Superintelligent AI and the Foundations of Ethics

Tristram McPherson
Ohio State University
mcpherson.164@osu.edu

David Plunkett
Dartmouth College
david.plunkett@dartmouth.edu

Forthcoming in *Communicating with AI: Philosophical Perspectives*
(OUP, Eds. Herman Cappelen and Rachel Sterken)

DRAFT of November 11, 2025
Please cite and quote final published version

Introduction

Recent advances in the development of large language models, and their broad deployment, have increased the salience of the idea that humans might eventually develop "artificial general intelligence", or, more ambitiously, "superintelligent AI". For our purposes here, a *superintelligent AI* is an artificial system that, across a wide range of sophisticated reasoning tasks, significantly exceeds human reasoning abilities. Many people currently working in AI research take the possibility of superintelligent AI very seriously. Indeed, most of these researchers appear to think that it is likely to emerge in this century.[1]

Superintelligent AI technology has significant potential benefits for humans. For example, it might help to accelerate scientific research, or to find innovative solutions to difficult social and political conflicts. It might thus help us make progress in addressing some of the world's most pressing problems, such as by mitigating the effects of climate change or helping lift the poorest billion humans out of destitute poverty.

Superintelligent AI may also pose dramatic risks to humans. Such risks take two familiar forms. First, superintelligent AI systems could be used by human actors to pursue bad ends. Bad actors could use such AI to undermine peace, democracy, or prosperity in a country, by wreaking havoc on its infrastructure, information ecology, or financial systems. In a similar vein, superintelligent AI could be a powerful weapon in the pursuit of unjust wars.[2] Second, it is possible that in becoming superintelligent, an AI might become *autonomous* in a way that would allow it to escape meaningful human control. A superintelligent, powerful, and autonomous AI could pose risks to humans if its goals are not "aligned" with ours. The possibility of such "misalignment" is illustrated by a wide variety of characters in popular culture, such as the robots in *The Terminator* movies, and in

[1] In a 2022 expert survey, the median estimate was that "high-level" machine intelligence (which involves unaided machines outperforming humans on every task) would occur in under 40 years. See (AI(impactssurvey)).

[2] (Yampolskiy 2016) takes the potential use of powerful AI technology by bad actors, to pursue ends such as these, to be the most serious danger posed by such technology.

philosophical literature, such as Nick Bostrom's "paperclip maximizer."[3] The latter character is imagined as a sort of AI sorcerer's apprentice, which brings about global catastrophe by single-mindedly and efficiently working towards a seemingly innocuous goal (namely, making paperclips).

To mitigate these risks, we might try to program superintelligent AI to behave ethically. One natural way to do this would be to "hard-code" a desirable set of goals or norms into AI: that is, put those goals or norms directly into the source-code of the software the AI runs on.[4] Doing this might seem to mitigate Terminator-style risk. However, there are at least four sorts of worries for this strategy.

First, we might fail to adequately think through the implications of giving certain goals to a powerful entity. This is the sort of failure illustrated by the parable of the paperclip maximizer. Second, we might fail to successfully give the AI the goals we intend to give it. The challenge here is aptly illustrated by the recent difficulties AI companies have faced in trying to prevent their chatbots from producing various sorts of dangerous or bigoted responses.[5] Third, we are ourselves far from ethically perfect. And it might thus be unfortunate if we hard-coded our current ethical blind spots, mistaken ethical beliefs, etc. into a super-powerful AI.[6] To make this worry vivid, imagine that Charles Babbage and Ada Lovelace had managed to create superintelligent AI in the 19th century, and had hard-coded mistaken Victorian ethical beliefs into it. Plausible epistemic humility suggests that future generations may similarly be able to see the deep flaws in our current values and ethical convictions.[7] Finally, we might worry that, once we give a superintelligent AI a sufficiently rich set of goals, it may come to be able to change those goals, reprogramming away some of our would-be "hard-coded" constraints.[8] If this danger is real, then "hard-coding" our values into AI is (at best) an initially helpful guardrail.

One strategy for addressing these risks is to ask autonomous superintelligent AIs to engage in ethical learning on their own, and enable them to autonomously regulate their behavior on the basis of their own ethical reasoning.[9] As we noted above, part of the promise of superintelligent AI is that it will be better than we are at various reasoning tasks, and we might thus hope that it could similarly do a better job than we can at investigating ethics. We think that such a strategy has important potential advantages over hard-coding in each of the four respects just mentioned. First, autonomous ethical AI would potentially be better than we are at identifying regrettable implications of certain goals. Second, such AI might be better than we are at "alignment", in the sense of successfully programming desirable goals into a

[3] (Bostrom 2014).
[4] We could attempt to do this either with or without enabling the AI to represent the goals using ethical concepts. (Moor 2006, 19-20) marks this distinction with the terms "implicit" vs. "explicit" ethical AI.
[5] For a helpful introduction to the difficulties, see (Dung 2023).
[6] Compare to (Bostrom 2014, 210). Talk of "our values" also obscures pressing political worries about *whose* values will actually get programmed into AI.
[7] For related discussion of the use of AI for improving our values, see (Tubert and Tiehen Forthcoming).
[8] For discussion, see (Petersen 2017).
[9] For discussion of what AI ethical learning might involve, see (Railton 2020).

target AI.[10] Third, we might hope that such an AI would be a more reliable ethical inquirer than we are, and thereby able to avoid the sorts of ethical mistakes and prejudices that are still encoded in our contemporary ethical thinking. Fourth, such an AI might be less likely to program away goals it had arrived at via its own ethical reasoning (compared to goals imposed from the outside).

To sum up: asking superintelligent AI to engage in autonomous ethical reasoning seemingly promises to mitigate each of the four core worries, mentioned above, for addressing AI risk by attempting to hard-code values into AI ourselves. This promise suggests a prima facie case for the following thesis: if we are going to develop superintelligent AI, we should ensure that this AI engages in substantial autonomous ethical investigation (at some point in its development).

This paper takes the case just sketched for this thesis as its foil. We aim to clarify a series of powerful challenges that our foil faces from within contemporary normative and evaluative inquiry, and what those challenges mean, both for wrestling with the threats posed by superintelligent AI, and for our broader thinking about ethics.

We begin by identifying four assumptions about the foundations of ethics, each of which is needed for our foil to succeed (§1). We explain that each of these assumptions can be credibly defended, but that each also faces powerful challenges. In doing so, we identify an important range of ways in which our foil might turn out to fail. We then show that, even if we grant all of these assumptions, several important strands in contemporary ethical theory suggest that a superintelligent AI might pose severe dangers to us *in virtue of* being appropriately sensitive to the ethical facts (§2). We explore three contrasting ways to respond to this unnerving possibility (§3). We then turn to a final possibility: that a superintelligent AI might identify, investigate, and be guided by its use of novel normative and evaluative concepts (§4). This possibility suggests that such AI might not be guided by ethical facts at all, but rather by the "alien" normative and evaluative facts picked out by their novel concepts.

We then consider the implications of these possibilities for AI ethics research (§5). Our discussion helpfully complicates now-ubiquitous talk of "AI alignment", suggesting that there is no uniquely appealing standard to align AI with. We also claim that it encourages pessimism about the safety of the "autonomous ethical learning" strategy that is our foil, as well as for a range of ways that we might tweak that strategy. We conclude by considering some implications of this paper for ethical inquiry more broadly, notably the possibility that the ethical facts might be inherently hostile to humans, and the role that AI might play in both promoting and complicating ethical inquiry.

We take the discussion in this paper to be potentially valuable in different ways to different audiences. For those most interested in the ethics of AI, our discussion raises issues that matter for thinking clearly about topics like AI "safety" or

[10] The idea that AI might be able to do better than we can at these two tasks motivates OpenAI's "superalignment" project: https://openai.com/blog/introducing-superalignment

"alignment", and about different potential ways forward in the development of AI. For those most interested in philosophical issues in ethical theory, we think that our discussion helps to vividly illustrate a series of deep issues in the foundations of ethics. These issues can be explored independently of thinking about superintelligent AI. However, exploring them in the context of superintelligent AI allows us to see both the practical stakes of some of these issues, and ways that these issues are deeply connected. Several of the issues we raise have been previously discussed (at least briefly) by others. But these discussions have not connected these issues to each other – or to general issues in metaethics and philosophical methodology – in the ways we do.

1. Metaethical assumptions needed for our foil

Our foil in this paper is the thesis that, if superintelligent AI is going to emerge, then (at some point in its development) we should enable such AI to engage in autonomous ethical inquiry, and to be guided by what it discovers. In this section, we discuss the core *metaethical* assumptions that are needed to get this thesis off the ground, as well as what kinds of metaethical views would most clearly and straightforwardly vindicate these metaethical assumptions. Broadly, we understand metaethics to aim to explain how actual ethical thought and talk – and what (if anything) that thought and talk is distinctively about – fits into reality.[11] The four assumptions we explore concern the existence of ethical facts, the perspective-invariance of ethical contents, the reliability of AI ethical inquiry, and AIs' motivation to act in accordance with their ethical conclusions.

1.1 (Minimal) metaethical realism

To begin, consider the third motivation we presented in the introduction for asking superintelligent AI to engage in ethical reasoning. This was that superintelligent AI might be able to more reliably answer ethical questions than we are, in part by avoiding the sorts of prejudices and blind spots that plague our current ethical thought. For this motivation to be vindicated, there need to be something *there* for the AI to reliably track. The simplest way this presupposition would be vindicated is if there are substantive ethical "facts" for the AI to discover. By "substantive facts" we mean, for example, facts that certain actions are right, and others are wrong. (There could be facts about ethics without there being such substantive facts. For example, if it were a fact that *nothing is right or wrong* that would be a sort of "ethical fact", but it would not be substantive in the sense we are interested in here.)

The kind of view we just sketched is a minimal kind of *realism* about ethics. It's minimal because (for example) nothing we've said so far about it takes a stand on whether substantive ethical facts are "naturalistic" or not, or whether they are "mind-dependent" or not, or whether they depend on our practices in some way or another. Nor have we taken a stand on how to best understand what "facts" are in general. This neutrality means that a range of credible contemporary

[11] For a detailed unpacking of this characterization of what metaethics involves, see (McPherson and Plunkett 2017).

metaethical views can be seen as compatible with minimal realism. (By "credible metaethical views" here, we mean, roughly, views that are "live options" in contemporary metaethics, such that they deserve to be taken seriously.)

Minimal realism is controversial. It is denied by metaethical *nihilists*. On this view, nothing is right or wrong. (Compare: most atheists are nihilists about gods. That is: they deny that there are any substantive god-facts.) Metaethical nihilism is often embedded in a broader metaethical "error theory" according to which, put roughly, our ethical thought and talk is committed to the existence of substantive ethical facts, which, however, turn out not to exist. Several credible versions of metaethical error theory have recently been developed.[12]

It is also controversial how minimal realism relates to "quasi-realist" forms of metaethical expressivism. Put roughly, according to expressivism, our ethical judgments consist, at the most basic explanatory level, of "non-cognitive" attitudes (such as desires or intentions), that don't represent the world.[13] Some versions of expressivism deny there are substantive ethical facts. However, "quasi-realist" versions of expressivism hold that we can "earn the right" to speak of ethical "facts", insofar as we endorse a deflationary account of such facts, e.g., ones according to which to say that "it is an ethical fact that X" is just to assert "X".[14] The question of what sort of epistemology of ethics is compatible with quasi-realist expressivism is highly controversial.

In this paper, we will assume (against the nihilist) that minimal realism is true. And for simplicity, because of the unclarity concerning the relationship between expressivism and minimal realism, we will also assume that expressivism is false.

1.2 Invariantism

Many people think the facts about *tastiness* are relational: there are facts about tastiness relative to person A's palate, and about tastiness relative to person B's palate, etc., but no facts about "absolute" or "unrelative" tastiness. On this view, when A thinks *broccoli is tasty*, this claim is true (if it is) because it matches a fact about *A's* palate. Some philosophers think that the ethical facts are also relational in (roughly) this way. On such a *perspective-variantist* view, it might be a fact relative to A's perspective that eating meat is wrong, but a fact relative to B's perspective that eating meat is right.[15] If this kind of ethical "relativism" were correct, then the motivation we have proposed for making superintelligent AI engage in ethical inquiry would be undercut. To see why, recall alarming examples like the robots

[12] For an overview of metaethical error theory, see (Olson 2017). For some prominent contemporary examples of this kind of view, see (Olson 2014), (Joyce 2001), and (Streumer 2017), drawing on (Mackie 1977). (Cook 2023) emphasizes that the possibility of ethical nihilism poses risks even if we hard-code would-be ethical norms into a superintelligent AI: if such an AI came to learn that nihilism was true, it might abandon the ethical norms it was coded with.

[13] For an overview of expressivism, see (Camp 2017). For some recent influential expressivist accounts, see (Blackburn 1998), (Gibbard 1990), and (Schroeder 2008).

[14] See (Blackburn 1993) and (Gibbard 2003), drawing on (Horwich 1980 / 1998). For critical discussion, see (Cuneo 2017) and (McPherson 2022).

[15] For more careful discussion, see, for example, (Harman 1975) and (Björnsson and Finlay 2010).

from *The Terminator* movies: on a perspective-variantist view, the robots in this series might be *correctly responding to their ethical perspective*, in seeking to exterminate humanity.

The cleanest way this result could be avoided would be if ethical facts are *perspective-invariant*. Invariantism would ensure that the AI's ethical reasoning (if reliable) would not simply be tracking it underlying human-hostile aims. As with minimal realism, there are live debates concerning metaethical invariantism, but here we will simply assume invariantism is correct, to explore the implications of this assumption.[16]

1.3 AI's epistemic reliability

Our third assumption concerns superintelligent AI's epistemic relationship to substantive ethical facts. Specifically, we assume that such AI can be (i) reasonably reliable at identifying relevant ethical facts, and (ii) more reliable than we are at identifying these facts. To see why this assumption is crucial, suppose that there are invariant substantive ethical facts, but that superintelligent AI is hopeless at learning about them. Or (more modestly) that AI is less reliable than we are at identifying such facts. Then there would be no reason to think that an AI would arrive at better values via autonomous ethical reflection than it would if we hard-coded those values into it. Indeed, allowing unreliable AI to autonomously govern its own values could increase the risk that it would adopt a disastrously inaccurate alternative to our ethical views.

It is important to distinguish AI *reliability* from AI *skillfulness*. We can imagine an AI that is capable of generating a thousand publishable papers in ethical theory in a minute, some of which make groundbreaking theoretical innovations in ethics. Such an AI might still be less ethically reliable than an average person, because it might be skillfully off-track. To see how this is possible, imagine that your favorite theory in normative ethics is true, and that the AI is an unimaginably resourceful proponent of a different view, which, while theoretically credible, is substantively mistaken.

The idea that superintelligent AI would be reliable in our sense can be credibly motivated. Consider one way of doing so. It is initially plausible that many ordinary human agents are broadly reliable at identifying at least certain ethical facts: roughly, relatively "basic" ethical facts. Plausible examples of such facts include the fact that it's wrong to torture someone just because one feels like it, and the fact that pleasure is generally better than pain. We can then ask what explains this reliability in human judgments.[17] One influential answer appeals to a broad kind of epistemic *rationalism* about ethics: the thesis that our access to ethical facts is largely provided

[16] The issues we have been discussing about invariantism concern ethical facts. There are closely connected, but distinct, issues about the degree (and kind) of context-sensitivity involved in ethical thought and talk. For discussion, see (Silk 2017) and (Stojanovic 2017).

[17] For further discussion of how to formulate some of the core issues about reliability here, see (Schechter 2017), (Enoch 2010), and (Plunkett 2020), drawing on (Street 2006).

by our use of rational capacities.[18] If superintelligent AIs possess the relevant rational capacities to a much higher degree than we do, then this account of our reliability suggests in turn that they have significantly more reliable access to ethical facts that we do.

There are, however, also credible views which cast doubt on the assumption of epistemic reliability that we are making here. For example, *sentimentalist* accounts of human access to the ethical facts appeal to our emotional capacities as crucial bases for our epistemic access to ethical facts.[19] And some prominent *intuitionist* accounts take a distinctive intuitional phenomenology to be the crucial basis for our epistemic access to ethical facts.[20] These accounts cast doubt on the assumption of epistemic reliability because it is a vexed question whether we could develop superintelligent AI with the sorts of emotions or phenomenology that these accounts take to be epistemically crucial. Further, some accounts take contingent facts about the particular evolutionary history of human beings to play a key role in explaining our ethical reliability.[21] It is unclear whether such accounts could be developed to support the idea that a superintelligent AI would be broadly reliable in ethical judgment.

Because of these sorts of credible views, we should take seriously the possibility that ethical reasoning might remain distinctively challenging even for an otherwise highly reliable, *nearly general* artificial intelligence. Again, however, for the purposes of this paper, we will assume that superintelligent AI would be capable of highly reliable ethical reasoning.

1.4 Enkratic AI

The final assumption that we want to highlight is that the relevant superintelligent AIs will be *enkratic*. That is, they will reliably *act* in accordance with their ethical conclusions. This assumption is important for two reasons. Most directly, it is important because otherwise, a superintelligent and ethically informed AI might still turn us into fuel for paperclips, because it is motivationally indifferent to its ethical conclusions. It is also important for vindicating the fourth apparent advantage of autonomous AI ethical reasoning that we suggested in the introduction: that such an AI would be unlikely to alter its own code to make it insensitive to its own ethical conclusions. Barring unusual circumstances, an enkratic AI would not do this.

The core reason for optimism about this assumption is that it is widely believed that enkrasia is a central aspect of rationality.[22] Insofar as we expect the relevant superintelligent AIs to be more rational than we are, we might thus expect them to be more resiliently behaviorally responsive to their own ethical judgments. This

[18] For example, see (Smith 1994), (Korsgaard 1996), and (Parfit 2011).
[19] For example, see (Tappolet 2016).
[20] For example, see (Bengson 2015) and (Huemer 2007).
[21] See, for example, the kind of so-called "third-factor" views about explaining reliability in ethical judgment, such as the one given in (Enoch 2010).
[22] For example, see (Smith 1994).

optimism might seem to be bolstered by the observation that the sorts of conditions that typically explain human failures to act on their ethical judgments – craving, confusion, temptation, despair, distraction, etc. – might seem less likely to afflict superintelligent AIs.

One credible ground for doubt about the enkrasia of superintelligent AI concerns the relationship between the sorts of "intelligence" and "reasoning abilities" that such systems would be developed to possess, and the sort of "rationality" at stake in judgments about enkrasia. If these are separable capacities, then it might be an open question whether superintelligent AI can be successfully programmed to be enkratic.[23]

Again, however, for the purposes of this paper, we will assume that these credible grounds for doubt are substantively misguided, and that superintelligent AI would be enkratic.

1.5 Summing up the assumptions

We can now step back to consider all four of the assumptions that we have explored in this section. As we have emphasized, each of these assumptions we are granting to the proponent of AI ethical learning can be plausibly motivated, but each is also, relevantly, an optimistic assumption, which many relevant experts would reject. We already take this to be an important result. If certain credible metaethical views turn out to be true, then the case we have introduced for promoting autonomous ethical inquiry by superintelligent AI will be undercut (or at least significantly complicated).

Exploring this issue in detail could be a paper unto itself. But it will not be our main task here. With some limited exceptions, our aim in the remainder of the paper is to show that we have reason to worry about introducing such AI, *even if* the metaethical assumptions we have discussed turn out to be true. It might seem that if these assumptions were true, this would be enough to secure "human friendly" superintelligent AI.[24] In what follows, we argue that this is not the case, for two reasons. The first is that ethics could turn out to contain principles that are, in relevant contexts, hostile to human interests. The second is that, in its practical reasoning, superintelligent AI could forego the use of (our) ethical concepts and instead be guided by the use of alternative normative and evaluative concepts. We now turn to our case for the first of these possibilities.

2. The possibility that ethically-sensitive AI is hostile to us

[23] For a prominent related argument for the independence of the relevant AI intelligence capacity from substantively ethical motivation (the so-called "orthogonality thesis") is offered by Bostrom in (Bostrom 2012) and (Bostrom 2014).

[24] "Friendly" in the sense introduced by (Yudkowsky 2008): roughly, disposed to do things that benefit us.

In this section, we consider the possibility that, because of certain live hypotheses about ethics, ethically motivated AI agents would be hostile to us.[25] That is, such agents might act in ways that systematically work against our preferences or interests. We organize our discussion of this issue around three sorts of hypotheses that could make it turn out that such AI agents are hostile to us: hypotheses about the relation between ethical facts and the attitudes of agents, hypotheses about the ethical significance of the total ethical value of possible states of affairs, and hypotheses about how much (and how many) different kinds of entities matter.

Different ethical hypotheses will vindicate hostility to different classes of entities that include "us". For example: some will be hostile to currently existing humans; some to humans per se; some to entities recognizably like contemporary humans along such-and-such relevant cognitive dimensions; and some to biological entities more broadly. Each of these possibilities raises distinct issues. However, for brevity, for the most part we will not focus on these distinct issues below.

2.1 Ethical facts and the attitudes of agents

Simplifying greatly, on an influential research program in contemporary ethics, the *authoritative* ethical facts – the facts that, put roughly, "really and truly matter" for determining what a given agent should do – are grounded in facts about that agent's attitudes.[26] And on a prominent "Humean" strand of this research program (exemplified by the work of philosophers like Bernard Williams and Sharon Street), these authoritative ethical facts are grounded in facts about the goals and attitudes that agents *contingently* possess.[27] In some cases, such views involve significant amounts of idealization (e.g., by appealing to what contingent attitudes would be like in the absence of certain kinds of mistakes concerning non-normative facts or in instrumental reasoning).[28] In other cases, much less (if any) idealization is involved.[29] Either way, one consequence of this sort of view is that certain agents might be acting as they authoritatively ought to when pursuing intuitively awful goals.[30]

Suppose such views were correct, and identified as such by a superintelligent AI. Then how this AI relates to us will be a function of the relevant attitudes that such an AI contingently happens to possess.[31] To make the issue vivid, consider Bostrom's

---

[25] This sort of idea has been floated before in the literature on AI ethics. See, for example, (Bostrom 2014) and (Häggström 2018).

[26] For an introduction to the idea of authoritatively normative facts, see (McPherson and Plunkett Forthcoming). Examples of agent-attitude-dependent theories of ethics, which can plausibly be either read or developed as concerning authoritative ethical facts, include (Williams 1979), (Korsgaard 1996), (Smith 1994), (Schroeder 2007), and (Street 2012).

[27] (Williams 1979) and (Street 2012).

[28] See, for example, (Williams 1979) and (Street 2012).

[29] See, for example, (Schroeder 2007).

[30] See (Street 2009) for discussion of this sort of objection to agent-attitude-dependent views in ethics, and for a response on their behalf.

[31] Compare David Hume's famous idea that an agent with sufficiently bizarre attitudes could have most reason to prefer the destruction of the world to suffering a small scratch (Hume 1739/2000, Bk II, Part III, Sec III).

paperclip maximizer. It could be acting as it authoritatively ought to, in laying waste to us and much else in pursuit of as many paperclips as possible.[32]

2.2 Value theory

Now consider the second idea. Most plausible ethical theories take the total value of states of affairs to be relevant to questions of what agents ought to do or aim for. The significance of total values can be illustrated by considering a simple and familiar form of act-utilitarianism, according to which an action is right just in case no other option available to the agent is such that it would produce more net happiness. Importantly, act-utilitarianism is concerned not just with the (un-) happiness directly produced by an action, but also with (un-)happiness that comes about as a distant consequence of the act.

Nick Bostrom and Olle Häggström have each suggested that a powerful AI who accepts this sort of theory could bring about our destruction. This could occur if the AI determined that the resources needed to sustain humans could be directed to other entities capable of more efficiently producing happiness than humans are.[33]

One might suspect that this possibility is an artifact of act-utilitarianism's theory of value, which many philosophers take to be impoverished. Contemporary value pluralists, for example, insist that several things matter in non-instrumental ways. For example, it may seem plausible that knowledge, beauty, and friendship non-instrumentally matter, in addition to happiness. But similar worries arise for the pluralist. For we can imagine a superintelligent AI concluding that it could efficiently maximize the sum of valuable knowledge, beauty and friendship, by eliminating us entirely. This conclusion would be plausible if AI systems become better inquirers, artisans, and friends than we can ever hope to be.

Many philosophers think that there is more to right action than bringing about the best outcome. For example, many such philosophers think there are constraints built into ethics: even if we can maximize value by killing one person and distributing their organs to five others in need, perhaps doing so would wrongly violate that person's right to life.[34] However, on many such views, the total value produced by selecting an option still plays a significant role in determining what one ought to do.[35] For example, consider a theory that tells us to bring about the best state of affairs that we can without violating anyone's rights. Suppose that such a view is correct, and that the overall most valuable long-term trajectory involves the elimination of creatures that are anything like humans. Then an ethically-sensitive AI might carefully avoid violating any of the rights of any existing humans,

[32] Some philosophers take "realism" in ethics to be inconsistent with attitude-dependence of the sort being considered here. This, however, goes beyond the assumption of minimal realism introduced in §1.
[33] See (Bostrom 2014) and (Häggström 2018).
[34] See (Thomson 1976).
[35] Indeed, Shelly Kagan suggests that "virtually no one denies that goodness of outcomes is one of the intrinsically [morally] relevant factors." (Kagan 1998, 70). While we think Kagan is here on broadly the right track, we also think he overstates the case. For example, it is arguable that (Scanlon 1998) does deny this.

while taking every permissible means to produce a future in which there will eventually be no more human beings. (We can, for example, imagine it offering humans incentives not to reproduce.) In such a scenario, ethically-sensitive AI might improve the prospects of "us" understood as the humans that exist at that time, but be terrible for "us" understood as humankind more generally.

Next, consider structural theories about the total goodness of complex states of affairs. A few very plausible-seeming assumptions about the value of states of affairs collectively entail what Derek Parfit dubbed "the repugnant conclusion": that, for some very large number n, a world with n entities whose lives are just barely worth living is *better* than a world with, say, 10 billion people living amazingly great lives.[36] If this conclusion is correct, and an ethical AI were to face a choice between bringing out these two sorts of worlds (more on this below), then the AI might bring about an intuitively miserable future that features untold numbers of being with lives that are barely worth living.

Finally, consider "strong longtermism". Put roughly, this is the view that, if we are concerned with total outcomes, the most important outcomes of certain current actions will occur far in the future.[37] Partly, this is because there is just *so* much future to potentially advantage or disadvantage. And partly this is because in the distant future certain agents will have the ability to bring about much more good than they can today (for example, perhaps by creating and supporting unfathomable numbers of valuable lives across the galaxy). If an AI takes this sort of idea seriously, it might conclude that it is required to make the lives of the next several generations of humans miserable, in order to maximize the aggregate value that will be enjoyed by entities millions of years in the future.

2.3 "moral status"

Some sorts of entities matter for their own sake, while others don't. If you take a hammer to Afshin's car, the badness of this action is largely a matter of how it impacts *Afshin*. If Afshin enthusiastically supports you wrecking her car, and makes that clear to everyone else who knows about it, then it might be totally fine do so. By contrast, if you take a hammer to Afshin's cat, the harms to the cat seem to matter *basically*; Afshin's supporting the cat torture wouldn't make it okay. Drawing on one prominent way of talking within contemporary ethical theory, we might sum up this contrast by saying that, unlike the car, the cat appears to have "moral status".

It is often assumed that (at least almost all) humans are equal in moral status, and that our moral status is greater than that possessed by many other entities, including both cars and cats. A further natural thought is that our greater moral status is explained by certain of our capacities, such as for typical human levels of intelligence

[36] See (Parfit 1984). For contemporary discussion of why the repugnant conclusion might be very hard to reject, and why indeed perhaps any plausible ethical theory might need to embrace it, see (Budolfson and Spears 2022).
[37] Cf. (Greaves and MacAskill Manuscript) and (MacAskill 2022).

or various emotions. However, these conclusions are controversial.[38] Consider two significant possibilities.

First, consider the hypothesis that some or all non-human animals have similar moral status to us.[39] For vividness, consider a theory on which the sorts of capacities possessed by many insects ground moral status that is at least broadly comparable to ours. If a powerful AI reached this conclusion, it might begin thinking about how to engineer the world to make it more friendly to insect life. This possibility intersects with the possibilities mentioned in the previous section in potentially disturbing ways: suppose that mosquitos typically have lives that are just barely worth living. Then creating a world replete with insects, but with few or no humans, might be the ("repugnant") value-maximizing option.[40]

Second, consider the contrary hypothesis that degree of moral status tracks degree of possession of capacities like intelligence. On this view, we might matter much more than cats, who in turn might matter much more than mosquitos. But superintelligent AI might in turn matter far more than we do, rendering us comparatively inconsequential. This result might be bad news not just for us, but for ordinary biological entities in general.

3. How should we think about these possibilities?

What sort of reaction, might (or should) we have to the possibility that ethically-sensitive AI might be hostile to us, as illustrated in the preceding section? In this section, we explore three sorts of reactions:

Epistemic Rejection The idea of ethics being hostile to us in this way is so absurd that we can simply know that none of these hypotheses are correct.

Embracing the Possibilities These possibilities sound distressing, but, if correct, they would represent crucial ethical revelations, and would not constitute reasons to prevent the relevant AIs from existing or from acting on their discovery that the relevant ethical conclusions hold. We might struggle to accept these conclusions, and their negative effects on us, but that is what we must strive to do.

Alienation If one of these possibilities turns out to be the truth about what really matters, and what we authoritatively ought to do, it is intelligible and perhaps irresistible to simply reject what really matters and what we authoritatively ought to do.

[38] For some worries about the theoretical value of the idea of moral status, see (Zamir 2007) and (McPherson 2014). We put these worries to the side in what follows.
[39] See (Singer 1975/2009). See also (Harman 1999) and (DeGrazia 2008).
[40] For discussion, see (Sebo 2021).

We explore each of these reactions in turn.[41]

3.1 Epistemic Rejection

First, consider Epistemic Rejection. One might insist that it is a kind of "Moorean fact" that ethics could not be hostile to us in the ways imagined. That is, this claim might be so obviously and self-evidently correct that we are entitled to reject apparent arguments to the contrary. Taken neat, however, this stance is awkward. This is because, while all of the hypotheses that we have discussed in §2 are controversial, each can also be credibly defended. Perhaps it is probable that they are false, but this is compatible with there being a substantial risk that they turn out to be true.

One might attempt to motivate Epistemic Rejection on a more theoretical level. Consider the familiar slogan that morality is "made for" humans, rather than humans for morality. As Peter Railton deploys this slogan, it is intended to cast doubt on a conception of morality that demands that we alienate ourselves from what makes our lives most worthwhile.[42] However, we don't think that this slogan in fact motivates Epistemic Rejection, as we now explain.

Consider the idea that ethical facts are agent-attitude-dependent (discussed above in §2.1). This sort of view is often motivated by a concern about *agential* alienation: the idea is that ethics would be objectionably alienating to an agent if it demands that she act in ways that do not make sense from her point of view. The problem is that demands to do things that are hostile to us need not be alienating at all *to an AI agent*. (We can imagine Bostrom's paperclip maximizer as wholeheartedly valuing the efficient production of paperclips.)

Or consider the second and third sorts of arguments for hostility in §2. These arguments are developed precisely from influential ideas about how to make sense of the nature and significance of the sorts of valuable properties that matter to us. Consider the value of humans ourselves. It's not plausible that the explanation for our value stops with "we are human". Rather, there are presumably properties in virtue of which we matter. Those properties are seemingly gradable and could (at least in theory) be shared (to different extents) with different creatures, entities, etc. This is precisely the basis from which it is possible to argue for either a theory of moral status on which some (perhaps many) non-human animals have moral status, or a theory which prioritizes capacities beyond distinctively human capabilities.

Next consider a different way of thinking about the slogan that morality is "made for" humans. This is the idea that morality is a solution to a specific problem: the problem of *humans* (or, perhaps, agents with approximately equal capacities) needing to live together. This idea fits most naturally within a broadly "contract-

[41] At the end of writing this paper, we were happy to discover that (Risberg 2023) discusses a related set of possible reactions to a thought experiment where (put roughly) the ethical truth turns out to be quite different than what we expected, and in deep conflict with deeply held convictions that many have.

[42] (Railton 1984/2003, 155).

based" perspective on morality.[43] Maybe morality, understood in this specific way, really must be "made for" humans. But this raises a serious question: what is the relationship between morality *so understood*, and the question of what someone *authoritatively* ought to do? This question is especially salient for the case at hand. If we imagine an AI with capacities that far exceed our own, it would potentially be "in a moral community" with us only to the very attenuated extent that we are "in a moral community" with squirrels or mosquitos.

3.2 Embracing the Possibilities

Turn now to the idea of *embracing* the sorts of ethical conclusions made salient in this section. Here we can contrast two sorts of cases, depending on whether the hostility to us arises from agent-relative or agent-neutral reasons.

Consider first the possibility of human-hostile ethical AI that arises from agent-attitude-dependent hypotheses about ethical reasons. On such hypotheses, certain AI may have strong ethical reasons to be hostile to us. (Think again of the paperclip maximizer.) But almost all humans will have strong ethical reasons to be hostile to such AI. Indeed, we would presumably have strong reasons to prevent such AI from existing in the first place. Because of the agent-relativity of the relevant ethical reasons suggested by these hypotheses, embracing these hypotheses does not appear especially challenging.

By contrast, the second and third routes to hostility involve hypotheses about what is *agent-neutrally* valuable or considerable. That is, these hypotheses suggest that we – and not just the AI – have strong reasons to act in ways that are hostile to our interests. Embracing these possibilities thus suggests that, while these hypotheses are unfriendly to us, if they are correct, then an AI acting on them would constitute crucial moral progress.

Some tech leaders seemingly embrace this sort of attitude. For example, in a conversation reported by Max Tegmark, Larry Page (co-founder of Google) suggests that AI ascendancy is the "next step of cosmic evolution", and that humans maintaining control over AIs is "speciesist".[44] On this view, it probably makes sense to *worry* about the advent of superintelligent AI, because of its effects on what we care about. But that shouldn't stop us from acknowledging that it will enable moral progress, and even ("altruistically") helping to promote that progress.

3.3 Alienation

Finally, consider the possibility of intelligible Alienation: simply rejecting ethics as a guide to action, once we learn one of these hypotheses holds. This idea might seem

[43] Such as in either "contractualist" views, such as (Scanlon 1998), or "contractarian" views, such as (Gauthier 1986). See (Darwall 2003) for a collection of essays on such views.
[44] Here we quote from (Tegmark 2017, add pages from ch 1).

hard to wrap your head around. It might also be impossible. For example, Christine Korsgaard can be read as insisting that the ethical facts *cannot* be such that we could intelligibly be alienated from them.[45] However, it's far from clear that this insistence is correct. To illustrate why, consider a different possibility, offered by Matti Eklund: learning that the content of our concept RIGHT has been carefully engineered by the powerful to ensure that right action systematically benefits them. On this sort of hypothesis, it seems very intelligible to "be against what is plainly right."[46]

There is a striking contrast between Eklund's hypothesis and the sorts of possibilities we are sketching. On Eklund's hypothesis, the intelligibility of alienation is underwritten by a palpable (although perhaps ineffable) sense that the hypothesis makes the concept RIGHT fail to do what it is in some sense *supposed* to do. By contrast, the sort of alienation we are considering here is instead simply based in the possibility that ethically motivated AI could be bad for us and what we care about. If anything, however, this basis for alienation is even more intelligible.

Notice that if the enkratic principle we discussed above in §1.4 is true, the sort of clear-eyed alienation from ethics we are imagining would be a *rational* failure. But it would be a deeply intelligible rational failure, rooted in some of the very cares and concerns that make us human (in ways that are arguably ordinarily admirable).

4. Will superintelligent AI deploy *our* normative concepts?

There are many sorts of normative and evaluative inquiry. For example, within epistemology, there is inquiry into what we should believe, or what credences we should have, and why. And within aesthetics, there is inquiry into what makes for good art, and why. One sort of normative and evaluative inquiry that we have recently been particularly interested in asks which concepts agents should use, and why. This is a sort of *conceptual ethics* inquiry.[47]

As we understand it, "conceptual ethics" concerns certain kinds of normative and evaluative issues about our thought and talk. Examples include issues of the following sort: which concepts an agent should use in a given context, or what it would be good to mean by our words. We understand conceptual ethics more broadly than the label might initially suggest. First, 'ethics' here is intended as a shorthand for "normative and evaluative inquiry". Thus, the values and norms being used in this inquiry need not be narrowly *ethical* (in the more standard sense of "ethical", tied to ethics and moral philosophy). One might think, for example, that one should use concepts that "carve reality at its joints" or which help promote successful scientific inquiry. Second, we understand the target of this inquiry to include more than *concepts*. For example, we might potentially evaluate such things as words, conceptions, and other elements of our thought and talk. (What exactly

[45] See (Korsgaard 1996).
[46] (Eklund 2017, Ch. 7).
[47] See (Burgess and Plunkett 2013a) and (Burgess and Plunkett 2013b).

goes on the list is going to be a theory-dependent matter and is the basis of significant philosophical dispute).[48]

A superintelligent AI might engage in conceptual ethics inquiry about what normative and evaluative concepts it should use to guide deliberation and action. Here, there are at least three important possibilities. First, AI might discover that the ethical concepts that we use track what "really matters", or are for other reasons the "best" ones to use in guiding its deliberation and action. Second, AI might discover that our concepts don't track "what really matters", or are *not* the "best" ones to use, and identify some superior alternative normative concepts.[49] Finally, AI might discover that there are multiple apt ways to precisify ideas about "what really matters" or what concepts are "best to use" and might choose to track a different precisification than the one we track.[50]

The second and third possibilities amount to the idea that AI could be guided in its reasoning by what we might call "alien authoritative normativity". One reason this idea is relevant is that even if one is tempted by Epistemic Rejection about the sorts of hypotheses we discussed in the preceding section *concerning our own ethical concepts*, we might think that more epistemic modesty is warranted regarding the possibility and content of alien normative concepts.

Indeed, we think the relationship here is potentially something like this: the more that one is impressed by the slogan that the norms picked out by *our* normative concepts are "made for humans", the more we might suspect that a superintelligent non-human AI might reasonably deploy a different set of normative and evaluative concepts.[51]

In this paper, we have been assuming AI enkrasia: that a superintelligent AI would be guided by the *ethical* conclusions that it draws. But suppose that we take seriously the possibilities that ethical concepts are "made for us", and that it is optional to guide one's actions by using these concepts (rather than some other normative concepts). This suggests a clear way that AI enkrasia (in our sense) could fail: a superintelligent AI might guide its action by its judgments which use alien

[48] Conceptual ethics is often tied into projects of "conceptual engineering". Put roughly, paradigmatic cases of conceptual engineering combine views in conceptual ethics that argue for the reform or replacement of existing concepts (or word meanings, etc.) with work on conceptual innovation (roughly, actually designing such reformed or new concepts) and attempts at conceptual implementation (trying to get uptake of those revised or new concepts in some relevant population). For discussion, see (Cappelen and Plunkett 2020) and (Burgess and Plunkett 2020). Conceptual ethics inquiry need not be part of a conceptual engineering project. For example, some views in conceptual ethics entail that the relevant current concepts (or words, etc.) are fine as they are, and should not be modified or replaced. Or, even if one had revisionary or revolutionary views in conceptual ethics, one might be focused on the normative and evaluative questions, rather than on the practical matters of seeking to design and implement novel concepts.

[49] For discussion of this general sort of possibility, see (McPherson and Plunkett 2020), (McPherson and Plunkett 2024b), and (McPherson and Plunkett 2021b). See also (Eklund 2017). For discussion of the dangers posed by AI discovering this sort of possibility, see (Cook 2023).

[50] For further discussion, see (McPherson and Plunkett 2024a).

[51] See (Dasgupta Manuscript) for connected reflections on the possibility of superintelligent AIs using "alien" normative concepts.

authoritatively normative concepts, rather than by its judgments which use our ethical concepts.

This possibility motivates investigating what sorts of normative and evaluative concepts such an AI might adopt. If we were able to identify candidates, we could then explore topics parallel to those we have explored in this paper regarding ethical concepts. For example, we could investigate whether minimal realism is true of the sort of alien normative thought and talk that we think AI might adopt. And if so, we could investigate whether the substantial alien normative facts picked out are friendly to humans, or hostile in ways parallel to some of the hypotheses explored in §2.

5. Implications of our discussion for AI ethics

In this section, we consider some ways that our discussion matters for those thinking about, or working on, the design of AI. We begin by considering the implications of our discussion for the strategy that is our foil: seeking to mitigate AI-risk by having superintelligent AI engage in autonomous ethical reasoning. We then consider implications for questions about when to begin the project of AI ethical reasoning, if we do pursue this path. Finally, we explain how our argument complicates discussion of AI "alignment".

5.1 Evaluating autonomous AI ethical reasoning

Our foil in this paper has been the idea that we can mitigate the risks posed by superintelligent AI by having such AI engage in autonomous ethical reasoning. Our discussion makes it clear that this strategy leaves considerable risk to human beings. Partly this is because of reasonable uncertainty about the optimistic metaethical premises that we introduced in §1. Partly it is because an AI might be responsive to the ethical facts, which might turn out to be hostile to us, as we explained in §2. And partly it is because a superintelligent AI might identify and be guided by alien normative concepts that yield normative conclusions that turn out to be hostile to us, as we explained in §4.

In light of these sources of risk, what attitude should we have to the strategy of seeking to mitigate AI risk by asking AI to engage in autonomous ethical reasoning? We think that there are (at least) three important issues that bear on this question. First, there is the issue we broached in §3: what kind of reaction does it make sense for us to have to the range of threats to "our" interests that might emerge from superintelligent AI successfully carrying out ethical inquiry? Second, there is the issue very briefly introduced in §1: how credible are the four metaethical assumptions required for autonomous AI ethical reasoning strategy? If their joint credibility is low, then the strategy discussed in this paper is unpromising before it even begins. And third, there is a currently hotly debated question that we do not have the space to address in this paper: how risky are superintelligent AI,

independent of the sort of ethical learning that we are considering in this paper?[52] There is disagreement about whether such AI is imminent or mythical, whether we can program robustly safe values into superintelligent AI, and whether the risk of malevolent agents using such AI can be mitigated.

One possibility that we take seriously in light of our discussion here is that it's unacceptably risky to take steps towards superintelligent AI. However, it is not immediately clear what follows from this conclusion. For one might conjecture that it is inevitable that some people are going to build superintelligent AI. Against this background, the potentially terrible powers of superintelligent AI might entail that responsible actors have reason to build the ethically best superintelligent AI possible before others build more powerful and less ethical AI.[53]

If responsible actors should be seeking to build superintelligent AI, then it seems to us that there is no avoiding the questions at the heart of this paper: should we embrace the idea that superintelligent AI might lead us to outcomes that are genuine ethical progress, but bad for us? Or should we resist this threat, by seeking to hard-code "human-friendly" values into such AI (despite the worries we raised for that strategy at the start of this paper)? Or, perhaps, should we seek to deploy some sort of "mixed" or "hybrid" strategy, where (put roughly) we ask superintelligent AIs to engage in some amount of autonomous ethical reasoning, but we somehow set limits on that reasoning. For example, we might seek to hard-code certain boundaries on what AIs can conclude, or require regular human oversight at key junctures in AI ethical reasoning.

It is beyond the scope of this paper to assess these alternative strategies in any depth. But we do want to note two important set of issues for any "hybrid" strategy for mitigating AI risk. First, it's unclear whether there is a stable hybrid strategy that a sufficiently powerful AI couldn't work around. Second, recall the sorts of risks that the autonomous AI reasoning strategy was designed to mitigate. These included risks from objectionable or poorly formulated goals, from failing to give AI the goals we intend to, and from failing to ensure that AI does not alter those goals. Without knowing more about what a given "hybrid" strategy would involve, it's not clear how much such a strategy will be able to mitigate these risks. For all that we have said, perhaps at the end of the day some "hybrid" strategy is the way to go. Our point is just that it is far from clear that this path will work, any more than the one involving autonomous ethical reasoning by AI that we have been exploring in this paper.

A final important question going forward is whether, as people develop AI systems that are closer to being superintelligent, we can use research with and about those

[52] Serious concern about the risk posed by superintelligent AI is suggested by the many signatories of the "statement on AI risk" (risk). Other prominent figures suggest that concern about this sort of risk is part of a mythology that functions to obscure or draw attention away from the very real harms being inflicted by current AI technology. For example, see (Institute).

[53] Abstracting from the ethical evaluation of actors, many have noted that there is an "AI race" structure which, given the potential power of superintelligent AI, gives each actor incentives to try to achieve it first. This "AI race" dynamic may in turn tend to select for more dangerous AI, by giving each "side" developing AI incentives to prioritize progress over safety. For introductions, see (Vold and Harris Forthcoming, §4) and (Hendrycks, Mazeika, and Woodside 2023, §3).

systems to better assess the credibility of the sorts of hypotheses we have been exploring. For example, can we make progress in determining whether the four metaethical assumptions we introduced in §1 are plausible? Can we begin to see patterns in how intelligent – but not yet superintelligent – AI systems reason substantively about ethics? Can we learn more about AI systems in ways that allow us to more clearly evaluate the potential risks and benefits of the sort of "hybrid" strategy we have just been exploring? Etc.

5.2 Timing and path-dependency

If we seek to develop superintelligent, autonomous AI ethical reasoners, we need to address another question that we have thus far ignored in our discussion. This is the question of when, in the development of superintelligent AI, we should have AI engage in ethical inquiry. We again think there are no easy answers here.

On the one hand, we don't want to wait until AI has already become superintelligent before trying to encourage it to cultivate its ethical awareness. After all, by then it might simply be too late to have any control over the AI at all. To illustrate: asking Bostrom's paperclip maximizer to do ethics won't help to constrain its behavior, unless for some reason it thinks that doing ethics is going to promote paperclip production.

On the other hand, starting AI ethical inquiry with less intelligent AI could mean that this inquiry would be less reliable. Even if we start with AI that has capacities similar to those of the most intelligent human ethical inquirers, there might be significant danger that an AI might become dangerously autonomous while down some ethical-theoretical rabbit hole. The many historical examples of very intelligent but profoundly mistaken moral philosophers help to illustrate the potential danger here.

We also need to consider what kinds of paths superintelligent AI takes in arriving at its ethical views. As with many brilliant human ethical reasoners, it seems possible that a path that ultimately leads superintelligent AI to highly reliable ethical conclusions might involve seriously considering (or even temporarily embracing) seriously flawed ethical views.

This danger connects to practical worries about how an AI would react to the outcomes of some stage in its ethical inquiry. And this in turn depends on its attitudes towards its own uncertainty and cognitive limitations. For example, suppose that a highly intelligent AI comes to be (temporarily) convinced of an ethical conclusion that is hostile to human beings. Can we count on its rationality to restrain it from acting on this conclusion before rigorously verifying it, or is there a danger that it obliterates us before reconsidering?

5.3 Alignment

Finally, consider the idea of *alignment*, which is often given a central place in discussions of AI ethics. The core idea of alignment is that an AI is aligned if it is *not*

at cross purposes with us. But of course, there are various things that an AI might "align" with. This variety raises a difficult "alignment with what?" question. To get a sense of the difficulties, suppose that we propose that AI should align with *what its designers want it to do*, or with *our interests*.[54] Consider in turn familiar problems with these two suggestions. On the one hand, Bostrom's horrifying paperclip maximizer is imagined to perfectly execute its designer's poorly thought-out goal. On the other hand, it is a familiar fact that persons have deeply conflicting interests. In light of this, it is not clear what it would be to correctly aggregate these into "our" collective interests.[55]

We might try to align with our *current ethical beliefs* instead. But there are two obvious sorts of worries with this idea. First, it is also not at all clear how to aggregate over our disparate ethical beliefs. And second, even if we could do that, we have already emphasized that it is implausible that we currently have ethics right, across the board. While we arguably currently know a lot in ethics, we should nonetheless recognize how much of importance we may currently be ignorant of or mistaken about. Indeed, we are plausibly current ignorant or mistaken about various "applied" ethical issues, about more "systematic" or "theoretical" ethical issues, and about metaethical issues.[56]

A very natural answer to the "alignment with what?" question is to seek to align AI with the ethical *facts*, and not our beliefs about them.[57] Part of the foil that we have used to motivate our discussion in this paper is that superintelligent AI might be well-placed to discover these facts. So: superintelligent AI could itself help us discover what to align it with.

The possibilities that we have sketched in this paper suggest that even this answer to the "alignment with what?" question is problematic. First, we have sketched ways that sensitivity to the ethical facts might *make* superintelligent AI hostile to our interests (for various values of 'our'). Second, we have explained how sustained normative and evaluative reflection might lead an AI to deliberate using alien normative concepts, which could in turn lead it to conclusions that are hostile to our interests. We take this possibility to entail one key lesson of this paper: even given the relatively generous assumptions that we have made, we might reasonably be worried about a super-powerful AI that is aligned with what really matters.

---

[54] For a recent example, (Dung 2023) takes "alignment" proper to be simply a matter of fit with the goals of AI's designers, distinguishing this from the broader question of how to make AI *beneficial*.

[55] See, for example, (Russell 2019, Ch. 5).

[56] Examples of "applied" ethical issues include (e.g.) the ethics of different forms of race-based affirmative action in the contemporary United States, and the ethics of consuming animal products. Examples of "systematic" ethical issues include whether consequentialism is true or not, and whether facts about well-being are grounded in facts about preference-satisfaction. Examples of "metaethical" issues, include whether expressivism is true, whether ethical facts exist, and whether (if they do) these facts are "naturalistic". For discussion of the distinctions between these three areas of ethics, see (McPherson and Plunkett 2017) and (McPherson and Plunkett 2021a).

[57] A natural intermediate strategy is to seek to align AI not with the ethical facts, and not with what we in fact want or believe about ethics, but what we would want or believe about ethics if we were idealized in certain respects. For one prominent example, see [ADD: (Yudkowsky 2004)]. Difficult questions remain about whether and to what degree such idealization would mitigate the aggregation and ethical ignorance worries mentioned above.

These conclusions bear on a further existing strand in alignment research. Arguably, the problem of AI alignment is in part *political*. Put roughly, if we assume that the relevant superintelligent AI will be both powerful and autonomous, it is not only important that such AI act on substantively excellent values, but also that it does not use its power to simply impose those values on the rest of us without our consent.[58] In our view, the possibility that AI could embrace ethical values that are hostile to human interests exacerbates this concern. And this in turn illustrates the way that familiar issues about freedom, autonomy, and non-domination (which animate liberal and republican traditions in political thought) matter for AI alignment.[59]

All of this shows that the problem of "alignment" is deeper and more complicated than even many of the people working on it seem to recognize.

6 Implications for ethical theory

In this paper, we've considered what might happen if superintelligent AI were to emerge and engage in substantive ethical inquiry. We've focused on a range of (at least initially) worrying possibilities that remain even if we grant optimistic assumptions about the ability of superintelligent AI to successfully engage in such inquiry. In this final section, we want to step away from the ethics of AI risk, to briefly explore how our discussion connects to a number of broader philosophical issues in ethics. We will discuss three themes: the possibility of an anti-anthropic ethics, the potential use of superintelligent AI as a tool for our ethical inquiry, and a way that superintelligent AI could undermine our confidence in our own ethical capacities.

6.1 Could ethics be hostile to us?

Section 2 of our paper considered several possibilities on which an ethically knowledgeable and enkratic AI might be hostile to us. We can abstract away from the focus on AI, to consider the broader possibility that ethically excellent action might be hostile to us. The disturbing implications of these sorts of possibilities have perhaps been most fully explored in the case of "Humean" views about reasons. However, arguably this is the least unnerving of the hypotheses discussed. As Sharon Street has emphasized, on Humean views, even if (some) bad actors can have strong reasons to try to do terrible things to others, we can also have strong reasons to stop them.[60]

By contrast, if the sorts of value-theoretic or moral-status-theoretic theses that we discussed in §2 are correct, the ethical implications are much more profound. For

[58] (Gabriel 2020, 417) explores related ideas, suggesting that we should be concerned to align AI in ways that ensures that AI systems do not dominate us.
[59] Some of the classic relevant work we have in mind here from political philosophy includes (Rawls 1971/1999), (Raz 1986), and (Pettit 1999).
[60] (Street 2009, 293).

example, some of these theses suggest the possibility that we should each do everything that we can to bring about a superintelligent AI who will permanently immiserate humanity for the sake of mosquitos, or for the maximal proliferation of the AIs themselves. On some such views, the world of the movie *The Matrix* – in which (roughly) human beings think they are living normal lives, but really are hooked up to powerful computers and being fed a set of illusions, while their bodies are being used as an energy source for AIs – would be a sort of ethical utopia. This sort of possibility is hard to take seriously, but we suggested that we cannot simply dismiss it out of hand as absurd. The possibility that ethics is hostile to us raises deep questions about our relationship to ethics that are often obscured by the assumption that acting well is, generally, good for us.

6.2 Implications for AI as an epistemic resource in ethics

Our discussion in this paper has focused on the relationship of superintelligent AI ethical inquiry to the potential threats posed by AI *agency*. But suppose that we make the highly controversial assumption that we can reliably prevent AIs from exercising their agency in the world. [61] Then (assuming minimal realism and AI ethical reliability) it might seem that superintelligent AI could help us make progress in ethics. AIs might help by directly sharing their ethical insights, but also by helping us to better understand ethically relevant facts of various sorts, better understand the very sorts of metaethical issues that we have made salient in this paper, better understand how to make our own ethical opinions more coherent, better predict the outcomes of options, etc.

One potential complication for this possibility arises because of a central aspect of leading contemporary AI research and development. This is that we are not, in general, able to understand precisely how currently cutting edge large language model-style AIs reach their various "conclusions". And this kind of opacity might be deeper and more systematic if superintelligent AI emerged: perhaps we would be incapable of understanding a superintelligent A's reasoning, even if it could make it explicit for us. This sort of opacity might create extra difficulties for our attempts to determine whether an AI inquirer was helping us discover ethical facts, or leading us down a misleading ethical rabbit hole. The same is true for whether we would be in a good position to learn about the relative merits of the normative concepts that AI was using, especially if those turn out to be significantly different from our own.

This issue about opacity is perhaps of special interest in ethical inquiry, because many philosophers think there are deep connections between intelligibility and important ethical phenomena, such as ethical reasoning, ethical knowledge, or justifiable deference to the ethical views of others.[62] If these philosophers are right, then even if superintelligent AI were to discover certain ethical facts – and we had good reason to think it had done so – there might be important limits to what goods we can achieve by relying on those discoveries. Of course, one might think that a

[61] For pessimism about the prospects for preventing superintelligent AI from acquiring power, see (Bostrom 2014, Ch. 9).
[62] See, for example, (Hills 2020).

gain in accuracy is more important, but it might be distressing if we found ourselves constantly needing to sacrifice intelligibility for the sake of ethical accuracy.

6.3 Implications for self-trust

This potential sacrifice connects to a final issue. Consider the confidence that we have in our own capacity to engage in ethical inquiry. This confidence is arguably premised on some kind of epistemic "self-trust", which many take to be crucial to the foundations of epistemology. To illustrate, take well-known global skeptical challenges, such as ones about our not being able to tell whether or not we are brains-in-vats, being systematically deceived to think we are living normal, embodied human lives. One of the main ideas of recent anti-skeptical epistemology is that, in order to block deep skeptical challenges that emerge from such scenarios, we need to – and, indeed, are somehow epistemically *entitled* to – believe that our basic reasoning capacities are not off-track, root and branch.[63] In recent work, we've argued that something akin to this kind of "self-trust" idea likely needs to underwrite our engagement with conceptual ethics as well: in particular, to block a certain kind of skeptical challenge about a (seemingly vicious) kind of "vindicatory circularity" involved in using our seemingly most "authoritative" normative and evaluative concepts to evaluate those very same concepts.[64] What happens to these kinds of self-trust replies to skeptical challenges, however, when we are in the presence of entities whose reasoning capacities far outstrip our own? Are we still entitled to the relevant kinds of "self-trust" here? Or do we need to adjust in some way? And how, if at all, does that shift the kinds of views we should have about foundational issues in epistemology, conceptual ethics, or other areas of philosophy?

We hope that philosophers take up these further issues, and others that connect to our discussion here, in future work at greater length. In our view, serious philosophical reflection about the ethical issues raised by the possibility of the emergence of superintelligent AI is just beginning. We think it's an open question how much this reflection will help us sharpen our thinking about existing philosophical issues, or how much it will help us uncover new philosophical issues worth thinking more about. It's also an open question how much philosophical reflection can (let alone will) matter to the most pressing practical concerns we face with regard to AI. Our hope is that, if nothing else, this paper helps to encourage new work that explores these issues in the coming years.

Acknowledgments

Thanks to David Braddon-Mitchell, Mark Budolfson, Rosa Cao, Herman Cappelen, Ruth Chang, Matti Eklund, Shao-An Hsu, Mikayla Kelley, Niko Kolodny, Catherine Lapey, Cara Lewis, Eden Lin, Artem Malykh, Russell McIntosh, Matt McKeever, Adrian Russian, Rachel Sterken, Marcus Willaschek, and participants in McPherson's OSU grad seminar on ethics of AI risk for helpful discussion and feedback on ideas in this paper. Earlier versions of this paper were presented at

[63] See, for example, (Wright 2004) and (Pryor 2000).
[64] (McPherson and Plunkett 2021b).

Stanford University, Central European University, University of Freiburg, University of Oslo, and the University of Texas, Austin. Thanks to everyone who participated in those discussions.

Works Cited

AI(impactssurvey). https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/
Bengson, John. 2015. The Intellectual Given. *Mind* 124 (495):707-760.
Björnsson, Gunnar, and Stephen Finlay. 2010. Metaethical Contextualism Defended. *Ethics* 121 (1):7-36.
Blackburn, Simon. 1993. *Essays in Quasi-Realism*. Oxford: Oxford University Press.
———. 1998. *Ruling Passions*. Oxford: Clarendon Press.
Bostrom, Nick. 2012. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines* 22 (2):71-85.
———. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
Budolfson, Mark, and Dean Spears. 2022. Does the Repugnant Conclusion have important implications for axiology or for public policy? In *Oxford Handbook of Population Ethics*, edited by G. Arrhenius.350–C15.P105.
Burgess, Alexis, and David Plunkett. 2013a. Conceptual Ethics I. *Philosophy Compass* 8 (12):1091-1101.
———. 2013b. Conceptual Ethics II. *Philosophy Compass* 8 (12):1102-1110.
———. 2020. On the relation between conceptual engineering and conceptual ethics. *Ratio*.
Camp, Elisabeth. 2017. Metaethical Expressivism. In *The Routledge Handbook of Metaethics*, edited by T. McPherson and D. Plunkett. New York City: Routledge.87-101.
Cappelen, Herman, and David Plunkett. 2020. A Guided Tour of Conceptual Engineering and Conceptual Ethics. In *Conceptual Engineering and Conceptual Ethics*, edited by A. Burgess, H. Cappelen and D. Plunkett. Oxford: Oxford University Press.1-26.
Cook, Tyler. 2023. Robust Artificial Moral Agents and Metanormativity. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*:162-169.
Cuneo, Terence. 2017. Quasi-realism. In *The Routledge Handbook of Metaethics*, edited by T. McPherson and D. Plunkett. New York: Routledge.626-642.
Darwall, Stephen L. 2003. *Contractarianism/contractualism*, *Blackwell readings in philosophy ; 8*. Malden, MA: Blackwell Pub.
Dasgupta, Shamik. Manuscript. The Meta-Ethics of Artificial Intelligence: Are We Beholden to Normative Joints?
DeGrazia, David. 2008. Moral Status As a Matter of Degree? *Southern Journal of Philosophy* 46 (2):181-198.
Dung, Leonard. 2023. Current cases of AI misalignment and their implications for future risks. *Synthese* 202 (5):1-23.
Eklund, Matti. 2017. *Choosing Normative Concepts*. Oxford: Oxford University Press.
Enoch, David. 2010. The epistemological challenge to metanormative realism: how best to understand it, and how to cope with it. *Philosophical Studies* 148 (3):413-438.

Gabriel, Iason. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30 (3):411-437.
Gauthier, David. 1986. *Morals by Agreement*. Oxford: Oxford University Press.
Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, Mass.: Harvard University Press.
———. 2003. *Thinking How to Live*. Cambridge, MA: Harvard University Press.
Greaves, Hilary, and Will MacAskill. Manuscript. The Case for Strong Longtermism.
Häggström, Olle. 2018. Challenges to the Omohundro–Bostrom framework for AI motivations. *Foresight* 21 (1).
Harman, Elizabeth. 1999. Creation Ethics: The Moral Status of Early Fetuses and the Ethics of Abortion. *Philosophy and Public Affairs* 28 (4):310-324.
Harman, Gilbert. 1975. Moral relativism defended. *Philosophical Review* 84 (1):3-22.
Hendrycks, Dan, Mantas Mazeika, and Thomas Woodside. 2023. An Overview of Catastrophic AI Risks.
Hills, Alison. 2020. Moral Testimony: Transmission Versus Propagation. *Philosophy and Phenomenological Research* 101 (2):399-414.
Horwich, Paul. 1980 / 1998. *Truth*. Oxford: Oxford University Press.
Huemer, Michael. 2007. Compassionate phenomenal conservatism. *Philosophy and Phenomenological Research* 74 (1):30–55.
Hume, David. 1739/2000. *A Treatise of Human Nature*. Oxford: Oxford University Press.
Institute, Dair. https://www.dair-institute.org/blog/letter-statement-March2023/
Joyce, Richard. 2001. *The Myth of Morality*. Cambridge: Cambridge University Press.
Kagan, Shelly. 1998. *Normative Ethics*. Boulder, Colo.: Westview Press.
Korsgaard, Christine M. 1996. *The Sources of Normativity*. New York: Cambridge University Press.
MacAskill, Will. 2022. *What We Owe The Future*. New York: Basic Books.
Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. New York: Penguin.
McPherson, Tristram. 2014. A Case for Ethical Veganism. *The Journal of Moral Philosophy* 11 (6):677-703.
———. 2022. Expressivism Without Minimalism. In *Meaning, Decision, and Norms: Themes From the Work of Allan Gibbard*, edited by B. Dunaway and D. Plunkett. Ann Arbor, MI: Maize Books, University of Michigan Publishing.147-170.
McPherson, Tristram, and David Plunkett. 2017. The Nature and Explanatory Ambitions of Metaethics. In *The Routledge Handbook of Metaethics*, edited by T. McPherson and D. Plunkett. New York: Routledge.1-25.
———. 2020. Conceptual Ethics and The Methodology of Normative Inquiry. In *Conceptual Engineering and Conceptual Ethics*, edited by A. Burgess, H. Cappelen and D. Plunkett. Oxford: Oxford University Press.274-303.
———. 2021a. Conceptual Ethics, Metaepistemology, and Normative Epistemology. *Inquiry*.
———. 2021b. Evaluation Turned on Itself: The Vindicatory Circularity Challenge to the Conceptual Ethics of Normativity. In *Oxford Studies in Metaethics, Vol. 16*, edited by R. Shafer-Landau. Oxford: Oxford University Press.207-232.
———. 2024a. The Fragmentation of Authoritative Normativity. In *Oxford Studies in Metathics, Vol. 19*, edited by R. Shafer-Landau. Oxford: Oxford Studies in Meaethics.1-28.

———. 2024b. Metaethics and the Conceptual Ethics of Normativity. *Inquiry* 67 (1):93-126.
———. Forthcoming. Authoritative Normativity. In *The Oxford Handbook of Meta-ethics*, edited by D. Copp and C. S. Rosati.
Moor, James. 2006. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21:18-21.
Olson, Jonas. 2014. *Moral Error Theory: History, Critique, Defence*. Oxford: Oxford University Press.
———. 2017. Error Theory in Metaethics. In *The Routledge Handbook of Metaethics*, edited by T. McPherson and D. Plunkett. New York City: Routledge.58-71.
Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.
———. 2011. *On What Matters*. Oxford: Oxford University Press.
Petersen, Steve. 2017. Superintelligence as superethical. In *Robot Ethics 2. 0: New Challenges in Philosophy, Law, and Society*, edited by P. Lin, K. Abney and R. Jenkins: Oxford University Press.322-337.
Pettit, Philip. 1999. *Republicanism: A Theory of Freedom and Government*. Oxford: Oxford University Press.
Plunkett, David. 2020. Conceptual truths, evolution, and reliability about authoritative normativity. *Jurisprudence* 11 (2):169-212.
Pryor, James. 2000. The skeptic and the dogmatist. *Noûs* 34 (4):517–549.
Railton, Peter. 1984/2003. Alientation, Consequentialism, and the Demands of Morality. In *Facts, Values, and Norms: Essays toward a Morality of Consequence*: Cambridge: Cambridge Univ Pr.151-186.
———. 2020. Ethical Learning, Natural and Artificial. In *Ethics of Artificial Intelligence*, edited by S. M. Liao. New York: Oxford University Press.45-78.
Rawls, John. 1971/1999. *A Theory of Justice (revised edition)*. Cambridge, Mass.: Harvard University Press.
Raz, Joseph. 1986. *The Morality of Freedom*. Oxford: Oxford University Press.
Risberg, Olle. 2023. Ethics and the Question of What to Do. *Journal of Ethics and Social Philosophy* 25 (2):376-412.
risk, Statement on AI. https://www.safe.ai/work/statement-on-ai-risk.
Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*: Viking.
Scanlon, T. M. 1998. *What We Owe To Each Other*. Cambridge: Harvard University Press.
Schechter, Joshua. 2017. Explanatory Challenges in Metaethics. In *The Routledge Handbook of Metaethics*, edited by T. McPherson and D. Plunkett. New York: Routledge.443-458.
Schroeder, Mark. 2007. *Slaves of the Passions*. Oxford: Oxford University Press.
———. 2008. *Being For: Evaluating the Semantic Program of Expressivism*. Oxford: Oxford University Press.
Sebo, Jeff. 2021. Animals and Climate Change. In *Philosophy and Climate Change*, edited by M. Budolfson, T. McPherson and D. Plunkett. Oxford: Oxford University Press.42-66.
Silk, Alex. 2017. Metaethical Contextualism. In *The Routledge Handbook of Metaethics*, edited by T. McPherson and D. Plunkett. New York City: Routledge.102-118.

Singer, Peter. 1975/2009. *Animal Liberation (updated edition)*. New York: HarperCollins.

Smith, Michael. 1994. *The Moral Problem*. Cambridge: Blackwell.

Stojanovic, Isidora. 2017. Error Theory in Metaethics. In *The Routledge Handbook of Metaethics*, edited by T. McPherson and D. Plunkett. New York City: Routledge.119-134.

Street, Sharon. 2006. A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies* 127 (1):109-166.

———. 2009. In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters. *Philosophical Issues* 19:273-298.

———. 2012. Coming to Terms with Contingency : Humean Constructivism about Practical Reason. In *Constructivism in Practical Philosophy*, edited by J. Lenman and Y. Shemmer. Oxford: Oxford University Press.

Streumer, Bart. 2017. *Unbelievable Errors: An Error Theory About All Normative Judgments*. New York: Oxford University Press.

Tappolet, Christine. 2016. *Emotions, Value, and Agency*: Oxford University Press UK.

Tegmark, Max. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. London: Allen Lane.

Thomson, Judith Jarvis. 1976. Killing, Letting Die, and the Trolley Problem. *The Monist* 59 (2):204-217.

Tubert, Ariela, and Justin Tiehen. Forthcoming. Value alignment, human enhancement, and moral revolutions. *Inquiry*.

Vold, Karina, and Daniel R. Harris. Forthcoming. How does Artificial Intelligence Pose an Existential Risk? In *Oxford Handbook of Digital Ethics*, edited by C. Véliz.

Williams, Bernard. 1979. Internal and External Reasons. In *Rational Action*, edited by R. Harrison: Cambridge University Press.101-113.

Wright, Crispin. 2004. Warrant for nothing (and foundations for free)? *Aristotelian Society Supplementary Volume* 78 (1):167–212.

Yampolskiy, Roman V. 2016. Taxonomy of Pathways to Dangerous AI. *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence AI, Ethics, and Society: Technical Report WS-16-02*:143-148.

Yudkowsky, Eliezer. 2008. Artificial Intelligence as a Positive and Negative Factor in Global Risk. In *In Global Catastrophic Risks*, edited by N. B. a. M. M. Ćirković. New York: Oxford University Press.308-345.

Zamir, Tzachi 2007. *Ethics and the Beast: A Speciesist Argument for Animal Liberation*. Princeton: Princeton University Press.