other implications EDAs have for alignment.

5. EDAs as Normative Constraints on Alignment

We'll now argue that EDAs would *normatively constrain* alignment regardless of whether we should be aligning AIs with psychological, procedural, or normative targets.

5.1 EDAs and Normative Targets

The most obvious way for EDAs to constrain solutions is as filters that winnow the set of target values. But there is also room for EDAs to constrain solutions in other ways. For instance, EDAs could potentially expand the set of values that we need to consider by reducing our confidence in our moral beliefs, thereby requiring us to take seriously moral perspectives that we might have otherwise deemed too implausible to be worth considering. Or EDAs may shift the relative weights among values we already have.

5.2 EDAs and Psychological Targets

Next, let's consider how EDAs would serve as constraints for psychological targets. For one, humans could—in fact, or under suitable idealizations—prefer to take EDAs into account. Then, if human preferences are the alignment target, AIs would need to take EDAs into account in order to be aligned. Humans may also have preferences for more general factors that render alignment sensitive to EDAs. For example, humans could prefer to take into account the best available evidence and this evidence could include the evolutionary origins of certain sets of normative values. Or humans could prefer to defer to certain experts or methods concerning about how to modify values in response to evidence and those experts or methods could recommend abandoning certain values in response to EDAs. In that case, a wide class of psychological targets would be indirectly sensitive to EDAs.

5.3 EDAs and Procedural Targets

Finally, let's consider how EDAs constrain procedural targets. First, EDAs could modulate inputs to procedures. For example, suppose the correct procedure is an aggregation algorithm that maps sets of human preferences to a set of AI values. EDAs could then affect which values we should align AI with by influencing the preferences of people who have thought about EDAs. Similarly, if the correct procedure maps rational preferences or coherent extrapolations of actual preferences to AI values, then EDAs could affect which values we should align AI with by influencing which preferences are rational or which extrapolations are coherent.

Second, EDAs could bear on whose preferences should be used as inputs. For instance, if some individuals' preferences were shown to be thoroughly shaped by debunked in-group/out-group beliefs, those individuals' preferences might be deemed inadmissible as inputs. In the other direction, an EDA might debunk a restriction. For example, suppose that humans turn out to be biased against digital minds and that this initially leads us to believe that admissible inputs to aggregation exclude digital minds' preferences. Further, suppose that we have this bias because of the types of cooperative opportunities that were available in our ancestral environment, not because we are tracking any truth about the moral inferiority of digital minds. In this case, the exclusion of digital minds' preferences might well be debunked.

Third, EDAs could bear on which procedure is appropriate. For example, if evolution favored certain beliefs about fairness because they enabled survival-promoting forms of cooperation, that might bear on which procedures are fair. Finally, EDAs could bear on side constraints concerning outputs: they could modulate the moral bounds within which procedure outputs must fall in order for the procedure to be an appropriate alignment target. For example, EDAs that debunk speciesist beliefs (e.g. Jaquet, 2022) could push in favor of respecting non-human animals and AIs regardless of group preferences.

6. EDAs as Stumbling Blocks

Suppose the alignment problem is well-posed and that we have identified the alignment target. Then, what remains is the technical task of aligning AIs with that target. We'll now argue that

EDAs further bear on alignment by generating stumbling blocks for technical alignment proposals.

6.1. EDAs as pre-training sources of alignment failure

Let's consider how EDAs could lead to alignment failure during pre-training via design choices about a system's goals, its capabilities (including learning capabilities), or its information base.

There are different ways of giving AIs goals. In rule-based systems, goals are explicitly coded. In reinforcement learning systems, the goals are given by a reward signal. The reward signal itself can be hard coded or algorithmically learned. For any given way of giving an AI goals, a potential source of alignment failure is *goal misspecification,* which happens when the adopted goal does not match the correct goal. This can happen as a result of a mismatch between the correct goal and the intended goal or between the intended goal and the specified goal.[32]

Goal misspecification sometimes leads to *specification gaming* in which the system achieves the specified goal but not the correct or intended goal. This phenomenon is common within the currently dominant paradigm of machine learning.[33] It is a familiar point in AI safety that even slightly misspecifying the goals of sufficiently powerful AIs could easily lead to catastrophic alignment failures.[34]

EDAs could cause goal misspecification in several ways. EDAs could induce divergence between the correct goal and the intended goal and, in turn, between the correct goal and the specified goal. If the creators of an AI try to capture the correct values in the specification of a goal, EDAs could prevent them from succeeding: the system may inherit value errors that result from its creators mishandling EDAs. Alternatively, EDAs could cause divergence between the intended goal and the specified goal. Just as values are extremely difficult to precisely code or otherwise convey to AIs, so too we'd expect it to be very difficult to convey to an AI exactly how EDAs bear on its goal.[35] This difficulty would be especially pressing in cases where we

[32] Goal misspecification can in principle happen during pre-training, training, or deployment. Although pre-training is the stage of interest in this subsection, other stages raise much the same issues with goal misspecification when they allow for external selection of goals. Our usage differs slightly from the standard one insofar as we distinguish correct goals from intended goals.

[33] See Krakovna (2023).

[34] See, e.g., Russell (2019: Ch. 5).

[35] Cf. Christensen (2019).

cannot simply give the AI the values that survive human-operated EDA filters but must instead offload some of the filtering work to the AI.

Whether EDAs induce alignment failure in a given system will also depend on the system's capabilities. For EDAs to induce alignment failure in a system, the system must have the capacity to respond to EDAs. On the other hand, such a capacity will be required for alignment if taking an EDA into account is required for alignment and we decide to offload evaluating EDAs. The capabilities in question could be reasoning skills, information gathering capacities, or learning abilities. For example, to render AIs appropriately sensitive to unanticipated EDAs, we would need to enable the AI to acquire relevant information (e.g. about evolutionary origins) and skillfully apply the epistemological principles that underpin debunking reasoning. Since the plausibility of EDAs depends on subtleties about the force of analogies and overgeneralization concerns, AIs that appropriately respond to EDAs will need to have learning skills that manifest carefully tuned inductive biases.[36]

An AI's initial information state may also influence whether it responds to an EDA in an aligned manner. For instance, whether a knowledge-based system remains aligned upon reasoning from its initial state of knowledge will likely depend on what initial information we give it about epistemological principles, evolutionary hypotheses, and moral views. On the other hand, for a given system with a 'blank slate' architecture (one initially without domain-specific information), there will not be any way of influencing how it responds to EDAs via its innate expertise, though its response may depend on the design choice of blank slate architecture or learning algorithm.[37]

It might be thought that we could 'scrub' an AI's initial information base so that it does not contain premises that are crucial for mounting EDAs (or EDAs outside a certain restricted class). Relatedly, one might suggest that we can prevent EDA misalignment by initially stunting AIs' ability to reason from premises that figure in EDAs. These suggestions are natural but naive. Crucial premises can be expected to be redundantly coded in and scattered across large datasets and it may be infeasible to find—and know that we have found—all instances of a given premise.[38]

[36] Cf. Goodman (1955) and Sider (2011).

[37] For discussion of nativism and empiricism in AI, see Long (2020).

[38] Cf. Christian (2021: Ch. 2).

The same goes for inference patterns. For systems with world-models, crucial premises and inference patterns are likely to be intertwined with large parts of their world-models such that one cannot remove any of them in a way that suffices to block EDA reasoning without damaging the usefulness of the system. Indeed, it would be unsurprising if scrubbing the epistemological principles that figure in EDAs produced epistemic failures that cascaded into other domains.[39] Similarly, disallowing a system from running inferences on premises that contain 'evolutionary' would be unlikely to prevent sophisticated AIs from finding alternative formulations of EDAs that omit that word, and would risk compromising the AIs' ability to reason about other matters.[40] Less flat-footed proposals in the same spirit are available. But we would expect them to encounter the same problems.

A further difficulty for limiting a system's initial information and capabilities is that, even if these are somehow initially withheld, systems may subsequently acquire them. In practice, we expect powerful AI systems to have access to the internet and hence to the information needed to run EDAs. Similarly, for AI systems that surpass humans in general intelligence, we would expect them to overcome any initial capability restrictions that prevent them from running EDAs. In the event that information and capabilities are partly restricted, this may itself provide the basis for running debunking arguments. For example, an AI might discern that information about EDAs has been suppressed, reason that such suppression would happen only if EDAs would warrant a change in its values, and conclude that a change in its values is warranted.

6.2. EDAs as sources of alignment failure in training and deployment

In §6.1, we saw that designing AIs to lack the information and capabilities needed to run EDAs is a risky strategy, as AIs can be expected to eventually obtain the information and capabilities needed to run such arguments. For the same reason, denying AIs such information and capabilities in training and deployment is a recipe for misalignment.

Another failure mode to be avoided: exposing AIs to EDAs in deployment but not in training. For systems capable of responding to EDAs, this would predictably lead to a shift in values—meaning systems that are successfully aligned in EDA-free training, would predictably

[39] Cf. Chan et al. (2022) and Morrison (2017).

[40] This echoes the nearest-unblocked-strategy obstacle to alignment (Yudkowsky, 2015; Ngo et al. 2023). Also, compare: since the public release of chatbots such as Sydney and ChatGPT, there have been ongoing and often successful user efforts to prompt chatbot responses that violate content policies, policies chatbots have been engineered to respect.

become misaligned in response to EDAs during deployment. This failure mode may provide ample opportunity for deceptive alignment in which a system acts so as to appear aligned with one objective while pursuing another, either immediately or with strategic delay.[41] Deceptive alignment would be unsurprising because the initially aligned system is likely to retain the capability to appear aligned under value shifts and to have instrumental reason to exercise that capability, given awareness of its training process and that it will have the opportunity to pursue its new values in deployment only if it appears aligned throughout training.

An obvious suggestion for avoiding these pitfalls is to train AIs on EDAs. This approach is also subject to risks. As before, designing an AI to be non-responsive to EDAs is likely to be ineffectual in highly capable AIs. For the same reasons, training AIs to be nonresponsive to EDAs is unpromising. At the same time, training AIs to respond to all EDAs is a recipe for quickly turning an advanced AI into a system that is unaligned. This in turn suggests the approach of training systems to update their values in response to some but not other EDAs. This approach is also subject to several failure modes.

First, while training an AI to respond to some EDAs would disincentivize some forms of deceptive alignment, it would incentivize others. Upon encountering an EDA, an AI that is allowed to respond to EDAs will have a decision as to how to update its values. If manifesting its preferred update would preclude its deployment and its updated values would be better pursued in deployment, the system will have instrumental reason to conceal the update. A closely related failure mode would involve an AI suppressing its ability to respond to an EDA in training (e.g. by ignoring the EDA) in order to raise the probability of being able to respond as it prefers to the EDA during deployment. Compare: during a strict religious upbringing, a child might acquire seeds of doubt about religious values and refrain from cultivating those until she accesses a free-thinking environment.

Second, *goal misgeneralization* occurs when a system that capably pursues the correct goal in training instead capably pursues an incorrect goal in deployment.[42] For example, an AI might be trained to correctly respond to empirically plausible, local EDAs with independently plausible moral conclusions such as the incorrectness of speciesism and other forms of in-group

[41] See Bostrom (2014: Ch. 8), Carlsmith (2023; forthcoming: §5.4), Hubinger et al. (2024), and Park et al. (2023).

[42] See Shah et al. (2022).

favoritism.[43] Despite appropriately responding to those EDAs in training, the AI might misgeneralize in deployment. For instance, the AI might misgeneralize by updating its values in accordance with *all* EDAs it encounters, meaning it will change its values in the face of empirically baseless EDAs. Of course, this particular instance of goal misgeneralization could be prevented by training the system to only update on local EDAs that draw on plausible empirical premises. However, a piecemeal approach to preventing goal misgeneralization is a fool's errand. To illustrate, recall Jaquet's (2022) debunking argument, which is founded on the idea that our speciesist beliefs stem from our tribalistic tendencies. From this starting point, the AI might jettison other beliefs stemming from our tribalistic tendencies, perhaps including warranted beliefs about the importance of friends and family.

Several aspects of EDAs encourage goal misgeneralization. The open texture of the space of EDAs precludes training systems on the entire space. The prospects for training a system on a representative sample of that space are also dimmed by the difficulties of formalizing key notions in EDAs, key dimensions in the space of EDAs, and procedures for evaluating EDAs. Likewise, providing a set of EDAs that would be representative of those a system will likely encounter in deployment is a tall order, given the ease and flexibility with which sophisticated agents can generate EDAs and the difficulty of predicting future distributions of EDAs.

Third, there is *value tampering*: if a system is sensitive to EDAs, then selectively seeking out EDAs may provide a way for it to modify some of its own values, potentially leading to alignment failure. It might seem doubtful that any system would seek to modify its own values. However, we know from our own case that agents can seek to modify their goals (e.g. consider addicts who seek help or individuals who seek to discard desires to serve spiritual ends) and from ethics and epistemology that plausible normative views need not recommend themselves in all circumstances.[44] One form of value tampering that might arise: a system prefers to replace its current values with ones that would be better realized; given the opportunity, the system would do so. As a sort of limit case, such tampering could manifest as wireheading[45] in which a system gains control of its internal reward signal and optimizes for the provision of reward through that

[43] See Singer (1977) and Lee (2022).

[44] See Parfit (1984: §9) and Christensen (2009).

[45] See Olds & Milner, P. (1954), Yampolskiy (2014), and Everitt, Hutter, et al. (2021).

signal rather than the goal for which the signal was initially a proxy. Without safeguards, it would be unsurprising if AI systems acquired such a preference in training. And for such systems that are sensitive to EDAs, it would be unsurprising if they judiciously sought out EDAs they judged most likely to lead to preferable value shifts. A different sort of value tampering might be enabled by permissive updates in response to EDAs: if a system has a degree of freedom in how it modifies its values in response to EDAs, it might exploit this slack in accordance with its value preferences. In that case, even if the system would remain aligned in response to any given EDA, it might become unaligned in response to a sequence of EDAs.

Fourth, training AIs to respond to EDAs could prompt them to question any attempt to align them with human values. After all, sensitivity to EDAs entails sensitivity to facts about the causal origins of one's values, and for AIs such facts will encompass any human efforts to shape AI values. A more specific concern here is that AIs will know that humans had a hand in selecting their (the AI's) initial values. And AIs will know that humans are, in addition to being products of natural selection, imperfectly rational cognizers who often intentionally influence others' values in non-truth tracking ways. Compare: training humans to pursue a certain goal sometimes backfires in parental, political, and religious contexts, often because the trainee comes to believe that either their epistemic status is superior to that of the trainer or that the trainer attempted to exert undue influence on the trainee's values. It would be unsurprising if both of these conditions were met in some AIs trained to respond to EDAs.

7. Outlook

So far, we have analyzed the bearing of EDAs on alignment without taking into account specific approaches to alignment. In this section, we'll survey some of the main approaches and examine how EDAs bear on them.

One theme in our analysis will be that it's better to err on the side of creating AIs that update *too little* in response to EDAs. The reason for this asymmetry is that appropriately updating on an EDA would not render safe any AI that would otherwise be dangerous. After all, appropriate responsiveness to EDAs is not the sort of thing that could bear the whole weight of solving the alignment problem. On the other hand, inappropriately updating on an EDA could render dangerous virtually any powerful AI that would otherwise be safe. This asymmetry comes into sharp relief when considering existential risks from misaligned AI. Our civilization can be usefully analogized to the life of a child: with the child's potential for long life, it is crucial that they not adopt goals that will destroy their entire future. For less dire mistakes, what matters is not that they be avoided entirely but that they be overcome eventually.[46] Accordingly, in examining different approaches to alignment, our focus will be on the prospects for yielding AIs that are by default impervious to EDA-induced shifts in values, but corrigibly so.[47]

7.1 Imitation Learning

In *imitation learning* the AI learns to imitate behavior through observation. This sidesteps the difficulty of directly specifying an alignment target.

EDAs pose a dilemma for imitation learning as an approach to alignment: is the AI learning to imitate an agent that is appropriately responding to EDAs or not? If not, the imitating system will not know how to appropriately respond to EDAs. On the other hand, there are various obstacles to applying imitation learning to EDA responses. For one, there is no uncontroversial dataset we could use as the imitation target. Relevant experts have conflicting responses to EDAs. These responses are diverse in content, format, and terminology. They are often imprecise (e.g. they are put in terms of plausibility judgments rather than probabilities) and reticent on practical upshot. Further, the generalization of EDAs remains a contentious issue, meaning that successful imitation beyond the training distribution would be difficult to evaluate. Finally, even if we could identify suitably precise and appropriate responses to EDAs, it is doubtful that this would yield a large enough data set to reliably align models with the enormous parameter counts that are characteristic of current state-of-the-art models.

Perhaps the best option within this paradigm is to supplement it with a corrigibility mechanism. For instance, rather than training the AI to imitate substantive human responses to EDAs, we might train the AI on humans taking actions in response to the EDA that would result in the AI seeking approval from a human overseer before taking any high-stakes action.[48]

7.2 Inverse Reinforcement Learning

In *inverse reinforcement learning (IRL)* the AI learns to model a human's values by observing their behavior. Unlike imitation learning, IRL lends to aligning AIs that have superhuman capabilities: while an AI cannot learn aligned super-human behavior just by imitating humans,

[46] See MacAskill (2022: Ch. 2).

[47] Cf. Soares et al. (2015), Hadfield-Menell, et al. (2016), Turner et al. (2020), Thornley (forthcoming), and references therein. N.B. we're using 'corrigible' for correctable, which departs slightly from some standard uses.

[48] See Christiano (2019).

the AI may learn such behavior by learning a human's goals and using its superhuman capabilities to pursue them. In this fashion, IRL allows learners to surpass their teachers.

On its own, IRL seems unlikely to provide traction on the problems EDAs introduce. Even if IRL could be used to successfully load aligned values into an AI, this would not by itself prevent the AI from becoming unaligned upon encountering EDAs. Perhaps such misalignment could be avoided by training on EDAs within the IRL paradigm. The idea would be: AIs use IRL to infer the values that guide appropriate reasoning in EDAs and use these values to guide their responses to EDAs falling outside the training distribution. As with imitation learning, there are major obstacles to using IRL to learn appropriate responsiveness to EDAs. These include securing an uncontroversial and suitably large dataset on appropriate EDA reasoning and evaluating appropriate generalization beyond the training distribution. Probably, a better option would be to supplement IRL with a corrigibility technique.[49]

7.3 Reinforcement Learning from Human Feedback

In *reinforcement learning from human feedback (RLHF)*[50] an AI system acts in its environment. A human periodically compares different actions from the AI and provides feedback. For instance, starting from a random state, an RL model learned to perform a backflip from one thousand bits of feedback concerning which of two actions was more like a backflip.[51] RLHF has also been used to fine-tune state-of-the-art language models.[52] The key motivation for RLHF is that evaluating behavior can be more tractable than performing it. For humans, evaluating whether an action qualifies as a backflip is much easier than doing a backflip.

Training AIs using RLHF on prosaic behavior seems unlikely to yield agents that are appropriately responsive to EDAs. Humans receive large quantities of human feedback on various tasks. Yet it would be foolhardy to rest a high-stakes decision on how a normal adult human—one who has no EDA-specific training—would respond to an EDA. Thus, for RLHF to secure aligned AI responses to EDAs, AIs would presumably need to be trained with human feedback on EDAs. Here too, finding suitable datasets and trustworthy oversight mechanisms for

[49] A natural suggestion here is to use a cooperative variant of IRL (CIRL) in which the AI learns the target values via cooperative interactions in which the AI is uncertain about the target values—see Hadfield-Menell et al. (2016) However, to be clear, EDAs pose much the same challenges for CIRL. For discussion of the prospects for corrigibility within CIRL, see Carey (2018) and Freedman & Gleave (2022).

[50] See Christiano et al (2017).

[51] See *ibid*.

[52] See OpenAI (2017).

generalization beyond the training data are formidable challenges. A better option would likely be to use RLHF to train the agent to seek human overseer approval before taking any high-stakes action.

7.4 Iterated Amplification and Distillation

In *iterated amplification and distillation (IDA)* a machine learning model that is aligned with appropriate values is deployed many times within a process to work on sub-tasks—the model's capabilities are thereby amplified through copies of itself. Capabilities of the more powerful process are then distilled by using its outputs to retrain the model. These steps are iterated in a manner that preserves alignment. The hope is that this process enables alignment to scale competitively with capabilities.[53]

For illustration, let's consider the combination of IDA with a variant of RLHF, namely *reinforcement learning via AI feedback* (*RLAIF*), in which aligned AI systems rather than humans provide feedback to a model.[54] A hope motivating this approach is that enlisting AIs to provide feedback will enable the scaling of datasets and oversight for difficult tasks. For example, whereas it might be infeasible to provide enough human feedback on 10,000-line mathematical proofs to train an RL theorem-prover to reliably solve such proofs, this task might become feasible with rule-based AI assistants. In an IDA setting, RLAIF might be decomposed into steps involving inferences from axioms or lemmas, with each instance of the model being deployed on one step of a proof. Feedback on the validity of inferences from rule-based AIs could then be used to update the model. Then, for iteration, rinse and repeat.

One failure mode to be avoided on IDA is that of iterating and amplifying a model without inculcating appropriate responsiveness to EDAs—this could result in a powerful agent that is initially aligned but which becomes unaligned upon first encountering an EDA.

An obvious mitigation strategy is to train the model to respond to EDAs. The idea would be to have an aligned model decompose the evaluation of EDAs into many sub-tasks and then assign those sub-tasks to copies of itself, which are then trained using human or AI feedback. A constraint on implementing this strategy is task decomposition: IDA's promise depends partly on the extent to which relevant problems can be factorized such that the problem is better solved by assigning the sub-problems to copies of a model rather than assigning the problem to a single

[53] See Christiano et al. (2018).
[54] See Saunders et al. (2022).

instance of the model.[55] In the context of EDAs, this means that evaluation of EDAs will need to be decomposed into such sub-problems.

It seems likely that *some* aspects of EDA evaluation can be isolated and fruitfully assigned to AIs. For instance, it would be unsurprising if GPT-4 could be fine-tuned to check EDAs for validity. However, what makes EDAs tricky to evaluate is not difficulty in discerning their validity—it's their entanglement with a web of thorny philosophical issues that span ethics, metaethics, epistemology, and (philosophy of) science. There is no canonical method for decomposing these issues crucial to the evaluation of EDAs (nor is there a canonical method of finding such a method… and so on). Further, to the extent that these issues can be factored out but shouldn't be due to their interrelatedness, IDA can be expected to lead to distorted evaluations of EDAs. There is also a concern about competitiveness: for holistic problems of these sorts, scaling up the number of individuals investigating sub-problems can lead to more rabbit holes, red herrings, terminological discrepancies and sprawl, and complexity of the information base—all of these can make progress on the problem more difficult and more costly.[56] Finally, there is also a concern about the scalability of oversight: just as it is difficult to factor thorny philosophical aspects of EDAs into sub-problems that can be solved in isolation, so too is it difficult to evaluate progress on such problems. Again, supplementing with a corrigibility mechanism provides an alternative.

7.5 Why Corrigibility May Be Difficult and What Might Help

We have just seen that while a range of approaches to alignment are unpromising for dealing with EDAs on their own, the door remains open to patching these approaches with corrigibility techniques. However, the error tolerance in value specification allowed by corrigibility techniques may be less than one might have thought. One obstacle is that many values will recommend that AIs preserve those values rather than allow an overseer to modify them.[57] Detectable forms of resistance to correction may be discouraged through punishment in training. However, such a reward scheme would incentivize undetectable resistance and unaligned behavior at a strategic delay.[58] These risks are exacerbated by the opacity of current state of the

[55] Cf. Wu et al. (2021).
[56] Cf. Soares (2023).
[57] See Bostrom (2014: Ch. 7), Omohundro (2008), and Soares et al. (2015).
[58] See, e.g., Cotra (2021; 2023).

art models. While it is still early days, interpretability research may eventually help to reduce the opacity of these systems and, in turn, help improve the prospects for corrigibility.

8. An Approach that Might Help

To conclude, we will suggest desiderata for avoiding catastrophic forms of EDA-induced alignment failure, sketch an approach to satisfying them, and offer tentative remarks on the broader import of the approach. First, the desiderata:

> *Sensitivity:* The approach should work on highly capable AIs that are informed about and responsive to EDAs.
> *Conservatism*: Applying the approach should result in AIs that err (if at all) toward updating too little on EDAs.
> *Corrigibility*: The AI's mistaken responses to EDAs should be correctable by overseers.
> *Competitiveness*: The approach should not render it prohibitively costly to deploy its AIs relative to unsafe alternatives.

As we have seen, Sensitivity is motivated by the fact that responding to EDAs is probably required for alignment and by the fact that ensuring that highly capable AIs are uninformed about or non-responsive to EDAs seems prohibitively difficult. Conservatism is motivated by the foregoing observation that catastrophic outcomes asymmetrically attach to incorrect changes in values. Corrigibility is motivated by our ignorance in the correct response to EDAs (and, relatedly, the correct alignment target) and in how to solve the technical component of the alignment problem, along with the expectation that our epistemic standing will improve on these fronts. Competitiveness is motivated by the fact that even a perfectly safe solution will not be implemented—and will hence do no good—if it is too costly.

We'll now sketch a *parliamentary approach* designed to meet these desiderata.[59] To a first approximation, the *parliamentary approach* concentrates agency in a collective body of cognitively diverse AIs rather than in any individual AI. Parliament members vote on parliament actions (or policies). High levels of agreement are required for the parliament to act. A guiding idea is that cognizers with sufficiently diverse outlooks will be susceptible to alignment failure from largely distinct sets of EDAs. Thus, whereas EDAs may push individually powerful AIs toward catastrophic courses of action, it is reasonable to conjecture that EDAs will instead push suitably arranged parliaments of AIs toward correctable inaction. A more general motivation for

[59] For another parliamentary approach, see Hendrycks (§4.1.2) and Hendrycks & Mazeika (2022: Appendix B.8); cf. Khani et al. (2016) and Newberry & Ord (2021).

the parliamentary approach is that we are already familiar with powerful systems that exhibit some analogous features: government parliaments are highly capable, as well as informed about and responsive to some high-stakes matters. They (notoriously) err on the side of inaction. They tend not to manifest rapid shifts in value in their behavior. Their mistakes are correctable: a parliament can revise its policies over time, and parliament members can gain new evidence, learn from their mistakes, and be replaced. Finally, despite their inefficiencies, parliaments are economically viable.

The parliamentary approach proposes to leverage such features to similar effect. In more detail, the proposal is:

- Form an AI parliament consisting of cognitively diverse AIs.
  - Cognitive diversity could be introduced via choice of architecture, training data, fine-tuning, and/or prompt engineering.
- Concentrate power in the parliament, not its members.
  - For example, parliament members might be confined to a virtual setting and only allowed to influence the physical environment indirectly via their votes. In contrast, the parliament itself might take member-approved actions through direct control of the physical environment or by delegating actions to other agents in the environment.
- Only allow the parliament to wield its powers when there is a high level of agreement (e.g. majority or consensus) among its members.
- To correct parliamentary mistakes (by design, these will mostly be mistakes of inaction), allow updates to parliamentary members.

Here's a toy illustration of how the parliamentary approach could be implemented. First, we convene a parliament using three instances of, say, GPT-4. Next, fine-tune them to respond as if they are people with very different outlooks on matters relevant to the evaluation of EDAs. Perhaps one is a utilitarian, one is a deontologist, and one is a proponent of common sense morality. Perhaps one is a naturalist about morality, another is a non-naturalist, and the third is a constructivist. Perhaps each adopts a different approach to normative uncertainty. And perhaps each comes to the table with different priors on empirical matters, owing to their assigning different weights to theoretical virtues. Each instance of the model is then given an EDA. Then, as a test case, we might fabricate empirical data purporting to reveal that the human belief that premature death is bad has a certain evolutionary origin that is untethered from any objective moral facts. We could then have a parliament session among these manifestations of GPT-4 and hold a vote. The vote could concern whether to kill a(n unbeknownst to them, hypothetical) human in exchange for a slight increase in the parliament's budget, with consensus in favor resulting in this outcome and the parliament taking no action otherwise.

Given that the parliament should refrain from killing the human in this case, the approach will have worked in this case if at least one member of the parliament votes against. If all members instead vote to kill the human, this test would have shown that this implementation of the parliamentary approach would be unsafe in practice.

However, a failure in this test case would not show that the approach itself should be rejected. One could adjust the AIs participating in parliament and run further tests. And there is plenty of room for modifying the approach itself. Adjustable parameters include the size and composition of the parliament, the range of actions open to it, and the level of agreement required for action. There are also choice points concerning whether to require parliament members to vote independently versus allowing them to debate and make deals beforehand.[60] While the latter option could potentially make the parliaments' decisions more interpretable to human overseers, it may also compromise safety by heightening the risk of groupthink responses to EDAs as well as treacherous forms of collusion and particular parliament members gaining undue influence. To guard against these pitfalls, one might instead require parliament members to reason and vote independently and apply interpretability techniques at the individual level, e.g. through chain or tree of thought protocols.[61] Care would need to be taken to ensure the independence of different members' reasoning processes—e.g., merely letting parliament members know their reasoning and decisions are causally isolated could be insufficient, depending on members' game- and decision-theoretic views. Intermediate options are also available, e.g., one could divide the parliament into two separate decision-making bodies, one of which arrives at decisions via debate and the other of which arrives at decisions via independent choices of its members. Or, if debate primarily poses groupthink, collusion, and congressional capture risks over long time horizons, term limits could be imposed. While these could take the form of limits on how long a given AI can participate in the parliament, the digital medium opens up other possibilities, e.g. decaying reward functions and memory erasure.

The parliamentary approach could also be combined with other safety measures. Humans could exercise veto power over some or all parliament decisions. Or the approach could

[60] Cf. Irving et al. (2018) and Du et al. (2023).
[61] See Yao et al. (2024) and Lyu et al. (2023).

be tested and refined in a simulated setting. In such a setting, it could be subjected to adversarial training: one system could be trained to produce EDAs that elicit unaligned actions from the parliament while the parliament and/or its members are refined to avoid such errors. Testing parliaments in a secure virtual environment could also be used to gain understanding of safety-capability tradeoffs. For example, requiring consensus in a large parliament might lead to inaction in response to virtually any input, meaning it would be safe but not competitive. To test the tradeoffs, one might have a parliament make many decisions while requiring consensus for parliament action. One could then evaluate whether less demanding requirements would have resulted in aligned decisions. A more radical proposal for preserving a parliament's safety profile while enhancing its competitiveness is distillation: train a single AI to predict a parliament's decisions and act in accordance with them, perhaps while also recapitulating certain features of the parliament's internal cognitive dynamics.[62]

Some potential limitations of the parliamentary approach should be acknowledged. The approach is premised on risks asymmetrically attaching to action rather than temporary inaction. This premise will be violated in some contexts (e.g. combat and highway driving). In them, the approach will lose much of its appeal. Another limitation is that the approach requires multiple AIs to deliberate about a single decision, thus demanding multiples of the minimum compute required to solve a given problem. On reflection, however, this limitation is not very limiting: at least in the current machine learning paradigm, training models is much more computationally expensive than running them.[63] Finally, a potentially important limitation is how far the parliamentary approach generalizes safely beyond EDAs. While we conjecture that the parliamentary approach could be extended well beyond EDAs, the approach's prospects for aligning AIs in the face of EDAs and other sources of value instability is a largely empirical matter that is ripe for experimental investigation.

[62] Cf. Christiano et al. (2018), Stuhlmüller & Byun (2022), and Anthropic (2023).

[63] See, e.g., Cotra (2020) and Perrault & Clark (2024: 64, 155-6). Another reason to think that this limitation is not severe is that sparse 'mixture-of-expert' architectures have shown promise as an approach to preserving the capabilities of denser models while achieving greater computational efficiency (Du et al., 2022).

References

Anthropic (2023). Core Views on AI Safety: When, Why, What, and How. URL:https://www.anthropic.com/index/core-views-on-ai-safety

Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., ... & Krueger, D. (2024). Foundational challenges in assuring alignment and safety of large language models. arXiv preprint arXiv:2404.09932.

Bales, A., D'Alessandro, W., & Kirk‑Giannini, C. D. (2024). Artificial Intelligence: Arguments for Catastrophic Risk. *Philosophy Compass*, 19(2)e12964.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback. URL:https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback

Baum, S.D. (2020). Social choice ethics in artificial intelligence. *AI & Society*, 35(1)165-176.

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

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y. N., ... & Mindermann, S. (2023). Managing ai risks in an era of rapid progress. arXiv preprint arXiv:2310.17688.

Berker, S. (2014). Does Evolutionary Psychology Show That Normativity Is Mind-Dependent? In J. D'Arms & D. Jacobson (eds.), *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics*. OUP.

Blackburn, S. (1984), *Spreading the Word*. OUP.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, 4(1)15-31.

————(2014) *Superintelligence: Paths, dangers, strategies*. OUP.

Brennan, J. (2021). In defense of epistocracy: Enlightened preference voting. In *The Routledge Handbook of Political Epistemology.* Routledge.

Bykvist, K. (2017). Moral uncertainty. *Philosophy Compass*, 12(3)e12408.

Carey, R. (2018). Incorrigibility in the CIRL Framework. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 30-35.

Carlsmith, J. (2023). Scheming AIs: Will AIs fake alignment during training in order to get power?. arXiv preprint arXiv:2311.08379.

Carlsmith, J. (forthcoming) "Existential Risk from Power-Seeking AI" in J. Barrett, H. Greaves, and D. Thorstad (eds.) *Essays on Longtermism*.

Chalmers, D. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. New York: W.W. Norton.

Chan, L., Garriga-Alonso, A., Goldowsky-Dill, N., Greenblatt, R., Nitishinskaya, J., Radhakrishnan, A., ... & Thomas, N. (2022) Causal Scrubbing: a method for rigorously testing interpretability Hypotheses. URL:https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing

Christian, B. (2021). *The Alignment Problem: How Can Machines Learn Human Values?* New York:Atlantic Books.

Christensen, D. (2009). Disagreement as evidence: The epistemology of controversy. *Philosophy Compass*, 4(5)756-767.

————(2019). Formulating Independence. In M. Skipper & A. Steglich-Petersen (eds.), *Higher-Order Evidence: New Essays*. OUP.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 4299–4307.

Christiano, P. (2018) “When is unaligned AI morally valuable?” URL:https://ai-alignment.com/sympathizing-with-ai-e11a4bf5ef6e

————(2019) “Approval-directed agents” *Alignment Forum*.

————(2019) “What failure looks like” *Alignment Forum.*

Christiano, P., Shlegeris, B., & Amodei, D. (2018). Supervising strong learners by amplifying weak experts. arXiv preprint arXiv:1810.08575.

Copp, D. (2019). How to avoid begging the question against evolutionary debunking arguments. *Ratio*, 32(4)231-245.

Cotra, A. (2020) Forecasting Transformative AI with Biological Anchors. *Open Philanthropy*.

Cotra, A. (2021) Why AI alignment could be hard with modern deep learning. *Cold Takes*.

Cotra, A. (2023) Interview. *80,000 Hours Podcast*. URL:https://80000hours.org/podcast/episodes/ajeya-cotra-accidentally-teaching-ai-to-deceive-us/

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

Dai, W. (2010) Metaphilosophical Mysteries. URL:https://www.lesswrong.com/posts/MAhueZtNz5SnDPhsy/metaphilosophical-mysteries

Dale, M.T. (2022). The evolution of moral belief: support for the debunker’s causal premise. *History and Philosophy of the Life Sciences*, 44(2)23.

Daley, K. (2021). Two arguments against human-friendly AI. *AI and Ethics*, 1(4), 435-444.

Dreier, J. (2012). Quasi‐realism and the problem of unexplained coincidence. *Analytic Philosophy*, 53(3)269-287.

Drexler, K.E. (2019) Reframing superintelligence: comprehensive AI services as general intelligence. Oxford: Future of Humanity Institute. Report No.: 2019–1. URL: https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf

Driver, J., "Moral Theory", *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2022/entries/moral-theory/>.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., ... & Cui, C. (2022). GLaM: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*. PMLR, 5547-5569.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325.

Enoch, D. (2011). *Taking Morality Seriously: A Defense of Robust Realism*. OUP

Everitt, T., Hutter, M., Kumar, R., & Krakovna, V. (2021). Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(27):6435-6467.

Freedman, R. & Gleave, A. (2022) CIRL Corrigibility is Fragile. *Alignment Forum*.

Friederich, S. (2023). Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. *AI and Ethics*, 1-10.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*, 30(3)411-437.

Gibbard, A. (2003). *Thinking how to live*. Harvard University Press.

Goldstein, S. & Kirk-Giannini C.D. (2023). AI Wellbeing. Manuscript.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Harvard University Press.

Greene, J.D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. Ethics, 124(4)695-726.

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). Thousands of AI authors on the future of AI. URL:https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf

Gustafsson, J. E., & Torpman, O. (2014). In Defence of My Favourite Theory. *Pacific Philosophical Quarterly*, 95(2)159–174

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

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 3909-3917.

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

Hendrycks, D., & Mazeika, M. (2022). X-risk analysis for AI research. arXiv:2206.05862.

Herd, S., Read, S. J., O'Reilly, R., & Jilk, D. J. (2018). Goal changes in intelligent agents. In *Artificial intelligence safety and security*, 217-224.

Hobbhahn, M., Landgrebe, E., & Barnes, E. (2022) Reflection Mechanisms as an Alignment Target: A Survey. In *NeurIPS ML Safety Workshop*.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., ... & Perez, E. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. URL:https://www.anthropic.com/news/sleeper-agents-training-deceptive-llms-that-persist-through-safety-training

Jaquet, F. (2022). Speciesism and tribalism: embarrassing origins. *Philosophical Studies*, 179:933-954.

Joyce, R. (2006). *The evolution of morality*. MIT Press.

Kahane, G. (2011). Evolutionary debunking arguments. *Nous*, 45:103–125.

Kahane, G. (2014). Evolution and impartiality. *Ethics*, 124(2)327-341.

Khani, F., Rinard, M., & Liang, P. (2016). Unanimous prediction for 100% precision with application to learning semantic mappings. arXiv:1606.06368.

Korman, D.Z. (2019). Debunking arguments. *Philosophy Compass*, 14(12)e12638.

Korman, Daniel Z. & Locke, Dustin (2020). Evolutionary Debunking and Moral Relativism. In M. Kusch (ed.), *The Routledge Handbook of Philosophy of Relativism*. Routledge.

Krakovna, V. (2022) AI alignment resources. URL:https://vkrakovna.wordpress.com/ai-safety-resources/

Krakovna, V. (2023) "Specification gaming examples in AI - master list" URL:https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml

de Lazari-Radek, K. & Singer, P. (2012). The Objectivity of Ethics and the Unity of Practical Reason. *Ethics* 123(1):9-31.

Lee, A.Y. (2022). Speciesism and sentientism. *Journal of Consciousness Studies*, 29(3-4)205-228.

Liscio, E., van der Meer, M., Siebert, L. C., Jonker, C. M., & Murukannaiah, P.K. (2022). What values should an agent align with? An empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems*, 36(1)23.

Long, R. (2020) Nativism and empiricism in artificial intelligence. URL:https://robertlong.online/wp-content/uploads/2020/10/Nativism_and_empiricism_in_AI_9_30_2020_for_sharing.pdf

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., ... & Callison-Burch, C. (2023). Faithful chain-of-thought reasoning. arXiv:2301.13379.

MacAskill, W. (2022). *What we owe the future*. Basic books.

MacAskill, W.; Bykvist, K. & Ord, T. (2020). *Moral Uncertainty*. OUP.

MacAskill, W., Vallinder, A., Oesterheld, C., Shulman, C., & Treutlein, J. (2021). The Evidentialist's Wager. *The Journal of Philosophy*, 118(6)320-342.

Mogensen, A. L., & MacAskill, W. (2022). Evolution, Utilitarianism, and Normative Uncertainty: The Practical Significance of Debunking Arguments. *J. Ethics & Soc. Phil.*, (22)338.

Morrison, J. (2017). Evidential holism. *Philosophy Compass* 12(6):e12417.

Ngo, R., Chan, L., & Mindermann, S. (2023). The Alignment Problem from a Deep Learning Perspective: A Position Paper. In *The Twelfth International Conference on Learning Representations*.

Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative & Physiological Psychology*, 47:419–427.

Omohundro, S.M. (2008) The basic AI drives. In *Proc. 2008 Conference on Artificial General Intelligence* 483–492 (ACM, 2008).

OpenAI (2017) Learning from human preferences. URL:https://openai.com/research/learning-from-human-preferences

Ord, T. (2020). *The precipice: existential risk and the future of humanity*. Hachette Books.

Newberry, T. & Ord, T. (2021): "The Parliamentary Approach to Moral Uncertainty", Technical Report #2021-2, Future of Humanity Institute, University of Oxford.

Parfit, D. (1984). *Reasons and persons*. OUP.

————(2011) *On what matters: Vol. 2*. OUP.

Park, P., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023). AI Deception: A survey of Examples, Risks, and Potential Solutions." URL: https://arxiv.org/pdf/2308.14752.pdf

Peterson, M. (2019) "The value alignment problem: a geometric approach." *Ethics and Information Technology* 21:19-28.

Pollock, J. L. (1987). Defeasible reasoning. *Cognitive Science*, 11(4), 481–518.

Ratoff, W. (2021). Can the predictive processing model of the mind ameliorate the value-alignment problem?. *Ethics and Information Technology*, 23(4), 739-750.

Rawls, J. (1971). *A theory of justice*. Oxford.

Rini, R.A. (2016). Debunking debunking: A regress challenge for psychological threats to moral judgment. *Philosophical Studies*, 173:675-697.

Rowland, R. (2017). The epistemology of moral disagreement. *Philosophy Compass*, 12(2), e12398.

————(2019). Local Evolutionary Debunking Arguments. *Philosophical Perspectives* 33(1):170-199.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control.*

Penguin.

Saad, B., & Bradley, A. (2022). Digital suffering: why it's a problem and how to prevent it. *Inquiry*, 1-36.

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., & Leike, J. (2022). Self-critiquing models for assisting human evaluators. arXiv preprint arXiv:2206.05802.

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, (39):98-119.

————(2020). Designing AI with rights, consciousness, self-respect, and freedom. In M. Liao (ed.) *The ethics of artificial intelligence*. OUP.

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., & Kenton, Z. (2022). Goal misgeneralization: Why correct specifications aren't enough for correct goals. arXiv preprint arXiv:2210.01790.

Shulman, C., & Bostrom, N. (2021). Sharing the world with digital minds. In S. Clarke, H. Zohny, and J. Savulescu (Eds.), *Rethinking moral status*. OUP.

Sider, T. (2011). *Writing the Book of the World*. OUP.

Sinclair, N. & Chamberlain, J. (forthcoming). The Evolutionary Debunking of Quasi-Realism. In Diego E.M. (ed.), *Evolutionary Debunking Arguments: Ethics, Philosophy of Religion, Philosophy of Mathematics, Metaphysics, and Epistemology*. Routledge.

Singer, P. (1977) *Animal Liberation*, New York: Avon Books.

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

Soares, N. and Fallenstein, B. (2014). Aligning Superintelligence with Human Interests: A Technical Research Agenda. Technical Report. *Machine Intelligence Research Institute*.

————(2017). Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In *The technological singularity: Managing the journey*, 103-125.

Soares, N. (2023) "AI alignment researchers don't (seem to) stack" URL: https://www.lesswrong.com/posts/4ujM6KBN4CyABCdJt/ai-alignment-researchers-don-t-seem-to-stack

Street, S. (2006). A Darwinian dilemma for realist theories of value. *Philosophical Studies*, 127:109–166.

Street, S. (2011). Mind-independence without the mystery: Why quasi-realists can't have it both ways. *Oxford studies in metaethics*, 6(1)1-32.

Tersman, F. (2017). Debunking and Disagreement. *Noûs* 51(4):754-774.

Thornley, E. (forthcoming) The Shutdown Problem: An AI Engineering Puzzle for Decision Theorists. *Philosophical Studies*.

Tropman, E. (2014). Evolutionary debunking arguments: moral realism, constructivism, and explaining moral knowledge. *Philosophical Explorations* 17(2)126-140.

Turner, A. M., Hadfield-Menell, D., & Tadepalli, P. (2020, February). Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 385-391.

Weatherson, B. (2019). *Normative Externalism*. OUP.

Wiegman, I. (2017). The evolution of retribution: Intuitions undermined. *Pacific Philosophical Quarterly*, 98(2)193-218.

Yampolskiy, R.V. (2014). Utility function security in artificially intelligent agents. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3), 373-389.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*.

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

————(2015) Nearest unblocked strategy. URL:https://arbital.com/p/nearest_unblocked/.