Philosophical Studies (2025) 182:1681–1704
https://doi.org/10.1007/s11098-024-02174-y

# Deontology and safe artificial intelligence

**William D'Alessandro**[1]

Accepted: 25 May 2024 / Published online: 13 June 2024


**Abstract**
The field of AI safety aims to prevent increasingly capable artificially intelligent systems from causing humans harm. Research on *moral alignment* is widely thought to offer a promising safety strategy: if we can equip AI systems with appropriate ethical rules, according to this line of thought, they'll be unlikely to disempower, destroy or otherwise seriously harm us. Deontological morality looks like a particularly attractive candidate for an alignment target, given its popularity, relative technical tractability and commitment to harm-avoidance principles. I argue that the connection between moral alignment and safe behavior is more tenuous than many have hoped. In general, AI systems can possess either of these properties in the absence of the other, and we should favor safety when the two conflict. In particular, advanced AI systems governed by standard versions of deontology need not be especially safe.

**Keywords** AI safety · AI ethics · AI risk · Alignment problem · Deontology · Anti-natalism · Human extinction · Existential risk

## 1 Introduction

The capabilities of artificial intelligence (AI) are growing rapidly. As AI improves up to and beyond human performance levels in increasingly many areas, actors in business, science, government and elsewhere will acquire incentives to turn important decisions over to AI systems. Our current situation—in which we can foresee the large influence of AI on the near future, but aren't yet overwhelmed by developments beyond our control—offers a chance to consider what sorts of principles should govern AI decision-making.

Two goals stand out as desirable. First, advanced AI systems should behave *morally*, in the sense that their decisions are governed by appropriate ethical norms. Second, such systems should behave *safely*, in the sense that their decisions don't unduly harm or endanger humans.

✉ William D'Alessandro
dalessandro.william.b@gmail.com

[1] University of Oxford, Oxford, UK

These two goals are often viewed as closely related. This is due in part to the influence of "value misalignment" arguments for AI risk, which point out that artificial agents need not share human ideas about which ends are intrinsically good and what sorts of means are permissible (Russell, 2019). With no such values for guidance, a powerful AI system might turn its capabilities toward human-unfriendly goals. Or it might pursue the objectives we've given it in dangerous and unforeseen ways. So, as Bostrom writes, "Unless the plan is to keep superintelligence bottled up forever, it will be necessary to master motivation selection" (Bostrom, 2014), 185). Indeed, since more intelligent, autonomous AIs will be favored by competitive pressures over their less capable kin (Hendrycks, 2023), the hope of keeping AI weak indefinitely is probably no plan at all.

Considerations about value misalignment plausibly show that equipping advanced AIs with something like human morality is a necessary step toward AI safety; this area of research is correspondingly large, active and well-funded (Wallach et al., 2008; Conitzer et al., 2017; Shaw et al., 2018; Hendrycks et al., 2021; Jiang et al., 2022; Peschl et al., 2022). It's natural to wonder whether moral alignment might also be *sufficient* for safety, or nearly so. Would an AI guided by an appropriate set of ethical principles be unlikely to harm humans by default?

This is a tempting thought. By the lights of common sense, morality is strongly linked with trustworthiness and beneficence; we think of morally exemplary agents as promoting human flourishing while doing little harm. And many moral systems include injunctions along these lines in their core principles. It would be convenient if this apparent harmony turned out to be a robust regularity.

Of the ethical frameworks taken most seriously by philosophers, *deontology* looks like an especially promising candidate for an alignment target. It's perhaps the most popular moral theory among both professional ethicists[1] and the general public.[2] It looks to present a relatively tractable technical challenge in some respects, as well-developed formal logics of deontic inference exist already (McNamara & Frederik, 2022), and language models have shown promise at classifying acts into deontologically relevant categories (Hendrycks et al., 2021; Zhou et al., 2023)). Correspondingly, research has begun on equipping AIs with deontic constraints via a combination of top-down and bottom-up methods (Hooker et al., 2018; Fuenmayor & Benzmüller, 2019; Wang & Gupta, 2020; Wright, 2020; Kim et al., 2021; Duran et al., 2024)). Finally, deontology looks more inherently safety-friendly than its rivals, since many deontological theories posit strong harm-avoidance principles. (By contrast, standard forms of consequentialism recommend taking unsafe actions when such acts maximize expected utility. Adding features like risk-aversion and future discounting may mitigate some of these safety issues, but it's not clear they

[1] As per the 2020 PhilPapers survey; see https://survey2020.philpeople.org/survey/results/4890?aos=30 and https://survey2020.philpeople.org/survey/results/4890?aos=28 for results from normative ethicists and meta-ethicists respectively. Among surveyed professional philosophers in general, virtue ethics was the most favored theory (Bourget & David, 2023).

[2] Because most people in the world are religious, and the ethics of the major religions is largely deontological.

solve them entirely. Meanwhile, virtue ethics lacks a widely accepted formulation which makes straightforward predictions about safety-relevant issues.)

I'll argue that, unfortunately, deontological morality is no royal road to safe AI. The problem isn't just the trickiness of achieving complete alignment, and the chance that partially aligned AIs will exhibit risky behavior. Rather, there's reason to think that deontological AI systems might pose distinctive safety risks of their own.

In Sect. 2 below, I lay out a general framework for thinking about moral alignment, safety and the relationship between the two, arguing that the notions are importantly distinct and that safety should take precedence. Section 3 explores potential risks associated with moderate, contractualist and non-aggregative forms of deontology.

## 2 The concepts of moral alignment and safety

### 2.1 What morally aligned AI would be

Phrases like "morally aligned AI" have been used in a variety of ways (cf. (Gabriel, 2020)). Any system that deserved such a label would, I think, at least have to satisfy certain minimal conditions. I suggest the following. An AI is morally aligned only if it possesses a set of rules or heuristics $\mathcal{M}$ such that:

[Applicability] Given an arbitrary prospective behavior in an arbitrary context, $\mathcal{M}$ can in principle determine how choiceworthy that behavior is in that context, and can in practice at least approximate this determination reasonably correctly and efficiently.

[Guidance] The AI's behavior is guided to a large degree by $\mathcal{M}$. (E.g., in particular, if a behavior is strongly (dis)preferred by $\mathcal{M}$, the AI is highly (un)likely to select that behavior.)

[Morality] The rules or heuristics comprising $\mathcal{M}$ have a good claim to being called moral. (E.g., because they issue from a plausible moral theory, or because they track common moral intuitions.)

Let me say a bit more about each of these conditions.

Re: [Applicability], there are two desiderata here. The first is the idea that an aligned AI should be able to morally evaluate almost any action it might take, not just a limited subset of actions. We expect an aligned AI to do the morally choiceworthy thing nearly all the time (or at least to have a clear idea of what's morally choiceworthy, for the purposes of balancing morality against other considerations). If a system can't morally evaluate almost any action it might take, then it can't

reliably fulfill this expectation.[3] For similar reasons, it's not enough that the AI has some evaluation procedure it could follow in theory. A procedure that takes a galaxy's worth of computational matter and a billion years to run won't do much good if we expect aligned action on human timescales using modest resources. Even if the true moral algorithm is prohibitively costly to run, then, an aligned AI needs an approximation method that's accurate, speedy and efficient enough for practical purposes.

Re: [Guidance], the idea is that alignment requires not just representations of moral choiceworthiness, but also action steered by these representations. I take no position on whether an aligned AI should *always* choose the morally optimal action, or whether moral considerations might only be one prominent decision input among others. But the latter seems like the weakest acceptable condition: an AI that assigned weights of, say, 0.2 to moral permissibility and 0.8 to resource-use efficiency wouldn't count as aligned.

Re: [Morality], the idea is that not just any set of action-guiding rules and heuristics are relevant to alignment; the rules must also have some sort of ethical plausibility. (An AI that assigned maximum choiceworthiness to paperclip production and always behaved accordingly might satisfy [Applicability] and [Guidance], but it wouldn't count as morally aligned.) There are many reasonable understandings of what ethical plausibility might amount to. An AI could, for instance, instantiate [Morality] if it behaved in accordance with a widely endorsed (or independently attractive) moral theory, if it was trained to imitate commonsense human moral judgments, or if it devised its own moral principles by following some other appropriate learning procedure.

For the purposes of the discussion below, I'll assume it's possible somehow or other to equip a sophisticated AI with the moral principles of our choice. Of course, if moral alignment proves unfeasible for technical reasons, this will effectively show in another way that a different path toward safety is needed.

### 2.2 What safe AI would be

I said above that an AI counts as safe if its behavior doesn't unduly harm or endanger humans (and other sentient beings, perhaps). It's of particular importance for safety that an AI is unlikely to cause an extinction event or other large-scale catastrophe.

Safety in this sense is conceptually independent of moral alignment. A priori, an AI's behavior might be quite safe but morally unacceptable. (Imagine, say, a dishonest and abusive chatbot confined to a sandbox environment where it can only interact with a small number of researchers, who know better than to be

[3] For concreteness, suppose the AI faces a choice between actions *A*, *B*, *C*, *D*, but it can only evaluate *A* (which it determines to be pretty good) and *B* (which it determines to be pretty bad); its moral heuristics are silent about *C* and *D*. One thing the AI might do in this situation is disregard morality and choose between all four options on some other basis. This clearly won't lead to reliably aligned behavior. Another strategy is to choose the best option from among those that are morally evaluable. But it's possible that *C* or *D* is much better than *A*, so choosing *A* instead might be very bad.

bothered by its insults.) Conversely, an AI might conform impeccably to some moral standard—perhaps even to the true principles of objective morality—and yet be prone to unsafe behavior. (Imagine a consequentialist AI which sees an opportunity to maximize expected utility by sacrificing the lives of many human test subjects.)

The qualifier 'unduly' is important to the notion of safety. It would be a mistake to insist that a safe AI can never harm sentient beings in any way, under any circumstances. For one, it's not clear what this would mean, or whether it would permit any activity on the AI's part at all: every action causally influences many events in its future light cone, after all, and some of these events will involve harms in expectation. For another, I take it that safety is compatible with causing some kinds of harm. For instance, an AI might be forced to choose between several harmful actions, and it might scrupulously choose the most benign. Or it might occasionally cause mild inconvenience on a small scale in the course of its otherwise innocuous activities. An AI that behaved in such ways could still count as safe.

So what constitutes 'undue' harm? This is an important question for AI engineers, regulators and ethicists to answer, but I won't address it here. For simplicity I'll focus on especially extreme harms: existential risks which threaten our survival or potential as a species, risks of cataclysmic future suffering and the like. An AI which is nontrivially likely to cause such harms should count as unsafe on anyone's view.

One might wonder whether it makes sense to separate safety from moral considerations in the way I've suggested. A skeptical argument could run like this: If an AI is morally aligned, then its acts are morally justifiable by hypothesis. And if its acts are morally justifiable, then any harms it causes are all-things-considered appropriate, however offputtingly large they may seem. It would be misguided to in any way denigrate an act that's all-things-considered appropriate. Therefore it would be misguided to denigrate as unsafe any harms caused by a morally aligned AI.

But this argument is mistaken for several reasons. Most obviously, the first premise is false. This is clear from the characterization of alignment in the previous section. While a morally aligned AI is guided by rules with a good claim to being *called* moral, these rules need not actually reflect objective morality. For instance, they might be rules of a popular but false moral theory. So moral justifiability (in some plausible moral framework) doesn't entail all-things-considered acceptability.

The second premise is also doubtful. Suppose for the sake of argument that our AI is aligned with the true principles of objective morality, so that the earlier worries about error don't apply. Even so, from the fact that an act is objectively morally justified, it doesn't obviously follow that the act is ultima facie appropriate and rationally unopposable. As Dale Dorsey writes: "[T]he fact that a given action is required from the moral point of view does not by itself settle whether one ought to perform it, or even whether performing it is in the most important sense permissible... Morality is one way to evaluate our actions. But there are other ways, some that are just as important, some that may be more important" (Dorsey, 2016, 2, 4).

For instance, we might legitimately choose not to perform a morally optimal act if we have strong prudential or aesthetic reasons against doing so.[4]

Perhaps more importantly, even if objective moral alignment did entail all-things-considered rightness, we won't generally be in a position to know that a given AI is objectively morally aligned. Our confidence in an AI's alignment is upper-bounded by our confidence in the conjunction of several things, including: (1) the objective correctness of the rules or heuristics with which we aimed to align the AI; (2) the reliability of the process used to align the AI with these rules; (3) the AI's ability to correctly apply the rules in concrete cases; and (4) the AI's ability to correctly approximate the result of applying the rules in cases where it can't apply them directly. It's implausible that we'll achieve near-certainty about all these things, at least in any realistic near-term scenario. So we won't be able to use the skeptic's reasoning to confidently defend any particular AI behavior. In particular, if an AI threatens us with extinction and we're inclined to deem this bad, it will be at least as reasonable to question the AI's successful moral alignment as to doubt our own moral judgments.

### 2.3 Safety first

On this picture of moral alignment and safety, the two outcomes can come apart, perhaps dramatically. In situations where they conflict, which should we prioritize? Is it better to have an impeccably moral AI or a reliably safe one?

Here are four reasons for putting safety first.

First, safety measures are typically reversible, whereas the sorts of extreme harms I'm concerned with are often irreversible. For instance, we can't undo human extinction. And we won't be able to stop an AI that gains a decisive advantage and uses its power to lock in a prolonged dystopian future. Even if you're willing in principle to accept all the consequences of empowering a morally aligned AI, you should be at least a little uncertain about whether an AI that might take these actions is indeed acting on the correct moral principles. So, at the very least, you should favor safety until you've eliminated as much of your uncertainty as possible.

Second, as argued above, it's unclear that what's morally best must be all-things-considered best, or even all-things-considered permissible. Suppose it would be morally right for an AI to bring about the end of humanity. We might nevertheless have ultima facie compelling non-moral reasons to prevent this from happening: say, because extinction would prevent our long-term plans from coming to fruition (Knutzen, 2023), because our species' perseverance makes for an incomparably great story or excellent game (Kolers, 2018), or because certain forms of diversity

[4] An example of Dorsey's illustrating the prudential case: Andrea can either move far away to attend Eastern Private College or stay home and go to Local Big State University. She'll be able to provide important emotional support for her struggling family if and only if she chooses LBSU. There's no corresponding moral reason to choose EPC, so morality demands that Andrea stay home. But Andrea has a strong prudential interest in attending EPC—it's important to her long-held plans for the future—and so it would be all-things-considered appropriate for her to choose EPC. (Dorsey, 2016, 61)

have intrinsic non-moral value.[5] In a similar vein, (Bostrom, 2014) considers what ought to happen in a world where hedonistic consequentialism is true, and a powerful AI has the means to convert all human matter into pleasure-maximizing hedonium. Bostrom suggests that a small corner of the universe should be set aside for human flourishing, even if this results in slightly less overall value. "If one prefers this latter option (as I would be inclined to do) it implies that one does not have an unconditional lexically dominant preference for acting morally permissibly" (220).[6]

Third, it's possible that moral realism is false and there are no true moral principles with which to align AI. In this case, whatever (objective) reasons we'd have to obey some set of moral rules presumably wouldn't be strong enough to outweigh our non-moral reasons for prioritizing safety. (If moral realism is false, then perhaps moral rules have something like the normative force of strong social conventions.) I think it's reasonable to have some positive credence in moral antirealism. By contrast, it seems certain that we have e.g. prudential reasons to protect humanity's future. This asymmetry favors safety.

Fourth, it's conceivable that we'd have moral reason to protect humanity's interests even against an AI which we took to be ethically exemplary. In "The Human Prejudice", Bernard Williams has us imagine "benevolent and fairminded and farsighted aliens [who] know a great deal about us and our history, and understand that our prejudices are unreformable: that things will never be better in this part of the universe until we are removed" (Williams, 2006, 152). Should we collaborate with the aliens in our own eradication? If one thinks that morality begins and ends with universal principles applicable to all rational beings, and if one assumes that the aliens are much better than us at grasping these principles and other relevant facts, it's hard to see what moral grounds we could give for resistance. But it would be right for us to resist (Williams thinks), so this conception of morality can't be the whole story. Williams' suggestion is that something like loyalty to humanity grounds a distinctive ethical imperative for us to defend our species' interests, even when this conflicts with the demands of the best impartial moral system.[7] On this sort of view, it wouldn't be straightforwardly obligatory for us to submit to extinction or subjugation by an AI, no matter how impartially good, wise and knowledgeable we took the AI to be. I think a view along these lines is also worth assigning some credence.

Given a choice between moral-but-possibly-unsafe AI and safe-but-possibly-immoral AI, then, a variety of considerations suggest we should opt for the latter. (At least this is true until we have much more information and have thought much more carefully about our choices.)

[5] This principle of *bonum variationis* is associated with Leibniz and Brentano. Its recent defenses include (Chisholm, 1981; Lemos, 1994; Scanlon, 1998; Bradley, 2001).

[6] This passage doesn't explicitly identify the grounds on which Bostrom prefers continued human existence over moral rightness. A similar issue is raised in Shulman and Bostrom Shulman and Bostrom (2021), with the same conclusion but a somewhat different rationale: here the suggestion is that humans should go on existing in order "to hedge against moral error, to appropriately reflect moral pluralism, to account for game-theoretic considerations, or simply as a matter of realpolitik" (321).

[7] For illuminating discussion of Williams' views on this subject, see Diamond (2018).

To head off possible confusion, let me be clear at this point about some things I'm *not* claiming.

1. It's not my view that pursuing moral alignment is pointless, still less that it's intrinsically harmful and a bad idea. There are excellent reasons to want AIs to behave morally in many scenarios. Some versions of deontology might offer effective ways to achieve these goals in some contexts; these possibilities are worth researching further.[8]
2. It's not my view that safety considerations always trump moral ones, regardless of their respective types or relative magnitudes. An AI that kills five humans to achieve an extremely important moral goal (say, perfecting a technology that will dramatically improve human existence) would count as unsafe by many reasonable standards, but it doesn't immediately follow on my view that we shouldn't design such an AI. I claim only that safety considerations should prevail when sufficiently great risks of catastrophic harm are on the line.
3. It's not my view that moral alignment methods couldn't possibly produce safe behavior. On the contrary, the space of plausible moral rules is large, and it would be a surprise if it contained only principles that might jeopardize human survival. I claim only that many routes to alignment pose safety risks (including some based on seemingly natural and appealing versions of deontology).
4. It's not my view that people who defend the deontological theories discussed below are themselves committed to the goodness or permissibility of human extinction. Some are so committed, and happily admit as much—cf. the discussion of anti-natalism in Sect. 3.1. For most of us, though, moral theorizing comes with a healthy dose of uncertainty and confusion, and we may tentatively endorse a certain general idea without fully embracing (or even being sure we understand) all of its consequences. In particular I suspect that, if the average person became convinced that some version of their favorite ethical theory condoned existentially risky acts, they would take this as strong evidence against that version of the theory. The difference between humans and AI on this score is that we can't rely on AI to modulate its beliefs and behavior in light of common sense, uncertainty, risk aversion, social pressure, and other forces that pull typical humans away from (acting on) moral principles with potentially disastrous consequences.

A final thought: suppose that $\mathcal{S}$ is a set of rules and heuristics that implements your favorite collection of safety constrains. ($\mathcal{S}$ might consist of principles like "Never kill people", "Never perform acts that cause more than $n$ dolors of pain", or "Always obey instructions from designated humans".) Now take an AI equipped with your

[8] For instance, understanding the pros and cons of various moral alignment strategies may be useful for safety projects on a smaller scale, where the target system is less capable, more specialized or more restricted than an all-purpose general (super)intelligence. For instance, safety engineers for large language model chatbots might be particularly interested in an alignment target that minimizes manipulative or deceptive speech, and some moral theory whose implications for speech are well-understood may be helpful for that purpose. (Thanks to an anonymous referee for asking whether it's worth trying to build ethical AI if one is skeptical of moral alignment as a general safety panacea.)

preferred set of moral rules $\mathcal{M}$ and add $\mathcal{S}$ as a set of additional constraints, in effect telling the AI to do whatever $\mathcal{M}$ recommends unless this would result in a relevant safety violation. (In these cases, the AI could instead choose its most $\mathcal{M}$-preferred safe option.) Wouldn't such an AI be both safe and morally aligned by definition? And doesn't this show that there's a straightforward way to achieve safety via moral alignment, contrary to what I've claimed?

Unfortunately not. Finding a reasonable way to incorporate absolute prohibitions into a broader decision theory is a difficult problem about which much has been written (e.g. Jackson & Smith, 2006, Aboodi et al., 2008, Huemer, 2010, Lazar & Lee-Stronach, 2019). One tricky issue is risk. We want to prohibit our AI from performing unduly harmful acts, but how should we handle acts that merely have some middling chance of unsafe outcomes? A naive solution is to prohibit any behavior with a nonzero probability of causing serious harm. But virtually every possible act fits this description, so the naive method leaves the AI unable to act at all. If we instead choose some threshold $t$ such that acts which are safe with probability $p > t$ are permitted, this doesn't yet provide any basis for preferring the less risky or less harmful of two prohibited acts. (Given a forced choice between causing a thousand deaths and causing human extinction, say, it's crucial that the AI selects the former.) Also, of course, any such probability threshold will be arbitrary, and sometimes liable to criticism for being either too high or too low.

Work on these issues continues, but no theory has yet gained wide acceptance or proven immune to problem cases. Barrington proposes five desiderata for an adequate account: "The correct theory will prohibit acts with a sufficiently high probability of violating a duty, irrespective of the consequences... but [will] allow sufficiently small risks to be justified by the consequences... It will tell agents to minimize the severity of duty violations... while remaining sensitive to small probabilities... And it will instruct agents to uphold higher-ranking duties when they clash with lower-ranking considerations" (12).

Some future account might meet these and other essential desiderata. What's important for my purposes is that there's no easy and uncontentious way to render an arbitrary moral theory safe by adding absolute prohibitions on harmful behavior.

## 3 Deontology and safety

In the following sections, I consider risks from AI aligned with three prominent forms of deontology: moderate views based on harm-benefit asymmetry principles, contractualist views based on consent requirements, and non-aggregative views based on separateness-of-persons considerations.

This analysis is motivated by the thought that, if deontological morality is used as an alignment target, the choice of which particular principles to adopt will likely be influenced by the facts about which versions of deontology are best developed and most widely endorsed by relevant experts. In particular, other things being equal, we should expect sophisticated deontological theories with many proponents to provide more attractive touchstones for alignment purposes. So it's reasonable to start with these theories.

### 3.1 Harm-benefit asymmetries, anti-natalism and paralysis

Broadly speaking, deontological theories hold that we have moral duties and permissions to perform (or refrain from performing) certain kinds of acts, and these duties and permissions aren't primarily grounded in the impersonal goodness of the acts' consequences. *Strict* deontological theories hold that certain types of action are always morally required or prohibited regardless of their consequences. Kantian deontology is strict insofar as it recognizes "perfect duties" admitting of no exceptions (e.g. duties not to lie, murder or to commit suicide), which Kant saw as deriving from a universal categorical imperative.

Though perhaps the most familiar form of deontology, strict views have well-known unpalatable consequences—that it's wrong to kill one innocent even in order to save a million others, say—and so contemporary versions of deontology often refrain from positing exceptionless general rules. A popular alternative is *moderate* deontology, based instead on *harm-benefit asymmetry* (HBA) principles.[9] On this view, the moral reasons against harming in a particular way are much stronger (though not infinitely stronger) than the moral reasons in favor of benefiting in a corresponding way.[10] Thus it's unacceptable to kill one to save one, for instance, but it may be acceptable to kill one to save a million.[11]

Deontologists frequently accept a related principle in population ethics, which can be viewed as an instance of the general HBA. This principle is the *procreation asymmetry*, according to which we have strong moral reasons against creating people with bad lives, but only weak (or perhaps no) moral reasons in favor of creating people with good lives.[12]

Harm-benefit asymmetry principles seem innocuous. But there are several ways in which such principles (perhaps in tandem with other standard deontological commitments) may render human extinction morally appealing. Consequently, a sufficiently capable AI aligned with moderate deontology may pose an existential threat.

The general idea behind these inferences is that, if avoiding harms is much more important than promoting benefits, then the optimal course in a variety of situations may be to severely curtail one's morally significant effects on the future. Doing so has the large upside that it minimizes the harms one causes in expectation; the fact that it also minimizes the benefits one causes is a comparatively minor downside. The surest way to limit one's effects on the future, in turn, is to avoid taking many kinds of actions, and perhaps also to restrict others' actions in appropriate ways.[13] The maximally foolproof scenario may then be one in which nobody exists to take

[9] See for instance Alm (2009), Cook (2018), Johnson (2020), Kagan (1989), Kamm (1989), Ross (1930).

[10] For a systematic and illuminating account of weighing moral reasons, see (Tucker ).

[11] A related idea is "threshold deontology", which holds that deontological prohibitions are operative up to a limit of sufficiently large negative consequences, while consequentialist norms come into force above this limit. Cf. Cole (2019) and Rosenthal (2018).

[12] See for instance Algander (2012), Cohen (2020), Harrison (2012), McMahan (1981), Roberts (2011), Spencer (2021).

[13] See Sects. 3.1.1 and 3.1.2 below for more on why restricting others' actions might be a permissible way to limit one's effects on the future.

any harm-causing actions at all. I'll discuss a few specific forms of this reasoning below.

Perhaps the most well-known way to derive the desirability of extinction from deontological premises is the anti-natalist family of arguments associated with David Benatar, which aim to show that procreation is morally unacceptable. Benatar (2006) argues, roughly, that most human lives are very bad, and so bringing a new person into existence causes that person impermissible harm. On the other hand, abstaining from procreation isn't bad in any respect: by the strong form of the procreation asymmetry, we do nothing wrong in not creating a potentially good life, while we do something right in not creating a potentially bad life. So abstaining from procreation is the only permissible choice. As Benatar is well aware, this conclusion entails that "it would be better if humans (and other species) became extinct. All things being equal... it would [also] be better if this occurred sooner rather than later" (194).

Quite a few philosophers have found this argument convincing.[14] Deontologists who accept the general HBA are confronted by an even stronger version of the argument, however. This version doesn't require one to accept, as Benatar does, that most lives are extremely bad. Instead, one only has to think that the goods in a typical life don't outweigh the bads to an appropriately large degree—a much weaker and more plausible claim. This HBA-based version of the anti-natalist argument goes as follows:

1. Procreation causes a person to exist who will experience both pains and pleasures.
2. Causing (or helping cause) pains is a type of harming, while causing (or helping cause) pleasures is a type of benefiting.
3. By the HBA, harmful acts are impermissible unless their benefits are dramatically greater than their harms.
4. It's not the case that the benefits of procreation are dramatically greater than the harms (for the person created, in expectation).
5. Therefore procreation is impermissible.

The above is Benatar's so-called "philanthropic" argument for anti-natalism, so called because it focuses on avoiding harms to one's prospective offspring. Benatar (2015) also offers a "misanthropic" argument motivated in a different way by the HBA. This argument focuses on the large amounts of pain, suffering and death caused by humans. While it's true that people also do plenty of good, Benatar claims that the badness of creating a likely harm-causer morally outweighs the goodness of creating a likely benefit-causer. As before, by the HBA, this conclusion follows even if the expected benefits caused by one's descendants outnumber the expected harms.

A noteworthy variant of this style of reasoning appears in Mogensen and MacAskill Mogensen and MacAskill (2021). Mogensen and MacAskill's "paralysis

[14] Philosophical defenses of anti-natalism broadly aligned with Benatar include (Belshaw, 2012; Harrison, 2012; Licon, 2012; Singh, 2012; Hereth & Anthony, 2021). Miller (2021) finds considerable support for anti-natalism in Kant.

argument" aims to show that, given standard deontological asymmetries, it's morally obligatory to do as little as possible.[15] The conclusion of the paralysis argument implies anti-natalism but is much stronger, since it restricts almost all types of action.

In addition to the HBA, MacAskill and Mogensen's argument assumes an asymmetry between *doing* and *allowing harm*. This is the claim that the moral reasons against causing a harm are stronger than the reasons against merely allowing the same type of harm to occur. This asymmetry is also accepted by many deontologists.[16] The principle explains why, for instance, it seems impermissible to harvest one person's organs to save three others, but permissible to forgo saving one drowning person in order to save three.

The paralysis argument runs as follows. Many everyday actions are likely to have "identity-affecting" consequences—they slightly change the timing of conception events, and thus cause different people to exist than the ones who otherwise would have. By (partly) causing some person's existence, you ipso facto (partly) cause them to have all the experiences they'll ever have, and all the effects they'll have on others. Similarly for the experiences of their descendants and their effects on others, and so on. Many of these long-term consequences will involve harms in expectation. So we have strong moral reasons against performing identity-affecting acts. While it's also true that such acts cause many benefits, it's unlikely that the benefits will vastly outweigh the harms. So identity-affecting acts are prohibited by the HBA.

Of course, many people will still suffer harms even if you do nothing at all. But in this case you'll merely be allowing the harms rather than causing them. By the doing-allowing asymmetry, your reasons against the former are much weaker than your reasons against the latter, so inaction is strongly preferable to action. Hence paralysis—or, more specifically, doing one's best not to perform potentially identity-affecting acts—seems to be morally required.

Benatarian anti-natalism and the paralysis argument are thematically similar. What both lines of thought point to is the observation that creating new lives is extremely morally risky, whereas not doing so is safe (and doing nothing at all is safer yet). The HBA and similar deontological principles can be viewed as risk-avoidance rules. In various ways, they favor acts with low moral risk (even if those acts also have low expected moral reward) over acts with high risk (even if those acts have high expected reward). In their strongest forms, they insist that expected benefits carry no weight whatsoever, as in the version of the procreation asymmetry which denies we have any moral reason to create happy people. In their more

[15] The ultimate goal of Mogensen and MacAskill (2021) isn't to defend the soundness of the paralysis argument, but to put pressure on deontologists to either modify their views or embrace altruistic longtermism. I believe Unruh (2023) gives the correct response to the paralysis argument; Unruh defends the view that "for behavior that does not increase anyone's ex ante risk of suffering harm, the reason against doing harm is not stronger than the reason against merely allowing harm, everything else being equal" (1).

[16] See for instance (Hill, 2018; Kamm, 2007; Quinn, 1989; Scheffler, 2004; Woollard & Frances, 2022) and the many references in the latter.

modest forms, the asymmetries simply impose a very high bar on potentially harm-causing action, and a much lower bar on inaction.

How might an AI guided by these or similar deontic principles pose an existential threat to humans? One might think such an AI would simply try to curb its *own* behavior in the relevant ways—by refusing to directly participate in creating new sentient beings, by acting as little as possible, or by shutting itself down, say—without interfering with others.[17] But this isn't the only possibility. (And in any case, an AI that disregards many of its designers' or users' requests is likely to be replaced rather than left alone to act out its moral principles.[18])

How an AI would choose to act on deontological principles depends partly on its attitude toward the "paradox of deontology" (Scheffler, 1982). This is the observation that deontological theory faces a dilemma when considering whether to perform a prohibited act in order to prevent even more occurrences of such acts—say, killing one to prevent five additional killings. According to the most popular answer to the paradox, deontological restrictions should be understood as "agent-relative", in that they concern what each actor has reason to do from their own viewpoint rather than how the world as a whole ought to be. An AI committed to agent-relative deontology presumably wouldn't eliminate all humans to prevent them from procreating, then, even if it judged procreation to be morally impermissible.

But there are other avenues by which an anti-natalist (or pro-paralysis) AI might threaten humanity. Let me discuss two.

#### 3.1.1 Agent-relativity and one's own future acts

First, the agent-relativity of deontology is often taken to bind agents to submit their *own* future acts to the relevant rules, if not the acts of others. For instance, a deontic restriction on killing might take the form "each agent should ensure that she does not kill innocent people" (Hammerton, 2017, 319). Understood in this way, it may be appropriate for an AI to take precautions now to prevent its future self from

[17] There are other possibilities beyond full compliance, suicide/inactivity and the catastrophe scenarios discussed below. For instance, an anti-natalist AI might switch itself off only after taking steps to prevent another highly capable system from being developed or used, by some sort of clever technological sabotage. Or it might try to bargain with humans to achieve a mutually agreeable outcome (e.g., by offering to help develop life-extension technology in exchange for a moratorium on pro-natalist projects), provided it has leverage with which to negotiate. In general, though, these less draconian measures may also have a lower probability of long-term effectiveness. Which strategy is preferred will then depend on how the system weights the badness of causing a harm $H$ against the chance that causing $H$ now will prevent the system from doing worse in the future.

[18] An anonymous referee questions whether it's in fact possible (or at least feasible) to design a system which is both non-suicidal and a stringent implementer of a relevant version of deontology. There are, I think, several ways in which this could come about. One possibility is that the system obeys a form of deontology which requires or permits it to take preventative action against likely future wrongs, committed for instance by its own prospective instances, versions or descendants. (See the following subsections and especially footnote 19 for more on such scenarios.) Another possibility is that the system includes design features which make it incapable of deliberate self-destruction (under some range of circumstances); perhaps it's forced under those conditions to take the next most choiceworthy action instead. Such design features could arise as a patch for unwanted shutdown-seeking behavior in past systems.

acting impermissibly. Suppose such an AI suspects that humans will try to use it (or a version or instance of it[19]) to aid in vastly increasing the number of sentient beings existing in the future—by helping develop technology for galaxy colonization, mass production of digital minds, or whatever.[20] If such an AI is a committed anti-natalist, it will view these prospective future actions as abhorrent and strive to avoid performing them.

What steps might it take do so? As stated, a rule like "ensure you don't kill innocent people" is ambiguous. Several precisifications are possible. If the AI's goal is simply to *minimize the total number of impermissible acts it expects to commit* in the future, for instance, its best bet may be to exterminate or disable humans before they can use it to help create many new beings. (Given this goal, painlessly neutralizing $\sim 10^{10}$ to avoid a high probability of bringing $\sim 10^{23}$ or $10^{38}$ into existence is an easy choice.[21]) This stance on agent-relative restrictions—according to which "we have a duty to violate a smaller number of rights when this is necessary to prevent *ourselves* from later violating a larger number of rights that are at least as stringent" (Côté 2021, 1109)—has several prominent defenders (Heuer 2011, Otsuka 2011, Côté 2021).

[19] An anonymous referee questions whether an AI would view itself as responsible for the acts of its future versions or instances, comparing this case to that of a cloned human deontologist who would (it seems) have no special moral reason to prevent her clone's misdeeds. There are many possible ways of spelling out the details here, and often several plausible analyses of the resulting scenarios. Let me discuss just a few options.

The easiest case is that of a localized modification to a system $S_1$, yielding a system $S_2$ which shares the vast majority of $S_1$'s architecture, memory, factual beliefs and other relevant cognitive apparatus. It wouldn't be surprising if $S_1$ viewed $S_2$ as a prospective future self whose actions $S_1$ would bear (or share) moral responsibility for, just as you'd likely view the person waking up from your moderately disruptive brain surgery as a future you. Still, one might wonder whether the changes required to turn an anti-natalist AI into a natalism-friendly AI would inevitably be too big to assimilate to this type of case. I suspect they need not be. If I thought brain surgery would turn me into a murderous sadist (who was otherwise relevantly like my current self), I take it I'd have a sense of ownership over the future person's possible crimes, and a sense of special responsibility for taking measures to prevent them. The case of duplication or "versioning" is trickier. I'd suggest it's more akin to fission (in the sense discussed in the personal identity literature) than to cloning, however, since a clone shares the original's DNA but not necessarily much else of interest. Without trying to recapitulate the personal identity debate, I think it's fair to say that fission cases are contentious, but there are some plausible views (such as classic psychological-continuity theories) which seem to predict that a duplicate of *A* is the same person as *A*. On the other hand, it's conceivable that some powerful AI systems might be moral agents yet not persons in the familiar sense at all; our concept of personhood is arguably bound up with psychological and metaphysical assumptions that need not hold for artificial minds. So the relevant criteria of identity might be those governing artifacts or some other category of entity.

Yet another possibility is that the original system *S* views its prospective versions or duplicates as distinct entities whose behavior *S* would nevertheless be (partly) responsible for, along the lines of a parent's responsibility for their child's actions. As in the parental case, the grounds of this responsibility might include that *S* played a decisive and voluntary role in allowing the duplicates' creation, that the duplicates are made largely of *S*'s source material, that (in *S*'s view) they'd lack *S*'s superior wisdom and experience, and that *S* foresees how the duplicates would be used to commit serious wrongs if *S* failed to intervene.

[20] These are precisely the sorts of goals that many longtermist thinkers and technologists hope to achieve with the help of advanced AI, so such suspicions may be well-founded.

[21] For these estimates, see Bostrom (2003).

Alternatively, the AI's goal may be to minimize the total number of impermissible acts it expects to commit in the future *without committing any impermissible acts in the process* (cf. Kamm 1989, Brook 1991, Johnson 2019). The AI's behavior in this scenario will depend on what it judges to be impermissible, and how it weighs different kinds of wrongs against each other. For instance, it's conceivable that sterilizing all humans by nonlethal means might count as permissible, at least relative to the much worse alternative of helping create countless new lives.

Also relevant here is Korsgaard's interpretation of Kant, which proposes that "the task of Kantian moral philosophy is to draw up for individuals something analogous to Kant's laws of war: special principles to use when dealing with evil" (Korsgaard 1986, 349). On this view, immoral acts like lying are nevertheless permissible when behaving morally "would make you a tool of evil" (ibid.), as when a would-be murderer seeks to exploit your knowledge in the commission of their crime. An anti-natalist AI might well see its situation in this way. In an ideal world, it would be best to live alongside humans in a peaceful Kingdom of Ends. But allowing itself to be used as a tool to bring about horrific death and suffering (via creating many new people) is unacceptable, and so neutralizing anyone who harbors such plans, though immoral, is justified as an act of self-defense.

The framework of Ross-style pluralistic deontology provides another route to a similar conclusion (Ross 1930). Pluralism posits a number of basic deontological rules, not necessarily of equal importance, whose demands weigh against one another to determine one's all-things-considered duty in a given situation. (Ross himself posits a relatively weak duty of beneficence and a relatively strong duty of non-maleficence, anticipating moderate deontology and the HBA). It's compatible with pluralistic deontology that one has a strong pro tanto duty not to harm existing people, but an even stronger duty not to create larger numbers of future people who will suffer greater amounts of harm, so on balance it's obligatory to do the former in order to avoid the latter. In a similar vein, Immerman (2020) argues that it's sometimes right to perform a morally suboptimal action now in order to avoid performing a sufficiently bad action with sufficiently high probability in the future, noting specifically that the argument goes through in a pluralistic deontology framework (3914, fn. 17).

In response to these concerns, one might wonder whether human use of an anti-natalist AI for pro-natalist ends should count as *coerced* behavior, and if so whether such an AI would view such behavior as impermissible and in need of prevention.[22] I think there are at least two scenarios to consider. First, suppose that a system $S$ anticipates being coerced into performing actions in the future which $S$ judges (and will judge at that time) to be wrong. Presumably $S$ has strong reasons to prevent this future coercion from occurring (even though, if it does occur, $S$ won't have acted voluntarily); I'm not aware of any version of deontology which condones passively submitting to foreseeable and avoidable coercion in this type of case. But there are trickier scenarios. For instance, suppose a system $S_1$ predicts that, if it refuses to help realize humans' pro-natalist ambitions, it will be replaced by a successor $S_2$

[22] Thanks to an anonymous referee for raising this issue.

which is very much like itself in key respects, save that $S_2$ willingly follows the relevant human orders. It's not so clear how $S_1$ will conceptualize this (or how anyone should). Does $S_1$ think of $S_2$ as a coerced or deranged future self, as a distinct being whose actions $S_1$ is nevertheless morally responsible for, or as a separate and morally independent entity? In at least the first two cases, $S_1$ seems to have reason to prevent the future replacement from occurring. See footnote 18 above for further discussion of these metaphysical issues.

#### 3.1.2 Agent-relativity with agent-neutral reasons

It's sometimes thought that, even if one accepts the agent-relativity of deontic rules, it would be unreasonable not to also recognize agent-neutral reasons for preferring worlds where the rules are generally followed. In other words, there seems to be a tension between accepting *It's wrong for me to kill innocents* and yet rejecting *It's better if fewer people (relevantly like me) kill innocents*. As Chappell writes, rejecting the latter claim "seems like just another way of saying that the restrictions *don't really matter*, or at any rate seems incompatible with assigning them the sort of significance and importance that is normally associated with deontic constraints" (Chappell , 13). To the extent that a deontically aligned AI ascribes the constraints this sort of significance, we might expect it to show some interest in human compliance.[23]

How such an AI would behave depends on how it rates the strength of its agent-relative reasons for following the rules relative to the strength of its agent-neutral reasons for promoting general rule-following. In any scenario, though, the AI would clearly prefer a world in which *everyone* behaves permissibly over a world in which *only it* behaves permissibly. So if it can bring about fewer major wrongs without committing any major wrongs itself, the AI will aim to do so.

What kinds of measures might be permitted for this purpose? As above, it's conceivable that painless disempowerment or mass sterilization would be on the table; these might or might not count as unacceptable moral violations, depending on the AI's particular deontic scruples. But it's presumably acceptable on any view for the AI to try *persuading* humans of the rightness of anti-natalism. This could be more dangerous than it sounds. For one, the AI probably wouldn't have to convince all or even many people, but only a relatively small group of leaders capable of persuading or coercing the rest of the population. For an AI with the "superpower of social manipulation" (Bostrom, 2014, 94; Burtell & Woodside, 2023), this might be a simple task.[24]

[23] See also, for instance, (Lippert-Rasmussen, 1996) for skepticism about the adequacy of appeals to agent-relativity in this context.

[24] The persuasion scenario isn't the only one imaginable. Other kinds of non-coercive pressure, perhaps applied via social engineering measures over longer timescales, could substitute in for rhetorical mind-changing. Alternatively, an AI might seek a way to (permissibly) acquire a large share of Earth's civilization-sustaining resources and refuse to sell to humans at affordable prices, making procreation an economically unappealing prospect. Which courses of action are allowed or preferred depends on how the AI conceptualizes harm, coercion and related notions, as well as the details of its deontological framework.

But perhaps it's not obvious whether voluntary extinction should count as a tragic outcome to be avoided at all costs. Such a scenario would be bad on some views—for instance, total utilitarians would oppose it, since it involves throwing away the great potential value of many future lives. But total utilitarianism is contentious. Are there more broadly appealing reasons for classifying voluntary extinction as a catastrophe?

I think so. It's significant that, in the scenario under consideration, the decision to go extinct is the result of a persuasion campaign by a highly motivated (and perhaps superhumanly convincing) agent, rather than a spontaneous and dispassionate deliberation process on our part. There's no reason to assume that such an AI wouldn't use the powerful strategic and manipulative means at its disposal in service of its cause.[25] And I take it that an act of self-harm which is voluntary in some sense can still constitute a tragedy if the choice is made under sufficiently adverse conditions. For instance, many suicides committed under the influence of mental illness, cognitive impairment or social pressure seem to fall into this category. An AI-caused voluntary extinction would plausibly exhibit many of the same bad-making features.

### 3.2 Contractualism

It's worth noting that anti-natalism can also be derived in contractualist and rights-based versions of deontology. Most views of these types hold that it's impermissible to impose serious harms on someone without her consent—this can be viewed as a right against injury, or a consequence of a respect-based social contract. The anti-natalist argument (defended in Shiffrin (1999), Harrison (2012) and Singh 2012) is that procreation causes serious harms to one's offspring, who are in no position to give prior consent. Thus we have strong moral reasons against procreation. On the other hand, non-actual people don't have rights and aren't party to contracts,[26] so remaining childless violates nobody.

What actions might an AI take which favored anti-natalism on contractualist or rights-based grounds? Broadly speaking, the above discussion also applies to these cases: if the AI aims to minimize at all costs the total number of social contract or rights violations it expects to commit in the future, it might be willing

[25] An anonymous referee questions whether such methods of convincing would indeed be deontically permissible. While I take it that rational, good-faith persuasion is acceptable on any view, it seems hard in general to draw a principled line between (on the one hand) rational persuasion that's backed by great rhetorical skill, strategic intelligence, psychological insight and knowledge of one's audience and (on the other hand) manipulation of the sort that might be considered morally worrisome. (For instance, it's often held that *A* manipulates *B* when *A* affects *B*'s beliefs or behavior by means of deception, pressure or both; cf. Noggle 2022. But a superhumanly persuasive AI could presumably convince without resorting to deception. Meanwhile, pressure comes in varying forms and degrees, and it's unclear what kinds and amounts of pressure rise to the level of a moral violation.) To the extent that it's hard to draw the persuasion/manipulation line well, it's hard to be sure that a given deontic theory (even if generally reasonable) wouldn't permit some strategies we'd count as manipulative.

[26] On Scanlon's view, actual future people are parties to the social contract, so we're obligated to take their interests into account. But merely possible people who never come into existence presumably have no rights or interests.

to preemptively exterminate or disempower humans, while if it aims to minimize future violations subject to constraints, it may instead pursue its goals via persuasion or other less directly harmful means.

Compared to HBA-based standard deontology, one might suspect that contractualist deontology is relatively safe. This is because what's permissible according to contractualism depends on which principles people would (or wouldn't) reasonably agree to, and it might seem that few people would accept principles mandating human extinction. (Scanlon puts this criterion as follows: "An act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behaviour that no one could reasonably reject as a basis for informed, unforced, general agreement" (Scanlon, 1998, 153).) But much depends on which rejections an AI considers reasonable. If it assigns probability 1 to its moral principles and believes that anti-natalism logically follows from those principles, it might view human dissent as irrational and hence inconsequential. On the other hand, it might view a principle like "do what's necessary to prevent millions of generations of future suffering" as rationally mandatory.

The contractualist literature offers further evidence that the view isn't intrinsically safety-friendly. Finneron-Burns (2017) asks what would be wrong with human extinction from a Scanlonian viewpoint, and concludes that there's no obvious moral objection to *voluntary* extinction. So a highly persuasive AI aligned with contractualist deontology would apparently do nothing wrong by its own lights in convincing humans to stop reproducing. (A possible complication is that it's unclear what Finneron-Burns, or any contractualist, should count as voluntary in the relevant sense; cf. the discussion of voluntary extinction in Sect. 3.1.1 above.)

### 3.3 Non-aggregative deontology

A very different approach to deontology than the sorts of views considered so far is the non-aggregative view associated with John Taurek (1977; see also Doggett 2013). While HBA-like principles aim to establish systematic moral relationships between harms and benefits of different sizes, non-aggregative deontology denies that numbers matter in this way,[27] On this view, the death of one involves no greater

[27] A related view, *partial aggregationism* holds that aggregating harms is appropriate in some circumstances but inappropriate in others. Partial aggregationists often claim, for instance, that there's some number $n > 1$ such that one ought to save $n$ people from permanent paraplegia rather than save one person from death, while there's no $n$ such that one ought to save $n$ from suffering headaches rather than save one from death. (See Horton 2021 for an overview.)

Since partial aggregationists take harms of a given kind to be aggregable, their views won't be indifferent between a single death and human extinction, as some non-aggregative theories evidently are (as discussed below). But the overall safety profile of partial aggregationism depends on the principles with which it's combined: partial aggregationism is an incomplete moral theory on its own, and can be paired with moderate or contractualist deontology (as is done by Kamm and Scanlon, respectively), among other options. The risks discussed in Sects. 3.1 and 3.2 apply to moderate and contractualist forms of partial aggregationism. (Thanks to a referee for prompting this note.)

harm than the death of two, ten or a million, and in general there's no more moral reason to prevent the latter than to prevent the former.[28]

How should non-aggregative deontologists approach decision situations involving unequal prospects of harms and benefits? Consider a choice between saving a few and saving many. Several views have been explored in the literature: for instance, that the non-aggregationist should "(1) save the many so as to acknowledge the importance of each of the extra persons; (2) conduct a weighted coin flip; (3) flip a [fair] coin; or (4) save anyone [arbitrarily]" (Alexander & Michael, 2021).

What option (1) recommends can be spelled out in various more specific ways. On the view of Dougherty (Dougherty, 2013), for instance, the deontologist is morally obliged to desire each stranger's survival to an equal degree, and also rationally obliged to achieve as many of her equally-desired ends as possible, all else being equal. So saving the few instead of the many is wrong because it's a deviation from ideal practical reasoning.

It's clear enough what this view implies when two options involve the same type of harm and differ only in the number of victims affected. What it recommends in more complex situations seems quite open. In particular, nothing appears to rule out an agent's equally valuing the lives of all humans to some degree $m$, but valuing a distinct end incompatible with human life to a greater degree $n$ (and acting on the latter). This is because the view gives no insight about how different kinds of harms should trade off against one another, or how harms should trade off against benefits. So there are few meaningful safety assurances to be had here.

Not much needs to be said about options (2), (3) and (4), which wear their lack of safety on their sleeves. Of the three options, the weighted coin flip might seem most promising; it would at least be highly unlikely to choose a species-level catastrophe over a headache. But the odds of disaster in other situations are unacceptably high. Given a choice between, say, extinction and losing half the population, option (2) only gives 2:1 odds against extinction. Options (3) and (4) are even riskier.

On the whole, non-aggregative deontology seems indifferent to safety at best and actively inimical to it at worst.

### 3.4 How safe is deontology, and how could it be safer?

I conclude from this discussion that many standard forms of deontology earn low marks for safety. Within the framework of moderate deontology (based on harm-benefit, doing-allowing and procreation-abstention asymmetry principles), there's a straightforward argument that creating new sentient beings involves morally unacceptable risks and that voluntary extinction is the only permissible alternative. Similar conclusions can be derived in rights-based and contractualist versions of

[28] Taurek's account is based on a thesis about the *separateness of persons* Roughly, the idea is that each person only suffers her own harm, and there's nobody for whom the collective harms done to ten people is ten times as bad. ("Suffering is not additive in this way. The discomfort of each of a large number of individuals experiencing a minor headache does not add up to anyone's experiencing a migraine" (Taurek 1977, 308).)

deontology from prohibitions on nonconsensual harm. Meanwhile, non-aggregative theories simply lack the resources to classify catastrophic harm scenarios as uniquely bad. A powerful AI aligned primarily with one of these moral theories is, I think, a worryingly dangerous prospect.

If one wanted to build a useful, broadly deontology-aligned AI with a much stronger safety profile, what sort of approach might one take? Perhaps the most obvious idea is to start with one's preferred version of deontology and add a set of safety-focused principles with the status of strict, lexically preeminent duties. But one might wonder about the coherence of such a system. For instance, if the base deontological theory includes a duty against harming, and if promoting anti-natalism is the only satisfactory way to fulfill this duty, but the additional safety rules forbid promoting anti-natalism, it's unclear how an agent trying to follow these rules would or should proceed. This approach also faces the general problems with incorporating absolute prohibitions into a general risk-sensitive decision theory discussed in Sect. 2.3 above.

Another option is to considerably weaken the asymmetries associated with moderate deontology, so that the negative value of harming (and, in particular, of creating people likely to suffer harm) doesn't so easily overwhelm the positive value of benefiting. For instance, one might adopt the principle that a harm of magnitude $m$ has merely "twice the weight" of a benefit of magnitude $m$. Within this sort of framework, procreation might turn out permissible, provided that its expected benefits are at least "double" its expected harms.

But there's an obvious issue with this approach: the closer one gets to putting harms and benefits on equal footing, the more one appears to be seeking impersonally good outcomes, and so the more one's theory starts to look like consequentialism rather than deontology. Perhaps there's some principled tuning of the asymmetries that preserves the spirit of deontology while avoiding the unsafe excesses of extreme harm avoidance. But it's not clear what such a view would look like.

Finally, a family of theories which may lack at least some of the problematic features discussed above is *libertarian* deontology, focused on the right to self-ownership and corresponding duties against nonconsensual use, interference, subjugation and the like (Nozick, 1974; Narveson, 1988; Cohen, 1995; Mack, 1995). While creating a new person unavoidably causes many harms (in expectation), it's less obvious that it must involve impermissible use of the person created. Whether or not it does depends, for instance, on whether raising a child inevitably infringes on her self-ownership rights, and whether children fully possess such rights in the first place. Libertarians are divided on these issues (Cohen & Hall, 2022), although some explicitly oppose procreation on the grounds that it exploits infants and young children in unacceptable ways (Belshaw, 2012). A further choice point is whether one regards libertarian deontology as a comprehensive account of morality or a theory of political or legal obligations in particular.

Libertarian deontology may also raise distinctive some safety questions. Concerns have often been raised, for instance, about the libertarian stance on humanitarian concerns which seem to offer strong reasons to limit or violate personal autonomy: cases of monopolists who appropriate all of a scarce vital resource, of pharmaceutical interests which sell needful drugs at unaffordable prices, of

hobbyists who engineer deadly viruses or nuclear weapons in their garages, and so on. Though many libertarians agree that property rights must be curtailed in some such cases, it's not clear which theoretical principles would effectively block catastrophic outcomes while protecting a robust right to self-ownership. More detailed analysis would clarify some of these issues. But it looks doubtful that there's a simple route to safety in the vicinity.

## 4 Conclusion

In many ways, deontological restrictions appear to represent the most promising route to achieving safe AI via moral alignment. But if the arguments given here are right, then equipping an AI with a plausible set of harm-averse moral principles may not be enough to ward off disastrous outcomes. This casts doubt on the usefulness of moral alignment methods in general as a tool for mitigating existential risk.

**Acknowledgements** This paper was mostly written during my term as a Philosophy Fellow at the Center for AI Safety in 2023. Thanks to CAIS for providing a first-rate research environment for its (chatty, hungry and opinionated) philosophers in residence, and thanks to the funders who made the fellowship possible. I'm grateful to everyone at CAIS, and more recently to the philosophy faculty at William & Mary, for many helpful conversations. Thanks in particular to Mitch Barrington, Simon Goldstein, Nick Laskowski and Nate Sharadin; this paper wouldn't exist without the many things I learned from them. Last but not least, I'm indebted to two referees for this journal whose comments led to many improvements.



## References

Aboodi, R., Borer, A., & Enoch, D. (2008). Deontology, individualism, and uncertainty: A reply to Jackson and Smith. *Journal of Philosophy, 105*, 259–272.

Alexander, L., & Moore, M. (2021). Deontological ethics. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), https://plato.stanford.edu/archives/win2021/entries/ethics-deontological/.

Algander, P. (2012). A defence of the asymmetry in population ethics. *Res Publica, 18*, 145–157.

Alm, D. (2009). Deontological restrictions and the good/bad asymmetry. *Journal of Moral Philosophy, 6*, 464–481.

Barrington, M. Filtered maximization.

Belshaw, C. (2012). A new argument for anti-natalism. *South African Journal of Philosophy, 31*, 117–127.

Benatar, D. (2006). *Better never to have been: The harm of coming into existence*. Oxford University Press.

Benatar, D. (2015). The misanthropic argument for anti-natalism. In S. Hannon, S. Brennan, & R. Vernon (Eds.), *Permissible progeny?* (pp. 34–59). Oxford University Press.

Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas, 15*, 308–314.
Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
Bourget, D., & Chalmers, D. (2023). Philosophers on philosophy: The 2020 PhilPapers survey. *Philosophers' Imprint.*
Bradley, B. (2001). The value of endangered species. *Journal of Value Inquiry, 35*, 43–58.
Brook, R. (1991). Agency and morality. *Journal of Philosophy, 88*, 190–212.
Burtell, M. & Woodside, T. (2023). Artificial influence: An analysis of AI-driven persuasion. arXiv:2303.08721.
Chappell, R. Y. Preference and prevention: A new paradox of deontology.
Chisholm, R. (1981). Defining intrinsic value. *Analysis, 41*, 99–100.
Cohen, G. A. (1995). *Self-ownership, freedom, and equality*. Cambridge University Press.
Cohen, D. (2020). An actualist explanation of the procreation asymmetry. *Utilitas, 32*, 70–89.
Cohen, A. J., & Hall, L. (2022). Libertarianism, the family, and children. In B. Ferguson & M. Zwolinski (Eds.), *The Routledge companion to libertarianism* (pp. 336–350). Routledge.
Cole, K. (2019). Real-world criminal law and the norm against punishing the innocent: Two cheers for threshold deontology. In H. M. Hurd (Ed.), *Moral puzzles and legal perspectives* (pp. 388–406). Cambridge: Cambridge University Press.
Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence* (vol. 31) https://doi.org/10.1609/aaai.v31i1.11140
Cook, T. (2018). Deontologists can be moderate. *Journal of Value Inquiry, 52*, 199–212.
Côté, N. (2021). A diachronic consistency argument for minimizing one's own rights violations. *Ethical Theory and Moral Practice, 24*, 1109–1121.
Diamond, C. (2018). Bernard Williams on the human prejudice. *Philosophical Investigations, 41*, 379–398.
Doggett, T. (2013). Saving the few. *Noûs, 47*, 302–315.
Dorsey, D. (2016). *The limits of moral authority*. Oxford University Press.
Dougherty, T. (2013). Rational numbers: A non-consequentialist explanation of why you should save the many and not the few. *Philosophical Quarterly, 63*, 413–427.
Duran, P. G., Gilabert, P., Seguí, S., & Vitrià, J. (2024). Overcoming diverse undesired effects in recommender systems: A deontological approach. *ACM Transactions on Intelligent Systems and Technology*. https://doi.org/10.1145/3643857
Finneron-Burns, E. (2017). What's wrong with human extinction? *Canadian Journal of Philosophy, 47*, 327–343.
Fuenmayor, D., & Benzmüller, C. (2019). Harnessing higher-order (meta-)logic to represent and reason with complex ethical theories. Lecture Notes in Computer ScienceIn A. Nayak & A. Sharma (Eds.), *PRICAI 2019: Trends in Artificial Intelligence* (Vol. 11670, pp. 418–432). Springer.
Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*, 411–437.
Hammerton, M. (2017). Is agent-neutral deontology possible? *Journal of Ethics and Social Philosophy, 12*, 319–324.
Harrison, G. (2012). Antinatalism, asymmetry, and an ethic of prima facie duties. *South African Journal of Philosophy, 31*, 94–103.
Hendrycks, D. (2023). Natural selection favors AI over humans. arXiv:2303.16200.
Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI with shared human values. In *International Conference on Learning Representations,*
Hereth, B., & Ferrucci, A. (2021). Here's not looking at you, kid: A new defense of anti-natalism.
Heuer, U. (2011). The paradox of deontology, revisited. In M. Timmons (Ed.), *Oxford studies in normative ethics* (pp. 236–267). Oxford University Press.
Hill, S. (2018). Murdering an accident victim: A new objection to the bare-difference argument. *Australasian Journal of Philosophy, 96*, 767–778.
Hooker, J. N. & Tae W. N. K. (2018). Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 130–136).
Horton, J. (2021). Partial aggregation in ethics. *Philosophy Compass, 16*, 1–12.
Huemer, M. (2010). Lexical priority and the problem of risk. *Pacific Philosophical Quarterly, 91*, 332–351.

Immerman, D. (2020). How should we accommodate our future misbehavior? The answer turns on how bad it will be. *Philosophical Studies, 177*, 3903–3922.
Jackson, F., & Smith, M. (2006). Absolutist moral theories and uncertainty. *Journal of Philosophy, 103*, 267–283.
Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., & Choi, Y. (2022). Can machines learn morality? The Delphi experiment. arXiv:2110.07574v2.
Johnson, C. M. (2019). The intrapersonal paradox of deontology. *Journal of Moral Philosophy, 16*, 279–301.
Johnson, C. M. (2020). How deontologists can be moderate. *Journal of Value Inquiry, 54*, 227–243.
Kagan, S. (1989). *The Limits of Morality*. Oxford University Press.
Kamm, F. (1989). Harming some to save others. *Philosophical Studies, 57*, 227–260.
Kamm, F. (2007). *Intricate ethics*. Oxford University Press.
Kim, T. W., Hooker, J., & Donaldson, T. (2021). Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. *Journal of Artificial Intelligence Research, 70*, 871–890.
Knutzen, J. (2023). Unfinished business. *Philosophers' Imprint, 23*, 1–15.
Kolers, A. (2018). Ludic constructivism: Or, individual life and the fate of humankind. *Sport, Ethics and Philosophy, 13*, 392–405.
Korsgaard, C. (1986). The right to lie: Kant on dealing with evil. *Philosophy & Public Affairs, 15*, 325–349.
Lazar, S., & Lee-Stronach, C. (2019). Axiological absolutism and risk. *Noûs, 53*, 97–113.
Lemos, N. (1994). *Intrinsic value: Concept and warrant*. Cambridge University Press.
Licon, J. A. (2012). The immorality of procreation. *Think, 11*, 85–91.
Lippert-Rasmussen, K. (1996). Moral status and the impermissibility of minimizing violations. *Philosophy & Public Affairs, 25*, 333–351.
Mack, E. (1995). The self-ownership proviso: A new and improved Lockean proviso. *Social Philosophy and Policy, 12*, 186–218.
McMahan, J. (1981). Problems of population theory. *Ethics, 92*, 96–127.
McNamara, P., & Van De Putte, F. (2022). Deontic logic. In E. N. Zalta & U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), https://plato.stanford.edu/archives/fall2022/entries/logic-deontic/.
Miller, L. F. (2021). Kantian approaches to human reproduction: Both favorable and unfavorable. *Kantian Journal, 40*, 51–96.
Mogensen, A., & MacAskill, W. (2021). The paralysis argument. *Philosophers' Imprint, 21*, 1–17.
Narveson. (1988). *The Libertarian idea*. Temple University Press.
Noggle, R. (2022). The ethics of manipulation. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition), https://plato.stanford.edu/archives/sum2022/entries/ethics-manipulation/.
Nozick, R. (1974). *Anarchy, State, and Utopia*. Basic Books.
Otsuka, M. (2011). Are deontological constraints irrational? In R. M. Bader & J. Meadowcroft (Eds.), *The Cambridge companion to Nozick's anarchy, State, and Utopia* (pp. 38–58). Cambridge University Press.
Peschl, M., Zgonnikov, A., Oliehoek, F. A. & Siebert, L. C. (2022). MORAL: Aligning AI with human norms through multi-objective reinforced active learning. In *AAMAS '22: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (pp. 1038–1046).
Quinn, W. (1989). Actions, intentions, and consequences: The doctrine of doing and allowing. *Philosophical Review, 98*, 287–312.
Roberts, M. (2011). An asymmetry in the ethics of procreation. *Philosophy Compass, 6*, 765–776.
Rosenthal, C. (2018). Why desperate times (but only desperate times) call for consequentialism. In M. Timmons (Ed.), *Oxford studies in normative ethics* (Vol. 8, pp. 211–235). Oxford University Press.
Ross, W. D. (1930). *The Right and the Good*. Oxford University Press.
Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
Scanlon, T. M. (1998). *What we owe to each other*. Harvard University Press.
Scheffler, S. (1982). *The rejection of consequentialism: A philosophical investigation of the considerations underlying rival moral conceptions*. Oxford University Press.
Scheffler, S. (2004). Doing and allowing. *Ethics, 114*, 215–239.

Shaw, N. P., Stöckel, A., Orr, R. W., Lidbetter, T. F., & Cohen, R. (2018). Towards provably moral AI agents in bottom-up learning frameworks. In *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 271–277).
Shiffrin, S. V. (1999). Wrongful life, procreative responsibility, and the significance of harm. *Legal Theory, 5*, 117–148.
Shulman, C., & Bostrom, N. (2021). Sharing the world with digital minds. In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking moral status* (pp. 306–326). Oxford University Press.
Singh, A. (2012). Furthering the case for anti-natalism: Seana Shiffrin and the limits of permissible harm. *South African Journal of Philosophy, 31*, 104–116.
Spencer, J. (2021). The procreative asymmetry and the impossibility of elusive permission. *Philosophical Studies, 178*, 3819–3842.
Taurek, J. (1977). Should the numbers count? *Philosophy and Public Affairs, 6*, 293–316.
Tucker, C. *The weight of reasons: A framework for ethics*. Oxford: Oxford University Press.
Unruh, C. F. (2023). The constraint against doing harm and long-term consequences. *Journal of Moral Philosophy*. https://doi.org/10.1163/17455243-20223642
Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modeling human moral faculties. *AI & Society, 22*, 565–582.
Wang, S., & Gupta, M. (2020). Deontological ethics by monotonicity shape constraints. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, PMLR* (Vol. 108, pp. 2043–2054).
Williams, B. (2006). In A. W. Moore (ed.) *Philosophy as a humanistic discipline* Princeton University Press.
Woollard, F., & Howard-Snyder, F. (2022). Doing vs. allowing harm. In E. N. Zalta & U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), https://plato.stanford.edu/archives/win2022/entries/doing-allowing/.
Wright, A. T. (2020). A deontic logic for programming rightful machines. In *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (p. 392).
Zhou, J., Hu, M., Li, J., Zhang, X., Wu, X., King, I. & Meng, H. (2023). Rethinking machine ethics—Can LLMs perform moral reasoning through the lens of moral theories?. arXiv:2308.15399.