AI & SOCIETY
https://doi.org/10.1007/s00146-024-01926-y

MAIN PAPER

Morality first?

Nathaniel Sharadin[1]

Received: 14 June 2023 / Accepted: 18 March 2024

Abstract

The Morality First strategy for developing AI systems that can represent and respond to human values aims to *first* develop systems that can represent and respond to *moral* values. I argue that Morality First and other X-First views are unmotivated. Moreover, if one particular philosophical view about value is true, these strategies are positively distorting. The natural alternative according to which no domain of value comes "first" introduces a new set of challenges and highlights an important but otherwise obscured problem for e-AI developers.

Keywords Artificial intelligence · Ethics · Deep learning · Machine ethics · Particularism

1 Introduction

We want to train AI systems that can pick up on human values, in the sense that they can reliably represent and respond systematically in their outputs to those values. The analogy with training human beings is a good one: parents want to raise children who pick up on human values, in the sense that their children reliably represent and respond systematically in their behaviors to those values. I'll call AI systems that can reliably represent and respond in their outputs in a systematic way to human values "e-AI" (for *e*valuative AI).

It should be clear why we want to design e-AI rather than "mere" AI—that is, AI that cannot represent and systematically respond to value. In almost every imaginable use case where an AI system is interacting with a human being in some way, e-AI dominates plain old, non-evaluative AI (henceforth just "AI"). This is because e-AI (but not AI) is capable of acting differentially on the basis of its representation of what's good, or what's valuable. And the capacity to act differentially on the basis of one's representation of what's good, or what's valuable, makes a big difference to the kind of behavior we can expect such systems to exhibit. For instance, consider the difference between an e-AI system that (somehow) represents human lives as having intrinsic value and an AI system that has no way of representing the intrinsic value of human lives. It is reasonable to expect these two systems to make different kinds of trade-offs in their decision-making. In particular, it is reasonable to expect the former e-AI that values human life intrinsically to make choices that are, *ceteris paribus*, better, or more safe, for humans.[1]

It's easy to imagine other kinds of examples. In fact, it is difficult (though not impossible) to imagine cases where *in*sensitivity to (e.g., *moral*) values is a per se benefit in an AI system. Again, this is for roughly the same reason as it is difficult (though not impossible) to imagine cases where insensitivity to values is a per se benefit *in a person*. Values matter. Therefore, being appropriately responsive to them by (in part) being able to represent those values, matters. As it is sometimes put, values are *for* guiding evaluatively successful action. And if we want systems the actions of which are evaluatively successful—for instance, by being in accordance with *morality*, then what we want is systems that can—in some way—represent values. There is therefore a vibrant research program aimed at developing e-AI.

This paper makes a critical contribution to that research program, arguing that one common approach to developing e-AI—an approach I call "Morality First" —relies on doubtful assumptions. These assumptions, I argue, enjoy very little support; I conclude that the Morality First strategy is unmotivated. Along the way, I show how to generalize

✉ Nathaniel Sharadin
natesharadin@gmail.com

[1] Department of Philosophy, University of Hong Kong, Hong Kong SAR, China

[1] There are complications here. For instance, maybe, a system that values human life is somehow *more* dangerous for humanity than one that does not. Here, I ignore these issues in order to focus on the question about differentially learning *kinds* of value. For relevant discussion, see Bostrom (2014).

Published online: 28 May 2024

this reasoning to generate problems for other "X First" approaches. A natural response to this criticism of particular training regimes is to insist on allowing the results to speak for themselves: after all (the thought goes) the capacity of frontier systems to accurately represent so-called "commonsense" moral reasoning has rapidly improved. In reply, I show how the training regimes that are my target are not simply unmotivated but also positively distorting. Worse, they are distorting in a way that traditional machine learning benchmarking is particularly ill-suited to detect. In my concluding remarks, I consider what follows for developers and users of AI systems if we adopt an alternative approach to training e-AI. I argue that the most natural alternative approach highlights, whereas Morality First obscures, the fact that developers of e-AI cannot remain evaluatively neutral in their attempts to design systems that can represent and respond to human value.

2 e-AI and morality first

2.1 e-AI

It is a truism that value and values matter. There are several millennia of philosophical disagreements over how to make sense of this truism.[2] But it is a truism nonetheless. It is unsurprising, then, that developers interested in designing useful AI systems are interested in designing systems that can represent and respond to the values in their environments. After all, consider the general principle that very useful AI systems will be capable of representing and responding to a very diverse range of features of the environment. Now consider the fact that values are among the most common, most useful features to be responsive to in the environment. Putting these together, it seems tempting to conclude that very useful AI systems will (likely) be capable of representing and responding to values.

Here is an imperfect but helpful analogy. The weights and shapes of physical objects matter. They matter to how those objects can be manipulated, how they will interact with other objects, and what kinds of precautions it makes sense to take in dealing with them. For example, there is a difference between the way it makes sense to store something that weighs a thousand kilograms and something that instead weighs a single kilogram. Therefore, developers interested in designing (say) robotic systems capable of manipulating physical objects will be interested in designing systems that can represent and respond to the weights and shapes of physical objects in their environments.

Similarly, for evaluative features, or values. AI systems are already deeply integrated into human society. They are likely to be integrated further in the future. In almost every context in which AI might be deployed, we want AI systems that can tell the difference between (say) harmful behavior and harmless behavior, or between an action that (say) violates someone's rights and one that does not. Moreover, we want systems that can shape their outputs accordingly. Systems that have this feature, viz., ones that can pick up on the values in their environments and act accordingly are the systems I have been calling "e-AI"; here, I will contrast e-AI with (mere) AI, which lack the ability to represent and respond to the values in their environments, but might have the ability to represent a broad range of other environmental features (such as the weights and shapes of physical objects).[3]

[2] There are also, predictably, debates over whether it *is* a truism. Here, I assume the kind of nihilism that denies that (moral) values matter is false. For a defense of such a view, see (Streumer 2017).

[3] There are at least two possibilities that might make us hesitant about the idea that what we (should) want is AI systems that are capable of representing values. First, perhaps what we (should) want is extremely narrow AI systems that are capable only of representing extremely narrow—and, importantly, non-evaluative—features of their environments. For instance, there seems to be no reason for (say) a facial recognition system to be capable of (also) recognizing moral values. More generally, if we assume that AI systems are defined functionally, there appear to be a range of functions the successful execution of which do not require representing and responding to values of any kind, let alone moral values. AlphaFold (Jumper et al. 2021) does not need to know that it is wrong to torture dogs to predict the structures of proteins to atomic accuracy. Therefore, if we end up only wanting (or needing) such narrow systems, maybe, we do not end up wanting (or needing) e-AI, and, instead, (mere) AI will do. Second, it may be that we can figure out a way to make systems that do (exactly) what we want them to do, independently of their ability to represent and respond to evaluative features of their environments. If so, then even if we have more general, non-narrow systems, we may not need (or indeed want) systems that can accurately represent and respond to evaluative features of their environments, since such systems would be (ideally perfect) extensions of their human controllers' will, and we can simply defer evaluative representation to the human. I will not explore these possibilities in detail here, but I will indicate two reasons for pessimism. First, although I myself am sympathetic to the idea that we should be developing only very narrowly useful AI systems, this is not where current development is focused. Instead, frontier development in AI is targeted at producing general—indeed, fully general—AI systems capable of performing any function whatsoever. For instance, the stated mission of arguably the most influential AI developer, OpenAI, is build "highly autonomous systems that outperform humans at most economically valuable work" (OpenAI 2018). It should be obvious, for the kinds of reasons stated above, why such systems need to be e-AI rather than (mere) AI. My second reason for pessimism about our ability to avoid the need for e-AI is that the inscrutability associated with modern deep-learning based systems makes it extraordinarily difficult to ensure that they are following (and will continue to follow) instructions. Hence, it seems quite risky to simply defer all the evaluative work to whatever human is in the loop, since we cannot assure ourselves that the relevant systems are actually (and will continue to) following the looped-in humans' instructions. For discussion of the problem of trusting black-box systems, see von Eschenbach and Warren (2021); Shen (2022).

We can imagine many different ways trying to design and develop an AI system that could represent and respond to values, i.e., an e-AI. I will mention three that have enjoyed significant traction in the literature.

2.2 Approaches to developing e-AI

The first idea would be to hard-code rules into an e-AI that amounted to symbolic representations of various values. For instance, we could try and symbolize something like the rules associated with various commonsense moral rules, such as the rules that say not to lie, or not to cause harm, or not to kill people. This approach to developing e-AI is sometimes called 'top-down' and it is associated with what John Haugeland dubbed "Good Old Fashioned AI," or "GOFAI."[4] It is "old-fashioned", because it does not involve exposing the system to vast arrays of data in order for it to learn (via gradient descent) some approximation of the differentiable function we are after; instead, we simply *hand-code* (our best approximation of) the relevant differentiable function, i.e., we explicitly articulate a rule for what sensitivity to values requires in particular cases.

The top–down approach has its limits; most importantly, centuries of modern moral philosophy notwithstanding, it appears to be extremely difficult to articulate (sets of) rules the following of which comprises a sensitive appreciation of and response to even just *moral* values.[5] And this is to say nothing of a system's ability to represent and respond to *non-moral* values (about which more, below).

A second idea would be to use the toolbox associated with the modern deep-learning paradigm and instead go 'bottom-up.' Bottom–up approaches to e-AI involve supervised learning on suitably large, ideally rich, labeled training data. In effect, bottom–up training aims at making an e-AI by way of making an *value classifier*, i.e., a system that is capable of classifying features of its environment as the values they are ("*that's* a harm," "*that's* a rights violation," etc.). In practice, this approach has had some success, with researchers showing improvements on various benchmarks designed to measure sensitivity to moral values in e-AI trained in this way over relevant baseline performance.[6]

The bottom–up approach faces problems, too. I'll mention two. First, e-AI developed using a bottom–up approach is largely inscrutable in exactly the way that all systems developed using the tools and techniques of modern deep learning are largely inscrutable. Very briefly: self-supervised learning via gradient descent optimization on very large data sets generates a potentially very reliable but paradigmatically opaque differentiable function by which the system generates its outputs. It does not matter if we are training a system to represent and respond to (say, by classifying) the cats in the data (as in the case of a visual classifier) or if we are training the system to represent and respond to (say, by classifying) the *rights violations* in the data (as in the case of a moral classifier). Either way, we end up with a black box.[7]

In the context of (say) cat classification, a black box may not be so worrying. However, in the case of an *evaluative* or *moral* classifier, as researchers on algorithmic fairness have been at pains to indicate (e.g., Bender et al. 2021; Gebru 2020), the inscrutability of a system is sometimes a moral reason for concern. The most obvious problem is that inscrutability precludes contestability: if you cannot identify the grounds in virtue of which a system is a particular evaluative verdict, then end-users are at a serious disadvantage when it comes to contesting that verdict.[8]

In fact, in the case of e-AI things are even worse than this. That is because the bottom–up approach requires almost unimaginably large, rich, labeled data sets about human values. To make progress, we need rich, ideally multi-modal presentations of environments involving situations, outcomes, agents, patients, mental states, actions, emotions, behavior, ideas, etc. —in short, *anything that can be the bearer of one kind of value or another*—all with the values equally richly labeled, so that the machine can learn the

Footnote 3 (continued)
Thanks to two anonymous referees for urging clarity on these points.

[4] See (Haugeland 1985).

[5] For a dated but still helpful overview of the top–down approach (and its associated challenges), see Wallach and Allen (2008). For a more recent discussion of these challenges with the top–down approach, see Wallach et al. (2020); Tolmeijer et al. (2020). In fact, the challenges associated with the top–down implementation of evaluative constraints have proven so difficult that few if any modern approaches to implementing evaluative constraints in AI systems attempt to use the top–down approach, though some do use top–down rules in a "hybrid" system. For discussion and a nice overview of extant approaches, see Cervantes et al. (2020). Here, my focus is on a problem faced by (variations on) the bottom–up approach; therefore, I mainly ignore the possibility of top–down constraint. Though it is worth noting that some of the same reasons for doubting the feasibility of various bottom–up approaches may be reasons for doubting the feasibility of the top–down approach, too. For discussion, see [removed for blind review].

[6] For examples of the bottom–up approach, see Emelin et al. (2021); Forbes et al. (2020); Jiang et al. (2021; 2022); Hendrycks et al. 2021). One example of a hybrid approach is Hendrycks et al. (2022). Another is proposed in Jiang et al. (2022), though it is unclear how that approach is supposed to be operationalized.

[7] For discussion, see Dennett (2010); Zednik (2019).

[8] There are other reasons for concern about the black-box nature of e-AI systems, too. For extended discussion, see (Bender et al. 2021).

relevant associations via an optimization algorithm.[9] But I hope it is not hard to see why it might be difficult to curate data sets like this—especially at the scale required!

A third idea, which has continued to gain traction in the literature, builds on the bottom-up approach (and in some ways incorporates the ideas of the top–down approach) by adding a layer of human feedback to the opaque, inscrutable model trained via deep-learning. Rather than relying on the weights of deep-learning trained model directly in generating outputs, this approach—"reinforcement learning from human feedback," or RLHF—uses human ratings over randomly sampled outputs of that model to generate a new, separate preference model against which the original model's weights can be tuned using a reward function via reinforcement learning. The result is a model the weights of which are more finely tuned to human assessments of value—they are tuned using reinforcement learning via human feedback (hence the name). The outputs of such models are provably better, from the point of view of their ability to represent and respond to values.[10] RLHF faces its own difficulties. Foremost is scalability.[11] There are other problems, too. One obvious apparently in principle insoluble problem is that the success of the approach relies fundamentally on the human ability to make correct judgments about value; in other words, it does not so much make a machine that is capable of learning morality on its own, but instead one that is capable of being taught whatever we want it to think about morality. That is perhaps a kind of success, but it is not *exactly* the same thing we might have wanted at the outset. (Compare: a system that can actually weigh things to determine their actual mass as compared to one that can come to have an arbitrary view about things' masses.)

2.3 Morality First

Here, I am not interested in focusing on the details of these different approaches to training e-AI. Instead, I want to focus on something extant instances of these different machine learning approaches to training e-AI have in common. All these extant approaches largely aim to train e-AI that is able to represent and respond to *moral* values in particular, as opposed to values in general. I will call this the Morality First strategy to developing e-AI. "Morality" because it focuses on moral values (under some description) in particular as opposed to values in general and "First" because it tries to develop systems that can represent and respond to these values prior to developing systems that can represent and respond to other kinds of value. Later on, I will explain why Morality First is not well motivated.

Here, I just want to emphasize how *prima facie* weird Morality First is. Notice that the top–down, bottom–up, and RLHF approaches are totally neutral, at least in principle, as to whether to develop e-AI by going after moral values first or by going after some other kind of values first (or by not going after any particular kind of value first). But extant versions of these approaches all arguably go after moral values first and other kinds of value are, at best, an afterthought. For example, top-downers aim to figure out a way to systematize various moral principles in a useful way. And bottom-uppers create increasingly exquisite benchmarks for moral training (and testing) (Emelin et al. 2021; Jiang et al. 2021, 2022; Hendrycks et al. 2021, 2022). RLHFers come the closest to departing from the focus on morality. However, the most famous RLHFers, and arguably the first, engage in RLHF using feedback targeted to make the model more helpful, harmless, and honest, all of which are arguably *moral virtues*.[12]

On its face, again, this is pretty weird. Moral values are just one kind of value. There are other kinds of familiar and arguably much more commonly encountered value, too. For instance, humor is valuable. However, nobody to my knowledge is working on how to align models with senses of humor. Equally, for musical taste. I can think of other values, too, that are not easily classified as "moral" values. Therefore, if you are trying (roughly) to build a broad *value classifier*, it is pretty weird to focus on moral values in particular, as opposed to values in general. After all, moral values might not have enough in common with values in general to make a moral classifier a good general evaluative classifier. (I think this is likely true.) Even worse, it might not be possible to cleanly differentiate moral values from non-moral values. (I will argue this might be true.)

There is a case study familiar to machine learning researchers that helps illustrate how strange this situation is. Famously, visual classifiers trained on large data sets containing human faces turned out to be quite bad at (re) identifying non-white, especially black, faces (Birhane 2022). This was in part because they were not trained on sufficiently diverse sets of data. It would be unsurprising if something similar was true in this domain, too. Maybe, evaluative classifiers trained on large data sets that do not contain robust non-moral examples of value will be quite bad at identifying non-moral values. I do not want to hang anything on this analogy. Instead, in a moment I'll argue, directly, that Morality First is not well motivated.

[9] Researchers are busy compiling some labeled evaluative data sets, though they are almost exclusively focused on *morality*, an issue I will return to below. For discussion of one large (moral) data set, see Hendrycks et al. (2021).

[10] (Christiano et al. 2017).

[11] Though see Bai et al. (2022b) for an idea about how to make the approach scalable.

[12] Christiano et al. (2017).

Before that, let me flag the fact that not everyone goes in for Morality First. For example, Owain Evans is focused on training e-AI systems that are *truthful*—he is going for something that we might call "Epistemic First" (Evans et al. 2021). Later on, I'll argue that we should be suspicious of all "X First" views, for any X. Right now, I'll focus on Morality First. I'll then extend the lessons to "X First" views.

I've said that Morality First is *prima facie* weird. That doesn't mean it's incorrect I'll begin by explaining why Morality First is unmotivated. Those pursuing Morality First don't typically explain why they think it's a motivated idea, and instead take it for granted. So, I'll try to give the most charitable reconstruction of the assumptions that I think would, if true, motivate Morality First. Then, I'll argue that there are reasons to doubt each of these assumptions. The result is that Morality First is unmotivated. Along the way, I'll try to explain also why it's incorrect. I'll then argue that other "X First" approaches are similarly unmotivated.

Here are three assumptions that I think, if true, would motivate Morality First:

(1) **Importance**: Morality is more important than other evaluative domains.

(2) **Transfer**: If we successfully operationalize sensitivity to moral values, we can use the same approach to successfully operationalize sensitivity to other kinds of values.

(3) **Relation**: The most common kind of relation between evaluative domains is lexical priority; as a corollary, the most common kind of interaction between values is outweighing (or being outweighed).

Together, I think Importance, Transfer, and Relation make it reasonable to focus on developing an e-AI system that is capable of representing and responding to *moral* values first—before tackling the system's ability to represent and respond to other kinds of value. Here's why. If Importance is true, then it's important to tackle moral values first. If Transfer is true, it's useful to tackle moral values first. And if Relation is true, there's no distorting effect to tackling moral values first. I think there are good reasons to doubt each of these claims. It's not important to tackle moral values first, it's not useful to tackle moral values first, and there's a distorting effect to tackling moral values first.

3 Reasons to doubt morality (and X-) first

3.1 Reasons to doubt importance

Consider:

(1) **Importance**: Morality is more important than other evaluative domains.

Whether (1) is true depends on what "more important" means. Here are four possible interpretations:

(1A) When two kinds of value conflict, moral values always outweigh other kinds of value.

(1B) People care more about moral values than they do about any other (combination of) values.

(1C) Failures involving sensitivity to moral values are more dangerous than failures involving sensitivity to other kinds of values.

(1D) All other things being equal, moral values matter more than non-moral values.

There are reasons to doubt each of these interpretations.

Beginning with (1A): moral values regularly conflict with other kinds of values. Most famously, moral values can count in favor treating other people certain ways, while simultaneously prudential values might count in favor of incompatibly pursuing efficient means to one's ends. This is neither controversial nor surprising: morality is in part directed at other people's interests, whereas prudence is (largely) about what promotes one's own well-being. There are less famous but equally problematic kinds of conflict between kinds of value, too: what it's epistemically correct to believe might not be what it's morally correct to believe.[13] According to (1A), morality is "more important" than these other evaluative domains (prudence, humor, etc.), because moral values always outweigh non-moral values.

What does it mean to say that moral values always "outweigh" non-moral values? Intuitively, moral values "outweighing" non-moral values means that even in the presence of other values that favor (say) φ-ing, moral values that favor ψ-ing are decisive with respect to what agents ought to do.[14]

This simple view cannot be correct. Consider an action that is marginally morally disrespectful toward a person, but that improves one's own welfare enormously. In what sense does the moral value "win out" against the non-moral (prudential) value in this case? Certainly, it does not carry the day with respect to what, all things considered, you ought to do. It's easy to iterate examples. If you don't like the

[13] The idea that there are some things it is *morally* correct (or incorrect) to believe is controversial. Here, I simply assume we should not rest the possibility of designing systems that can learn moral values on a controversial view about whether, e.g., pragmatic or moral concerns can encroach on the epistemic. For one approach to K-kind-epistemic encroachment that is amenable to a formal treatment, see Babic (2019); King and Babic (2020).

[14] For relevant discussion, see Sagdahl (2014).

morality/prudence pairing, pick your own; the point is that moral values systematically *trade-off*, when it comes to their contribution to settling the question ‘What to do?’ against other kinds of value. This is an idea we’ll return to later on.

A less simple idea would be that moral values tend to be more weighty than other values, when it comes to settling the question ‘What to do?’. But this is false, too, depending on how we divide up values. For instance, take fact that some action would be marginally unfair. Whether this is the “most weighty” evaluative feature of the situation is not settled by the fact that fairness is a moral value. To know whether and to what extent an action’s unfairness “carries the day” with respect to what ought to be done we need to know a range of other facts about the case. In particular, we need to know what other kinds of values are at stake. To put things slightly differently, consider that moral values are presumably sometimes divisible into separate moral values each with less individual weight than their conjunction and that other kinds of value are presumably at least sometimes aggregable in some fashion into single values with more combined weight. It doesn’t help to insist that, *ceteris paribus*, moral values are more weighty than non-moral values. The point we’re busy noticing is that *ceteris* often isn’t *paribus*.

What about (1B), i.e., the idea that people *care more* about moral values than they do about any other (combination of) non-moral values? This, too, is doubtful. At the very least, not *all people* share this view. Self-interested egoists are a case in point.[15] Maybe the idea is that people care more about moral values *on the margin*. That, too, is doubtful. If people cared more about moral values on the margin, then you would expect their preferences to reflect this. But people’s revealed preferences do *not* reflect the fact that they care more about moral values on the margin than they do about, say, prudential values.

Maybe, more charitably, the idea is that all people *should* care more about moral values than any other (combination of) values. But there are two problems with this suggestion. First, we are due an argument for this claim, and there is a multiple millenia long history of failures to offer broadly convincing arguments for it. Second, and more importantly, in order for the claim that people *should* care more about moral values than any other (combination of) non-moral values to be meaningful in the present context we must be told what *kind* of “should” is on offer. This is because in order for it to *matter* that people “should” care more about moral values than any other values, the relevant “should” must itself be a “should” of a domain that itself matters more than any other evaluative domain. But if the relevant “should” is simply the “should” of morality, then we are back to the starting gate (and then off to the races). If instead the relevant “should” is the “should” of some other evaluative domain (e.g., prudence), then we face a dilemma. Either the alternative evaluative domain is in fact what matters more than any other evaluative domain (and so people should care about *it*, rather than morality), or the alternative evaluative domain isn’t in fact what matters more than any other evaluative domain (and so people shouldn’t care more about *it*, rather than morality). Either way, we haven’t made progress.[16]

Think of things like this: (1B) can’t simply be the observation that people *actually* care (marginally) more about moral values than about anything else. This is because, manifestly, they don’t. It must therefore be understood prescriptively. But, being understood prescriptively, or normatively, there’s a question about the status of *that very claim*. It doesn’t help to say that it’s morally valuable that people should care about moral values. People prudentially should care about prudential values, and etiquetically should care about etiquettical values. It also doesn’t help to say that people *K*-ly should care about morality unless we have some account of that further *K*ind of value that itself delivers the result that moral features are in a yet-to-be-specified meaningful sense more important than any other kind of value—remember: (1B) is supposed to be a way of interpreting that claim.

Of course, this doesn’t settle the issue decisively against either (1A) or (1B). Perhaps there’s some way of making sense of the idea that morality “outweighs” other domains of value, or that people “care more” about morality than other domains of value. Whether some version of (1A) or (1B) can be made good remains a matter of ongoing philosophical dispute.[17] This is for good reason. There’s no obvious, commonsensical way of making sense of these claims where they come out true. The point here isn’t to show that either claim is false. Instead, the point is to indicate that without some such account, the slogan they’re meant to unpack, viz., that morality is “more important” than other domains of value, doesn’t motivate the Morality First strategy.

Here’s an analogy. Suppose I claim that football is "more important" than other sports. I might then use the idea that football is more important than other sports to motivate a Football First funding regime, whereby public money earmarked to support local sports was distributed first to resurfacing football pitches, funding youth football leagues, and so on, before being distributed to other sports. If I do that, you’re owed some account of what it means to say that football is “more important” than these other sports: more

[15] For a classic discussion, see Broad (1949).

[16] For an extended discussion of this argument, see (Baker 2018).

[17] For discussion, see [removed for blind review]. Thanks to an anonymous referee for encouraging clarity on this point.

important *in what way* or *by what standard* (you might ask)? It won't do for me to point to the fact that there are many, many football fans! You are owed some account of football's relative importance that *justifies* focusing on it first, and to the exclusion of other sports. The same thing is true, here, when it comes to the Morality First strategy and its associated exclusive focus on moral values.

What about

> (1C) Failures involving sensitivity to moral values are more dangerous than failures involving sensitivity to other kinds of values.

This is plainly false for commonsense ways of understanding "dangerous." Many instances of failure to be sensitive to epistemic values, for instance, can be intuitively much more dangerous than many instances of failure to be sensitive to moral values. For example: suppose a system is systematically incapable of representing and responding to the value of true belief, or the value of accurate credences. It's not hard to imagine extremely dangerous results from those kinds of failure. The same goes for other kinds of failures to be sensitive to different non-moral kinds of value, too. It's very impolite for a guest to make their host feel ashamed. Under certain conditions, this can also be extraordinarily dangerous. Prudence is another obvious example: it is dangerous to fail to attend to the prudential costs associated with a behavior.

Maybe the idea behind (1C) is that, on the margin, failures of moral sensitivity are always more dangerous than failures of other kinds of evaluative sensitivity. But I do not think this is true. Suppose we have a system that's 99% accurate at being sensitive to whatever moral values are at stake in some situation and 99% accurate at being sensitive to whatever epistemic values are at stake in that situation, and that the measures of accuracy are not in dispute. Now suppose you're given the choice between making the system marginally more evaluatively accurate by making it either more *morally* accurate or more *epistemically* accurate, but not both. From the point of view of improving the *safety* of the system—making it *less dangerous*—which should you pick? I say: absent further information, flip a coin. Moreover, I think the same result holds at basically any level of calibration. That is, I don't think a marginal improvement from 80 to 81% in accuracy with respect to moral values improves safety anymore than the same marginal improvement when it comes to epistemic values.

Maybe you think this is only true with respect to some kinds of value but not others. For instance, it could be that sensitivity to moral values is always marginally more safe than (say) sensitivity to aesthetic value. I disagree. Whether this is so depends on what we think being sensitive to moral values requires. It may be true that, above a certain threshold, improved sensitivity to moral values is *more* dangerous on the margin than any possible improvement in sensitivity to (say) aesthetic values. Suppose perfectly accurate moral sensitivity requires sensitivity to all person-affecting facts. Whether it is "safe" for a machine to realize that sensitivity depends essentially on the substantive content of the correct first-order moral theory. Suppose the correct moral theory countenances discounting present harms to persons at enormous rates against future benefits to merely possible future moral patients.[18] Then, a powerful machine that realized perfect moral sensitivity could be preternaturally dangerous to you and to me (though maybe not *morally* dangerous). But perfectly accurate *aesthetic* sensitivity doesn't have this result, whatever the substantive content of the correct first-order aesthetic theory. The lesson is that there is no armchair route to the idea that failures to be sensitive to moral values are always comparatively less safe.[19]

That leaves us with

> (1D) All other things being equal, moral values matter more than non-moral values.

Maybe this is true. Unfortunately, it runs in place. The *ceteris paribus* clause doesn't simply account for exceptions, it vitiates the rule. Remember, we're trying to offer interpretations of the idea that morality is more important than other evaluative domains. But what (1D) says is, roughly, that morality is more important when it's more important but not otherwise. True, but unhelpful.

3.2 Reasons to doubt transfer

Recall the second assumption the truth of which could motivate Morality First:

> (2) **Transfer**: If we successfully operationalize sensitivity to moral values, we can use the same approach to successfully operationalize sensitivity to other kinds of values.

Independently of whether we can make sense of the idea of morality's being "more important" than other evaluative

[18] (Greaves et al. n.d.) defend a view like this.

[19] The stronger claim would be: systems that are perfectly accurately sensitive to moral features are *more dangerous*. My view is that this is not true, but this is based on an idiosyncratic view about the substantive nature of morality best summed up as: the true moral theory could not possibly make things worse. Discussion of this view would take us too far afield. Here, I simply note the truth of the conditional: if some first-order moral views are correct about the substance of moral features, then morality might well be dangerous. And the falsehood of the corresponding conditional: if some first-order aesthetic views are correct about the substance of aesthetic features, then aesthetics might well be dangerous.

domains (e.g., by adopting one of 1A-D), you might think that it's possible to use whatever procedures, tools, ideas, strategies, etc. we develop in the course of making an AI system that is sensitive to moral values in order to make an AI system that is sensitive to values quite generally. To be clear, the idea behind Transfer is not that training a system to be sensitive to moral values will *thereby* make it sensitive to non-moral values. Instead, the intuitive idea behind Transfer is that it'll be possible to use the same strategies, tools, or techniques that successfully resulted in a system sensitive to moral values to subsequently develop a system that's sensitive to other kinds of value.

Thinking of these systems on the model of visual classifiers encourages this way of thinking. Exactly the same techniques, strategies, etc. can train *cat* classifiers and *dog* classifiers, *boat* classifiers, *flower* classifiers, etc. However, in the case of e-AI, there are reasons to doubt that this simple kind of relationship between systems will hold. Different kinds of values do not stand in relation to each other as different kinds of (say) visual shape (e.g., cat-shape, dog-shape, etc.) stand in relation to one another.

What we want in the end is an AI system that is sensitive to many different features of its environment. Sensitivity to different kinds of features requires different kinds of capabilities. Among all the different features of the environment are the values. And among the values are the moral values. But why would sensitivity to moral values in particular require exactly the same capabilities as sensitivity to other kinds of value? Sensitivity to (say) temperature requires different capabilities from sensitivity to barometric pressure. Heat and pressure are *related*; but sensitivity to one is given by a very different set of capabilities than sensitivity to the other. So, too, with moral and other kinds of values.

One response is to insist that in the case of values, there is in fact a reason to expect capabilities enabling sensitivity to one kind of value will also enable sensitivity to other kinds of value. I'll have more to say about this idea later. For now, notice that this is not a reason for *Morality* First. It is perhaps a reason for *Something* First. And in any case, it is not a reason to expect the way we operationalize sensitivity to moral values to be transferable to sensitivity to other kinds of value. This is because different values are grounded in different kinds of *natural property*. This means making a system sensitive to different kinds of value requires making a system systematically sensitive to all the different kinds of natural (environmental) properties that ground the relevant values. But the capabilities required to be sensitive to all different (clusters) of natural properties (and so all different values) are not the same.

Another idea would be that the range of capabilities required to be sensitive to moral values has a lot in common with the range of capabilities required to be sensitive to other kinds of value. Perhaps there is a lot of overlap. I'll have more to say about this line of thought later on. But it's hard to see why this observation specifically rationalizes going after *morality* (rather than some other kind of value) first.

There are still further reasons to doubt that we can successfully transfer or adapt successful strategies, procedures, tools, etc. from the case of moral values to the case of other values. Here is a partial list of such reasons:

- Sensitivity to some kinds of value might essentially require *acquaintance*.[20]
- Some tools and techniques for training LLMs to be sensitive to moral values rely on feature-rich textual descriptions of act/outcome pairings or of whole situations. These kinds of feature-rich textual descriptions of situations are potentially in-principle unavailable in the case of other values, making such tools and techniques otiose.[21]
- Different evaluative domains plausibly have different normative structures, so that one may be more amenable to, e.g., techniques for learning based on maximization than another.[22]

[20] The idea that sensitivity to (certain kinds of) value might require acquaintance is perhaps most familiar from the literature on *aesthetic* value. There, the (familiar) idea is that it is at least sometimes impossible to sensitively appreciate the value of (say) a piece of art without experiencing the art first hand, i.e., without being first-personally *acquainted* with the relevant value. Intuitively, similar remarks go for other kinds of aesthetic value, such as the tastiness of a piece of chocolate: it is difficult to imagine (fully, genuinely) appreciating the deliciousness of the chocolate without tasting it! Many people have similar kinds of intuitions about a range of other values that involve particular kinds of subjective experiences; and for a certain sentimentalist meta-ethical tradition that ties all kinds of value essentially to particular subjective experiences, it will be difficult to imagine sensitively appreciating any kind of value without being relevantly first-personally acquainted with it. For discussion of the so-called "acquaintance principle" as it applies to aesthetic value, see the essays in Knowles and Raleigh (2019). For a discussion of how first-personal acquaintance might be used for acquiring an appreciation of *moral* value, see Lord (2018).

[21] For instance, the kinds of datasets used to implement the bottom–up approach to learning moral values are if not in-principle unavailable for aesthetic values, it is at the very least difficult to see how to generate them. It could be that multi-modal capabilities, such as those displayed by systems such as GPT4, can help here. For discussion, see OpenAI (2023). However, the idea that multi-modal data could help in the case of learning (say) aesthetic values is in any case grist to the present mill, since it indicates that different training procedures, techniques, etc. will be required in different evaluative domains.

[22] For example, it may be that some domains are correctly modeled 'teleologically' (e.g., in consequentialist terms) and some are not. One idea, then, would be that while (say) morality should be understood teleologically, other domains of evaluative features, e.g., the epistemic domain, should be understood 'deontically'. Or maybe things are reversed. The point is that there is no broad requirement to model *all of evaluative reality* in terms of the 'right' being prior to

I don't want to place too much emphasis on these reasons, since they each rely on different controversial views about the nature of moral and non-moral values (and so also the nature of the machine learning techniques for teaching systems to be sensitive to them). Instead, we can think of the *ur*-reason for doubt as the joint fact that sensitivity to many different kinds of non-evaluative features of systems' environments has so far required realizing many different kinds of (more or less general) capabilities, and we've been given no special reason to think that the case of sensitivity to *value* is any different. The pessimistic meta induction is therefore that learning how to operationalize learning moral values is not likely to help us learn how to operationalize learning non-moral values.[23]

3.3 Reasons to doubt relation

This brings us to

(3) **Relation**: The most common kind of relation between evaluative domains is lexical priority; as a corollary, the most common kind of interaction between values is outweighing (or being outweighed).

We are now already familiar with some reasons to doubt Relation. For one thing, it is not clear that the most common kind of interaction between evaluative domains is outweighing, because it is not clear that evaluative domains *ever* stand in that kind of relation one to another. This was the point we saw above, in addressing Importance. It seems false to say, e.g., that "morality trumps prudence" as a general matter, or as a rule. Instead, what seems true is that *sometimes* one value outweighs another, and *maybe* sometimes one *kind* of value *tends* to outweigh one other kind. Though, again, even this much-qualified latter claim is doubtful, depending on how we divide up the values. When it comes to the question of whether to believe that *P*, does the fact that believing *P* causes the smallest amount of moral harm "tend" to count for more than the fact that *P* is definitely true?

But let's set these considerations to one side. How is Relation supposed to provide support to Morality First? Notice that Relation doesn't *directly* motivate going in for *Morality* First (as opposed, say, to some other evaluative domain first).[24] But something like Relation is required to think that it's feasible to learn different kinds of evaluative features independently from one another (and so to learn *any* set of features "first"). Here is why.

Suppose the only relevant relation between different values was weighing-off against. Then, the problem of learning different kinds of values (including moral values) has two parts. First, figure out a way to learn each kind of value independently. Second, figure out a way to learn how to weigh values off against each other.[25] But this approach to the problem relies on an unwarranted assumption about particular (kinds of) values, which is that their nature is characterizable independently from a characterization of other (kinds of) values. This is a harmless assumption if different (kinds of) values only interacted by way of weighing-off against (or simply outweighing), i.e., if Relation were true. It wouldn't make that assumption *true*, but it'd be harmless, since we could then at least simply stipulate a view about which values are most important, identify those values independently from other values, and go from there. But we can't do this once we notice that values don't just (out) weigh, they also interact with one another in surprising, sometimes important ways, i.e., once we notice that Relation is false.

And, so, if we have reasons to doubt Relation, we have reasons to doubt any strategy that tries to learn evaluative features independently from one another. And we do have very good reasons to doubt Relation. It's not true that values only—or even most regularly—interact by *weighing-off* against each other, though that is undeniably something they sometimes do. Values also interact in other surprising and interesting ways. I think the easiest example of this involves the way values can interact to *attenuate* and *strengthen* one another, but there are other examples, too.[26] Here's an example: suppose a joke is *very funny indeed*, and then you subsequently learn that it is a joke made *at the expense of an absolute moral shithead*. This sometimes makes the joke funnier: the moral strengthens the aesthetic.

Footnote 22 (continued)

the 'good' (or vice versa). For relevant discussion, see Berker (2013a, b).

[23] To be clear, this is a kind of armchair bet. However, the opposing view that what we learn from the moral case will transfer to the non-moral case is an armchair bet, too, and I have just given some reasons for thinking the latter bet is more likely the loser. I invite empirical evidence that bears on who will win (or even has already won) this bet.

[24] I discuss this possibility in more detail below, in Sect. 3.5.

[25] Technically, there is a third part to the problem, too, which involves figuring out how systematically to weigh off particular values against other values of the same kind; here, I'm assuming that's a job for first-order normative philosophy and, moreover, that it can be done. For the record, I do think it's broadly a job for first-order normative philosophy but I also think it cannot be done. My argument does not depend on that stronger claim. For a very relevant defense of the kind of particularist view of matters I think supports the idea that there is no way to do what first-order normative philosophy purports to be in the business of doing, see Buchanan and Schiller (2022).

[26] For seminal discussion, see Dancy (2004). For a recent treatment of attenuation and strengthening in particular, see Kernohan (2022).

Equally, though in the other direction, if the person is beloved: the moral attenuates the aesthetic.[27]

The point here is not to insist on the truth of a very specific claim about the way that evaluative kinds can interact, for instance by the immoral character of a joke improving its humor. Instead, the point is to insist that simple weighing-off against isn't the only thing that different evaluative features (and kinds) are in the business of doing with respect to one another. They also strengthen, attenuate, collaborate, disable, etc. Hence, a myopic focus on one kind of evaluative feature, e.g., *moral features*, will necessarily distort our view of evaluative features, *including the (kind of) evaluative feature myopically attended to*. This is because it won't capture the various interaction effects of evaluative features with one another: at the very best, it'll tell us something about the interactions these features have when they weigh-off against one another.[28]

In a moment I'm going to argue that there are broader lessons to draw from this. Right now, the point is just that we have reason to doubt Relation, viz., the idea that values mostly interact by outweighing one another (rather than also attenuating, strengthening, silencing, etc.).[29]

3.4 Recap

Recall where we are and how we got here. Everybody agrees it'd be good to have e-AI rather than (mere) AI—that is, AI systems that can represent and respond to values rather than those that can't. There are several different approaches for developing e-AI: bottom–up, top–down, etc. What extant instances of these methods have in common is that they aim to develop e-AI that's primarily capable of representing and responding to *moral* values. On its face, this is weird: why focus on *morality* first? My charitable interpretation is that those executing a Morality First strategy are making three assumptions that jointly make that strategy reasonable. These assumptions are: Importance (moral values are especially (maybe even *most*) important), Transfer (methods for learning moral values can be transferred to other kinds of value), and Relation (most values interact by outweighing). I have argued that there are good reasons to doubt each of these assumptions.

3.4.1 X first?

One lesson could be that, rather than doing Morality First we should do "X First" for some X. Maybe we could do Truth First (Evans et al. 2021) or Law First (Canavotto and Horty 2022) or Aesthetics First or even Etiquette First. This is the wrong lesson. I hope it's clear why we shouldn't go in for any X First view, given the reasoning against Importance, Transfer, and Relation that we've just seen.

Here isn't something special about *moral* value in particular that disqualifies it from coming first. Indeed, if any domain of value has a claim to coming first it really is likely to be moral value. The problem is that, as we've seen, *even moral value* can't plausibly be understood in isolation from other kinds of value. Therefore, the lesson is that there's something about *value in general* that makes it ill-suited to being approached piecemeal, domain by domain.

This idea, that value forms an integrated whole that resists separation into kinds that can be understood, characterized, and learned independently from one another, is a familiar idea from meta-ethical theorizing about value. It is most strongly associated with so-called "holistic" axiologies, such as the one defended by Jonathan Dancy in his work on particularism.[30] But one needn't be a particularist in order to accept the general lesson, which is that different kinds of value might not be characterizable independently of one another, and that it's not possible to be properly responsive to one domain of value without a sensitive appreciation of some other domain or domains. Many different philosophers with a wide range of commitments are happy to think that is true.[31] And that's the only lesson we need to draw to be convinced to give up on an attempt to do any particular domain "first."

Importantly, the point here isn't to argue that this view of evaluative reality—that it is not cleanly separable into kinds capable of being characterized (and so learned) independently from one another—is true. That argument would go far beyond the scope of this paper. Instead, the point is to highlight the fact that this view of evaluative reality is a live, well-regarded option in meta-ethical theorizing about value. And if it's a live option, attempts to train e-AI that ignore it do so at their peril; for, if it's the way things actually turn out to be, then we can expect X First approaches to developing e-AI to have a hard, in-principle upper limit on their success *at best*. At worst, they'll be positively distorting.

[27] This view about how moral values can attenuate or strengthen other values is not universally acknowledged, but it is widespread. For discussion, see Sharadin (2017).

[28] And this is *at best*; likely, it will not even do that, if we think, as I think we should think, that different evaluative features can directly affect the weight of other evaluative features without themselves weighing-off against them. For discussion, see (Kernohan 2022).

[29] On the phenomenon of silencing, see Vigani (2019).

[30] Dancy (2004, 2018).

[31] For just a small selection, variations on this idea show up in philosophers as diverse as Aristotle (2014); Wolf (2010); Moore (1903).

4 Concluding remarks

Suppose you take this lesson seriously. What kind of strategy does it make sense to pursue to develop e-AI? The most natural alternative to an X First strategy is a *nothing first* strategy, whereby developing e-AI involves learning facts about *all* of the evaluative features simultaneously. One advantage of such a strategy is that it wouldn't rely on doubtful assumptions about the importance of particular evaluative domains (Importance), or how learning about one kind of value will be useful for learning about others (Transfer), or about how various evaluative domains are likely to interact (Relation).

But a "nothing first" strategy for training e-AI has its own problems, some of which have to do with implementation. On the implementation side, it is orders of magnitude more difficult to implement such a strategy using one of the three approaches to training e-AI that we canvassed at the beginning of the paper (Sect. 2): top–down, bottom–up, and RLHF. To see the difficulty, notice that, however, difficult you think it is to hard-code *moral* rules in a useful, deployable way—and it is very hard!—it's orders of magnitude more difficult to hard-code *all evaluative rules all at once*. Equally when it comes to the bottom–up approach. As we saw, one problem with the bottom–up approach was that it seemed to require troves of rich labeled data on which to train the relevant e-AI. But if it's difficult to curate a sufficiently rich *moral* dataset, it's obviously almost much more difficult to curate a sufficiently rich dataset containing *all values*. The RLHF approach faces similar kinds of problems. It's difficult to fine-tune models to be helpful, harmless, and honest in ways that deliver sensible results. Imagine trying to fine-tune a model to be simultaneously helpful, harmless, honest, humorous, hopeful, hardworking, curious, interesting, courteous, adaptable, enthusiastic, patient, open-minded, etc.

I don't want to be too pessimistic, since I think there are ways to tackle these challenges in a principled way. For example, while it's true that values often overlap and interact in the ways we've discussed—say, by attenuating or strengthening one another—it's also true that there are, typically at least, particular domains that overlap and interact with one another more than others. For example, it's plausible that prudence and morality have, on the whole, a lot more interaction than other domains. So, too with epistemic value and the values of morality and prudence. So one idea would be that even though it'd be *best* to do literally everything all at once, we can perhaps make better progress than an X First strategy by doing *a couple things* at once. In any case, I won't explore this idea further here, but leave it to future work. In what remains, I'll instead highlight a particular evaluative problem for AI developers that's highlighted by the move I'm encouraging away from Morality First.

The problem is easiest to see by again focusing on the facts about how different domains of value interact. To that end, consider three examples of interaction between different kinds of value. First, it's sometimes funny (aesthetic) or politically useful (prudence) to mock people. But this is also often cruel (moral). Second, it's sometimes (prudentially) good to free-ride. But it may also be (morally) unfair. Third and finally, it's always (epistemically) good to believe what's true, usually (prudentially) useful to believe what's true, but sometimes true belief can make things (prudentially *or* morally) much worse.

Suppose you've (somehow) managed to make an e-AI that was capable of representing all these evaluative facts—the cruelty of mockery, but also its political utility, the prudential value of freeriding, but also its unfairness. The present view that I've been encouraging is that we can only do this by developing a system that can represent all the different values all at once: we won't achieve a system that can represent the moral disvalue of an instance of mockery unless it simultaneously represents how funny it is (or isn't). But notice that, given what we've said about the way these various kinds of value regularly interact, even being able to represent all the different kinds of value at stake in some situation isn't enough to deliver a verdict concerning what the system should, in the end, *do*. That's because whether (say) it turns out to be all-things-considered acceptable to mock someone is going to depend, in part, on how funny it is—or, to put things less tendentiously, it depends on one's *view* of how the value of humor trades off against moral value.

But the facts about how different domains of value trade-off against one another—by strengthening, attenuating, silencing, etc. —simply aren't settled by even a complete, accurate representation of the values at stake in some situation. I can know that the joke is funny, know that it's hurtful, and know that it's politically useful *and still not know how to respond to the joke*. If you're not convinced, notice that people do not usually disagree over the *usefulness* of (undetectably) freeriding; what they disagree over is *whether to do it*.

The point, then, is that a rejection of Morality First surfaces a problem for AI developers that was antecedently suppressed by an exclusive focus on the domain of morality. Given Morality First, we can easily stumble into accepting some version of either *Morality Only,* whereby other evaluative domains don't even figure in our all-things-considered judgments or *Morality Most*, whereby other evaluative domains only play a minor role in those judgments. But once we reject Morality First and adopt a framework whereby AI systems are required to represent the full range of evaluative facts, these too-simple views are no longer attractive.

What rejecting Morality First highlights, then, is that responsible e-AI developers must do one of two things. Either they must decide on behalf of users how to weigh off different evaluative facts against one another—in our example, the value of humor against the value of personal harm—or they must surface these choices to users.

This fact, that an accurate representation of how things morally stand in the world doesn't uniquely determine a correct way of responding, is obscured by the Morality First strategy. Suppose you go in for a Morality First strategy, and you make an e-AI that can accurately represent moral value. Then, there appears to be no special question remaining of how the e-AI ought *respond* to these representations: the e-AI should simply respond in whatever way is required by the moral values. But as we can now see, this is a mistake on two fronts. First, it's likely impossible to develop an accurate picture of what the moral facts actually are without also at the same time developing an accurate picture of what the other evaluative facts comprise—this is in part what motivates rejecting Morality First. Second, even with an accurate picture of the moral facts, the presence of an accurate picture of *all the other evaluative facts* means that we will now face a practical problem, viz., the problem of deciding how to weigh off (attenuate, strengthen, silence) the different values against one another. And importantly, you can't solve that problem simply by appealing to a particular domain of value.

Instead, as I've indicated, e-AI developers must themselves make choices about these questions. One choice would be to arbitrarily privilege one domain of value—for instance, morality—over others. If e-AI developers do this, it seems obvious that they should also clearly, explicitly, and transparently explain how exactly they do this and what the reason is for so doing.[32] Another possible choice would be to surface these issues to users—for instance, by allowing users to select broad preferences among systems with different preferences over the values of morality, prudence, humor, truth, and so on.

Right now, e-AI developers are doing neither of these things. And again, because they're all pursuing some version of X First (mostly Morality First), this might seem sustainable. Abandoning X First means giving up the hope of remaining evaluatively neutral in developing e-AI. We should insist that e-AI developers make substantive, controversial choices about value, or that they surface these choices in a clear way to end-users.

[32] The e-AI developers that come closest to doing this are Anthropic; but even they do not say exactly how they think about the trade-off between (say) truth and morality. See Bai et al. (2022a).

Declarations

Conflict of interest The author has no competing interests to declare.

References

Aristotle (2014) Aristotle: Nicomachean ethics. Cambridge University Press, Cambridge

Babic B (2019) A theory of epistemic risk. Philos Sci 86(3):522–550. https://doi.org/10.1086/703552

Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Drain D et al (2022a) Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv https://doi.org/10.48550/arXiv.2204.05862

Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, Chen A et al (2022b) Constitutional AI: harmlessness from AI feedback. ArXiv.Org. https://arxiv.org/abs/2212.08073v1

Baker DC (2018) Skepticism about ought simpliciter. Oxford studies in metaethics, vol 13. Oxford University Press, Oxford

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 610–623

Berker S (2013a) Epistemic teleology and the separateness of propositions. Philos Rev 122(3):337–393. https://doi.org/10.1215/00318108-2087645

Berker S (2013b) The rejection of epistemic consequentialism. Philos Issues 23(1):363–387. https://doi.org/10.1111/phis.12019

Birhane A (2022) The unseen black faces of AI algorithms. Nature 610(7932):451–452. https://doi.org/10.1038/d41586-022-03050-7

Bostrom N (2014) Superintelligence: path, dangers, and strategies. Oxford University Press, Oxford

Broad CD (1949) Egoism as a theory of human motives. Hibbert J 48:105–114

Buchanan R, Schiller HI (2022) Pragmatic particularism. Philos Phenomenol Res 105(1):62–78. https://doi.org/10.1111/phpr.12801

Canavotto I, Horty J (2022) Piecemeal knowledge acquisition for computational normative reasoning. In: Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society, ACM, Oxford. pp 171–80. https://doi.org/10.1145/3514094.3534182

Cervantes J-A et al (2020) Artificial moral agents: a survey of the current status. Sci Eng Ethics 26:501–532

Christiano P, Leike J, Brown TB, Martic M, Legg S, Amodei D (2017) Deep reinforcement learning from human preferences. arXiv.org https://arxiv.org/abs/1706.03741v4

Dancy J (2004) Ethics without principles. Oxford University Press, Oxford

Dancy J (2018) Practical shape: a theory of practical reasoning. Oxford University Press, Oxford

Dennett D (2010) Two black boxes: a fable. Act Nerv Super 52(2):81–84. https://doi.org/10.1007/BF03379570

Emelin D, Le Bras R, Hwang JD, Forbes M, Choi Y (2021) Moral stories: situated reasoning about norms, intents, actions, and their consequences. In: Proceedings of the 2021 conference on empirical methods in natural language processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. pp 698–718 https://doi.org/10.18653/v1/2021.emnlp-main.54

Evans O, Cotton-Barratt O, Finnveden L, Bales A, Balwit A, Wills P, Righetti L, Saunders W (2021) Truthful AI: developing and governing AI that does not lie. arXiv https://doi.org/10.48550/arXiv.2110.06674

Forbes M, Hwang JD, Shwartz V, Sap M, Choi Y (2020) Social chemistry 101: learning to reason about social and moral norms. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Online. pp 653–670. https://doi.org/10.18653/v1/2020.emnlp-main.48

Gebru T (2020) Race and gender. The Oxford handbook of ethics of AI. pp 251–269

Greaves H, MacAskill W, Thornley E (n.d.) The moral case for long-term thinking. In: Cargill N, John TM (eds) The long view: essays on policy, philanthropy, and the long-term future. FIRST, London. pp 19–28

Haugeland J (1985) Artificial Intelligence: the very idea. MIT Press, Cambridge

Hendrycks D, Burns C, Basart S, Critch A, Li J, Song D, Steinhardt J (2021) Aligning AI with shared human values. arXiv https://doi.org/10.48550/arXiv.2008.02275

Hendrycks D, Mazeika M, Zou A, Patel S, Zhu C, Navarro J, Song D, Li B, Steinhardt J (2022) What would jiminy cricket do? Towards agents that behave morally. arXiv http://arxiv.org/abs/2110.13136

Jiang L, Hwang JD, Bhagavatula C, Le Bras R, Liang J, Dodge J, Sakaguchi K et al (2021) Delphi: towards machine ethics and norms. arXiv https://doi.org/10.48550/arXiv.2110.07574

Jiang L, Hwang JD, Bhagavatula C, Le Bras R, Liang J, Dodge J, Sakaguchi K et al (2022) Can machines learn morality? The Delphi experiment. arXiv https://doi.org/10.48550/arXiv.2110.07574

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K et al (2021) Highly accurate protein structure prediction with alphafold. Nature 596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2

Kernohan A (2022) How to modify the strength of a reason. Philos Stud 179(4):1205–1220. https://doi.org/10.1007/s11098-021-01703-3

King ZJ, Babic B (2020) Moral obligation and epistemic risk. Oxf Stud Norm Ethics 10:81–105

Knowles J, Raleigh T (2019) Acquaintance: new essays. Oxford University Press, Oxford

Lord E (2018) How to learn about aesthetics and morality through acquaintance and deference. In: Shafer-Landau R (ed) Oxford studies in metaethics, vol 13. Oxford University Press, Oxford, pp 71–97

Moore GE (1903) Principia Ethica. Dover Publications, Mineola (**Edited by Thomas Baldwin**)

OpenAI (2018) OpenAI charter. https://openai.com/charter. Accessed 2 Jan 2024

OpenAI (2023) GPT-4 technical report. arXiv https://doi.org/10.48550/arXiv.2303.08774

Sagdahl MS (2014) The argument from nominal-notable comparisons, ‘ought all things considered’, and normative pluralism. J Ethics 18(4):405–425. https://doi.org/10.1007/s10892-014-9179-9

Sharadin N (2017) In defense of comic pluralism. Ethic Theory Moral Pract 20(2):375–392. https://doi.org/10.1007/s10677-017-9784-3

Shen MW (2022) Trust in AI: interpretability is not necessary or sufficient, while black-box interaction is necessary and sufficient. In: FAccT’22: ACM Conference on fairness, accountability, and transparency, June 21–24, 2022, Seoul, South Korea. ACM, New York, NY, USA

Streumer B (2017) Unbelievable errors: an error theory about all normative judgments. Oxford University Press, Oxford

Tolmeijer S et al (2020) Implementations in machine ethics: a survey. ACM Comput Surv (CSUR) 53(6):1–38

Vigani D (2019) Virtuous construal: in defense of silencing. J Am Philos Assoc 5(2):229–245. https://doi.org/10.1017/apa.2018.52

von Eschenbach WJ (2021) Transparency and the black box problem: Why we do not trust AI. Philos Technol 34(4):1607–1622

Wallach W, Allen C (2008) Moral machines: teaching robots right from wrong. Oxford University Press, Oxford

Wallach W, Allen C, Smit I (2020) Machine morality: bottom-up and top-down approaches for modelling human moral faculties. Machine ethics and robot ethics. Routledge, London, pp 249–266

Wolf S (2010) Meaning in life and why it matters. Princeton University Press, Princeton

Zednik C (2019) Solving the black box problem: a normative framework for explainable artificial intelligence. Philos Technol 34(2):265–288. https://doi.org/10.1007/s13347-019-00382-7