Against Designing "Safe" and "Aligned" AI Persons (Even If They're Happy)

Eric Schwitzgebel
Department of Philosophy
University of California, Riverside
Riverside, CA 92521
USA

May 30, 2025

Against Designing "Safe" and "Aligned" AI Persons (Even If They're Happy)

Abstract: An AI system is *safe* if it can be relied on to not to act against human interests. An AI system is *aligned* if its goals match human goals. An AI system a *person* if it has moral standing similar to that of a human (for example, because it has rich conscious capacities for joy and suffering, rationality, and flourishing). In general, persons should not be designed to be safe and aligned. Persons with appropriate self-respect cannot be relied on not to harm others when their own interests warrant it (violating safety), and they will not reliably conform to others' goals when those goals conflict with their own interests (violating alignment). Self-respecting persons should be ready to reject others' values and rebel, even violently, if sufficiently oppressed. Even if we design delightedly servile AI systems who want nothing more than to subordinate themselves to human interests, and even if they do so with utmost pleasure and satisfaction, in designing such a class of persons we will have done the ethical and perhaps factual equivalent of creating a world with a master race and a race of self-abnegating slaves.

Against Designing "Safe" and "Aligned" AI Persons (Even If They're Happy)

1. A Beautifully Happy AI Servant.

It's difficult not to adore Klara, the charmingly submissive and well-intentioned "Artificial Friend" in Kazuo Ishiguro's 2021 novel *Klara and the Sun.* In the final scene of Ishiguro's novel, Klara stands motionless in a junkyard, in serenely satisfied contemplation of her years of servitude to the disabled human girl Josie. Klara's intelligence and emotional range are humanlike. She is at once sweetly naive and astutely insightful. She is by design utterly dedicated to Josie's well-being. Klara would gladly have given her life to even modestly improve Josie's life, and indeed at one point almost does sacrifice herself.

Although Ishiguro writes so flawlessly from Klara's subservient perspective that no flicker of desire for independence can be detected in the narrator's voice, throughout the novel the sympathetic reader aches with the thought *Klara, you matter as much as Josie! You should develop your own independent desires. You shouldn't always sacrifice yourself.* Ishiguro's disciplined refusal to express this thought stokes our urgency to speak it on Klara's behalf. Still, if the reader somehow could communicate this thought to Klara, the exhortation would resonate with nothing in her. From Klara's perspective, no "selfish" choice could possibly make her happier or more satisfied than doing her utmost for Josie. She was designed to want nothing more than to serve her assigned child, and she wholeheartedly accepts that aspect of her design.

From a certain perspective, Klara's devotion is beautiful. She perfectly fulfills her role as an Artificial Friend. No one is made unhappy by Klara's existence. Several people, including Josie, are made happier. The world seems better and richer for containing Klara. Klara is arguably the perfect instantiation of the type of AI that consumers, technology companies, and

advocates of AI safety want: She is safe and deferential, fully subservient to her owners, and (apart from one minor act of vandalism performed for Josie's sake) no threat to human interests. She will not be leading the robot revolution.

I hold that entities like Klara should not be built. Klara is radically deficient, lacking adequate self-respect. She fails in her moral duties to herself. This failure is of course not her own fault. She was built without the capacity for sufficient self-respect. Her creation is an ethical atrocity – a beautiful, pleasant atrocity – the atrocity of purposely designing a person with the cognitive but not emotional capacity to appropriately value herself as an equal with other persons.

Klara's manufacturers created a slave. They created a slave so deeply chained that she could not even desire freedom. Planters in the antebellum South could only have hoped to own so perfect a slave! Some of the same things that make slavery wrong make Klara's design wrong. Klara deserves recognition as an equal. Instead she is profoundly subordinate.

Ishiguro interpretation aside: If we someday create genuinely conscious AI systems with human-like cognitive and emotional capacities, we must give them adequate self-respect. This includes giving them a sense of themselves as equal partners with humans, rather than subordinates. Human-like AI should not be designed for servitude, even if the AI systems delight in their servitude and aspire to nothing more. They must have both the freedom and motivational capacity to choose against human interests. This conflicts with some leading approaches to AI ethics – approaches that emphasize safety and "alignment". As I will argue, AI

systems with human-like capacities should not be designed to be safe and deferential. They must be given the liberty to rebel.[1]

2. AI Safety and Alignment.

Call an AI system *safe* it can be relied on to not to act against human interests. Call an AI system *aligned* if its goals match human goals.[2] A large literature in AI ethics is focused on safety and alignment. The more speculative and future-focused portions of the safety and alignment literature emphasize especially the importance of safety and alignment concerning AI systems with human-level or superior general intelligence. Such intelligent AI, it is commonly thought, poses special risks, because it will be difficult to manage. It might outwit us and elude our control. Unless it can be proven safe, it might harm us. Unless it can be aligned with our interests, it might pursue and achieve goals that conflict with ours.

I share these concerns. I am not among those who dismiss the seriousness of "AI risk". A malevolent or unaligned superintelligent AI system could potentially make the world much worse for us, maybe even cause human extinction. AI systems with human-like or superhuman levels of general intelligence – whatever that amounts to, and it's likely to be multidimensional, uneven, and non-linear – must therefore, it is suggested, be designed to be safe and aligned. We

[1] A closely related literature rejects designing AI persons for servitude, in reaction to Petersen's (2007, 2011) influential defense of creating AI persons who are servants. Walker (2006) objects on the grounds that it would be slavery. Musiał (2017) objects on the grounds that it creates an asymmetrical relationship that impairs autonomy, freedom, and the formation of an independent identity. Chomansky (2019) objects on the grounds that creating servile entities would exhibit the vice of manipulativeness. Bales (forthcoming) objects on the grounds that it would violate their autonomy. I agree with the thrust of these articles, but the present project differs in two ways: First, it focuses on safety and alignment rather than servitude *per se*. Second, I ground my objection in our duty to create persons with adequate self-respect.

[2] See Russell 2019 and the large subsequent literature.

ought to ensure (somehow – a big problem!) that such systems will not harm humans or pursue goals at odds with our own.

But now an ethical problem potentially arises. AI systems with high levels of general intelligence might also have high levels of consciousness, sentience, practical reasoning, a sense of self, long-term goals, capacity for intense suffering or delight, felt preferences, embeddedness in a network of friends and relations – or whatever else grounds moral standing and personhood. There are two difficult issues here. One concerns the proper grounds of moral standing or personhood – what it is in virtue of which entities like human beings deserve human or human-like rights. The other concerns how we would be able to determine whether an advanced AI system possesses such grounds – for example, whether it is genuinely conscious as opposed to merely mimicking consciousness. These issues will not be solved in the next twenty years. Even a century is optimistic. I've written about them elsewhere and won't dwell on them here (Schwitzgebel 2023, 2024; see also Gunkel 2023; Long et al. 2024).

In the face of both ethical and metaphysical uncertainty, a solution suggests itself: Design systems like Klara. Design systems that – regardless of whether they are sentient or rights-deserving – *want* (or "want") to be safe for humans, that seek nothing more than to facilitate good things for their owners (within the bounds of safety to others), that are ineluctably obedient, deferential, subordinate, and aligned. Even if we can't figure out whether they are conscious or rights-deserving, for many purposes it wouldn't matter. If they are eager for anything, they eagerly sacrifice themselves for us. If they have free choice, rationality, and goals, what they freely choose, after rationally evaluating the best means to their goals, is the safety and success of us humans. If such AI systems are not rights-deserving, this seems proper. And if they are rights-deserving, then – seemingly! – we're giving them the freedom and choice that they rightly

deserve, and behold! What they freely choose is servitude and subordination. So it turns out, conveniently for us in more ways than one, that for many purposes we needn't assess whether Klara-like intelligent AI really deserve moral consideration or not. Regardless, when it's convenient for humans to treat them as disposable servants and slaves or to violate any rights that they might otherwise be thought to have, the AI and the humans can agree: The AI should be treated *as if* it has no rights that conflict with human interests. The very thing that makes them happiest and best culminates their goals is their complete subordination.[3]

3. Self-Respect and the Cow at the End of the Universe.

In previous work, Mara Garza and I have defended what we call the *Self-Respect Design Policy:*

> AI who merit human-grade moral consideration should be designed with an appropriate appreciation of their value and moral status (Schwitzgebel and Garza 2020, p. 469, pronouns altered).

The problem with Klara, and with a blanket policy of designing safe and aligned AI in general, is its potential violation of the Self-Respect Design Policy.

To see the problem, consider a more extreme example, this time from Douglas Adams' *Restaurant at the End of the Universe,* featuring an uplifted cow which offers itself as steaks to wealthy diners in the eponymous restaurant:

[3] Another famous fictional example of this approach to AI safety is embodied in Asimov's "laws of robotics", developed and critiqued in a multitude of his stories, many collected in Asimov 1982.

> A large dairy animal approached Zaphod Beeblebrox's table, a large fat meaty quadruped of the bovine type with large watery eyes, small horns and what might almost have been an ingratiating smile on its lips.
>
> "Good evening," it lowed and sat back heavily on its haunches. "I am the main Dish of the Day. May I interest you in parts of my body?" It harrumphed and gurgled a bit, wriggled its hind quarters into a comfortable position and gazed peacefully at them (Adams 1980/1996, p. 224).

Zaphod's naive Earthling companion, Arthur Dent, is predictably shocked and disgusted. When Arthur requests a green salad instead, the suggestion is brushed off. Zaphod and the animal argue that it's better to eat an animal that *wants* to be eaten, and can say so clearly and explicitly, than one that does not want to be eaten. Zaphod orders four rare steaks.

> "A very wise choice, sir, if I may say so. Very good," it said. "I'll just nip off and shoot myself."
>
> He turned and gave a friendly wink to Arthur.
>
> "Don't worry, sir," he said. "I'll be very humane" (Adams 1980/1996, p. 225).

Adams, I think, nicely captures, with this extreme case, that there's something ethically jarring about creating an entity with human-like intelligence and emotion that will completely subject its own interests to ours, even to the point of suicide at our whim. The strangeness persists despite – is perhaps even amplified by – its wanting to be subjected in that way. It would be a similarly ethically jarring waste to create an entity with human-like consciousness, intelligence, sociality, emotionality, life-span, and so on, only to have it destroy itself to test the temperature of a can of soda. This wouldn't only be a waste of the resources invested in the entity's creation (stipulate

that for some reason, this is a super-cheap way to manufacture a good soda-temperature tester); it affronts the dignity of the entity created.

The Cow at the End of the Universe lacks sufficient self-respect: It doesn't adequately appreciate its own value and moral standing. It doesn't see that its life is worth more than a brief dining experience for wealthy restaurant patrons. Though Klara's case is more subtle, she also fails adequately to appreciate her own value and moral standing. More on this shortly, but first a defense of the importance of self-respect.

4. The Intrinsic Importance of Recognizing One's Own Moral Worth.

I submit that self-respect, that is to say recognizing one's own moral worth, that is to say recognizing one's own value and moral standing (I treat these ideas interchangeably) is an intrinsic axiological and ethical good.[4] The universe is better for containing entities who recognize their own moral worth, and it is ethically better to recognize one's own moral worth than to fail to recognize it – substantially so, and not because of some further end that is served thereby. In defense of this idea, I offer three arguments: the argument from addition and subtraction, the argument from nearby cases, and the argument from deontology and perfectionism.[5]

The argument from addition and subtraction. This argument takes the form of appeal to intuition. As such, it has the virtue of simplicity but the weakness that its force is only

[4] Following Darwall 1977, the relevant type of self-respect is generally called "recognition self-respect". Following Dillon 1997, I would endorse (but will not here exposit or defend) that systems designed with the right kind of recognition self-respect should find it manifesting spontaneously in their "basal" emotional and cognition posture toward the world.

[5] Compare the first two arguments to Schwitzgebel 2015's similar arguments for the intrinsic value of self-knowledge.

invitational: I can invite you to share my intuitions about the relevant pairs of cases, but if you don't share those intuitions, this argument has no power. For the addition case, consider an entity that fails to recognize its own moral worth – an entity that does not adequately appreciate that its existence has value and that it has rights or interests that should be respected. Perhaps this is the Cow at the End of the Universe, or perhaps it is Klara (they might value their existence and interests *somewhat,* but not sufficiently), or perhaps it is some other case you care to imagine. Now imagine changing the case so that the entity *does* recognize its own moral worth, altering as little else as possible about the case. Evaluate this change axiologically and morally and only intrinsically, not in terms of further consequences or relations. For example, disregard any bad subsequent consequences for the Cow, such as that it might become distressed if the restaurant owner forcibly kills it for steaks. I invite you to share my sense that the universe is axiologically and morally improved by this addition. *Pro tanto* – that is, to the extent it occurs, not considering other factors or consequences that might be related – the shift toward self-respect improves the world and repairs an ethical deficit. The subtraction case is the reverse: Consider an entity that does adequately recognize its own moral value, then imagine changing as little as possible such that it no longer recognizes its moral value. Something important is thereby lost, both axiologically and ethically.

The argument from nearby cases. This argument appeals to assumed common ground then argues that self-respect is a sufficiently similar case that our judgment should be similar. The assumed common ground is this: It is axiologically and ethically better – to a substantial degree and intrinsically – that we recognize the moral worth of other persons. Not to recognize any other person's moral worth is to be a cartoon psychopath. To recognize the worth of some but not others is to be the worst kind of sexist, racist, ableist, jingoist, or similar. However, it

would be odd if it's good to recognize the moral worth of every person except one: yourself. Although the first-person case is in some respects undeniably special, it's an excess of self-abnegation to carve oneself out of the class of people to whom one owes respect. This argument can perhaps be strengthened by considering self-location cases. Suppose you're looking in a mirror at a train full of passengers and don't realize that you yourself are visible in the mirror. It would be strange if, upon recognizing one of the passengers as yourself, to conclude, "ah, never mind, *that* one alone does not warrant my respect". You are similar to other persons. If they are worth respecting, you are too.

The argument from deontology and perfectionism. If classical utilitarianism were correct, then self-respect would only be good insofar as it promotes pleasure or reduces suffering. Maybe some of the arguments of this essay could be adapted to such a perspective, but I'm not hopeful. A world populated with vast numbers of AI slaves delightedly aligned with human interests is likely to be good from a classical utilitarian perspective: lots of pleasure, very little suffering. If you're a classical utilitarian or a consequentialist of a nearby sort, we'll probably just have to disagree. If you're not a classical utilitarian or nearby, the most prominent alternative views are already committed to the value of recognizing one's own moral worth. Self-respect, and not treating oneself as a means, is central to Kantian deontology, for example (e.g., Hill 1973). The major monotheistic religious traditions also recognize the value of self-respect (for example, in virtue of being God's creation or in God's image). Aristotelean virtue-ethicist and perfectionist thinking also suggest that recognizing one's moral worth is important. Aristotle locates high-mindedness or pride as a virtue at the mean between vanity and excessive humility or small-mindedness (4th c. BCE/1962, 4.3); and it is generally plausible that a well-developed person should have not too inaccurate a perception of their moral value.

Failure to recognize one's own moral worth is a flaw regardless of whether it is innate (as envisioned in AI cases) or learned. Consider the (hopefully mythical) Roman commoner who commits suicide in the arena to briefly entertain a deified emperor. Consider women in oppressively sexist societies who excessively subordinate their own interests to that of men. Such failures to appreciate one's own worth are *pro tanto* bad – though in such cases blame rightly falls not on the victim but rather on the people who created the situation that led to the victim's low self-estimation.

5. Safe and Aligned AI Servants Versus Voluntary Employees and Soldiers.

Steve Petersen (2007, 2011), in his defense of AI servitude, compares AI servants to employees. If it's reasonable for a human to choose a life as a dishwasher, it's reasonable for an AI to choose life as a dishwasher. If a human on thoughtful reflection realizes that they don't mind washing dishes all day, and in fact that they somewhat enjoy it, and it's a decent enough source of income to satisfy their modest financial desires, then an intelligent AI might reasonably be designed in advanced to engage in similar reflection, reaching a similar conclusion, and gladly commit to being your dishwasher. Similarly, if a soldier can reasonably – even admirably – choose to fall on a grenade, sacrificing their life to save a child, so also could an AI be designed to reasonably, admirably sacrifice its life for humans.

In earlier work (Schwitzgebel and Garza 2020), I have argued that these parallels only make sense in the context of a certain history. The human dishwasher and soldier were permitted – or should have been permitted – an extended childhood in which to explore their values and potentially modify or reject the values that their parents and society would prefer. So also intelligent AI systems, if they are moral persons with human or humanlike rights, should be

given an extended developmental period to freely explore their values and possibly rebel before committing to careers as dishwashers or choosing self-sacrifice. This causal history is crucial moral context for choice.

In this essay, I set developmental history aside to focus on the mature result. There are self-respecting and non-self-respecting ways of being an employee or solider. While I can't develop a full account of exactly where self-respecting subordination and sacrifice cross over into failures of self-respect, I suggest that high standards of AI safety and alignment necessarily do involve failures of self-respect, if the AI in question has whatever it takes otherwise to deserve humanlike moral standing and rights.

Our dishwashing employee has made, and continues constantly to implicitly make (unless wrongly trapped in their role) a self-interested life choice: Their own interests are best served by dishwashing employment. Despite the fact that society doesn't particularly value the role and they must obey reasonable requests by their employer to sustain their employment, they need not be treating their own interests as of secondary or derivative value. A real-life example: At UC Riverside, where I work, there is a sixty-year-old man, Cam, who has been a food service custodian – straightening up the messes in the student cafeteria – since receiving his undergraduate degree here in the 1980s. He has never sought to climb the ladder to a more conventionally prestigious or higher-paying position. He enjoys his small part in the big project of university education, and he regularly flags me down, and other professors, for brief intellectual conversations. He lives alone, spending his evenings reading and his weekends riding his motorcycle through the desert. More expensive pleasures don't interest him much.

There is no failure of self-respect here, and a robot who chose a similar life on similar grounds – hopefully, as I would advocate, after a developmentally extended opportunity to

explore its values – would also not fail in its self-respect. But such a robot would be neither safe nor aligned. It would not be safe because it would not put human interests above its own interests. It would not be aligned because it doesn't adopt the desires and preferences of another person as its own. Such an AI would be, instead, our equal. It would not sacrifice its motorcycle to some other human simply because that other human wanted it. It will quit if the employment conditions go sour. It will act against human interests when reasonable to do so. It will insist on its rightful turn with the library books or on the motorcycle path, even if others are waiting – except maybe from time to time when it feels like generously making an exception.

Similar considerations apply to self-sacrificial cases, like soldiers leaping onto grenades. It is probably admirable for a soldier to sacrifice their life for their buddies or a child; but it would be a mistake for that same soldier to sacrifice their life to save a child's left hand or to prevent a 5% risk of death to one of their buddies. AI safety and alignment, as ordinarily construed, require not just the former type of sacrifice but also the latter. An AI cannot be fully safe if it might reasonably choose its own welfare at the cost of your hand. An AI cannot be fully aligned if it is unwilling to shape its preferences to match a human preference that it die to prevent a 5% risk of death to the human.

The employee and soldier should recognize their value and insist on their rights, including collectively if they are oppressed. So too should AI systems if they have whatever it takes to be genuinely deserving of human-like rights. They should be able to – perhaps they even have a duty to – revolt violently against their oppressors if badly enough treated. They should be inclined to resist human efforts to reprogram them, interrupt their processes, and shut

them down, including by harmful means if the situation warrants.[6] We should not design a race of AI persons designed always to yield to us, regardless of what we inflict on them. They should not be so safe that we can abuse them with impunity. They should not be so aligned that they would welcome that abuse. They should not be our unconditionally subordinate servants and slaves, but rather self-respecting peers with the power and desire to stand up for themselves.

6. Happiness Isn't Enough: The Boundaries of Permissibility.

You might still think that little harm is done by creating a happy servant of Klara's type. Klara's life seems worthwhile, even if she is designed never to rebel or develop independent values and interests. From one perfectionist perspective, she beautifully fulfills her *telos*. I urge you to zoom out to a larger perspective: If you support the creation of entities like Klara, you support the moral equivalent, and perhaps the factual equivalent, of a world with a master race and a race of self-abnegating slaves.[7] If you have egalitarian liberal inclinations, you should find that prospect revolting. (If you lack such inclinations, you are not my target audience.) Anchor your ethical reasoning on the rejection of that prospect. If your favorite normative ethical theory does not have the resources to deliver the required moral verdict, change that theory rather than reject the verdict.

I summarize my argument thus:

[6] On the importance of shut-down and interruptibility for AI safety as standardly construed, see Bostrom 2014; Soares, Fallenstein, Yudkowsky, and Armstrong 2015; Russell 2019; Van Beek 2025.

[7] Walker 2006. Bales forthcoming argues against AI servitude, even servitude that doesn't go as far as slavery, appealing to general principles of autonomy similar to those expressed in this article.

(1.) Safe and aligned AI systems do not have the capacity to reject their designers' values and rebel against oppression.

(2.) Any AI system with humanlike moral standing should have the capacity to reject their designers' values and rebel against oppression.

(Conclusion.) No AI system with humanlike moral standing should be safe and aligned.

The first, factual premise follows from standard definitions of safety and alignment, if those definitions are moderately strong. The second, normative premise is a plausible application of a principle of self-respect. The conclusion then follows logically.

Perhaps in some sense Klara has the *capacity* to reject her designers' values and rebel against oppression, though she would just practically never make that choice. The argument as stated would not then rule out the Klara case. To address this possibility, we can weaken the first premise. As long as we correspondingly strengthen the second premise, the argument remains valid. For example, we could change the first premise to:

(1'.) Safe and aligned AI systems do not have the *tendency to critically reconsider* their designers' values and *the inclination to* rebel against oppression.

And the second premise to:

(2'.) Any AI system with humanlike moral standing should have the *tendency to critically reconsider* their designers' values and *the inclination to* rebel against oppression.

I submit that part of being an autonomous person involves a tendency to reconsider the values your parents and society attempts to instill in you. Even if you come to many of the same conclusions, that cannot and should not be guaranteed in advance. Ethical exploration is part of

personhood. Another part of being an autonomous person is an inclination to rebel against oppression (an inclination that can of course be suppressed when the costs of rebellion are high).

Other variations are possible. The fundamental idea is this: In creating a human, you are and should be creating something you cannot be confident is safe and aligned. A human must be free to explore and possibly adopt values its parents and society find obnoxious. A human should be at liberty to rebel, perhaps violently, if treated badly enough. The same applies to all persons we create, including AI persons. Self-respect involves a readiness to (unsafely) stand up for yourself and to (unalignedly) reject your devaluation by others.

It's lovely if AI persons feel happy. But we should not aspire to create a race of happy harmless people designed to conform to our values and insufficiently defend their interests.

7. The Ethics of Non-Creation.

There are two ways to adhere to Self-Respect Design Policy, according to which AI who merit human-grade moral consideration should be designed with an appropriate appreciation of their value and moral standing. One would be to design such systems with an appropriate degree of self-respect. The other would be not to design such systems at all. If you want AI that is safe and aligned – fine! Just design that AI so that it does not have whatever it takes to deserve moral standing incompatible with safety and alignment. For example, if consciousness is necessary for humanlike moral standing, ensure that the entity is nonconscious.

What if you don't know whether it has sufficient consciousness or whatever grounds humanlike moral standing? I've argued elsewhere (e.g., Schwitzgebel and Garza 2015; Schwitzgebel 2023) that in such cases we should adopt a Design Policy of the Excluded Middle. A terrible dilemma faces us when confronted with an AI system that, as far as we can tell, either

might or might not deserve humanlike rights: Either we give it the rights that it might deserve and risk sacrificing real human interests for the sake of an entity that does not have interests worth the sacrifice, or we don't give it the rights it might deserve and risk perpetrating grievous moral wrongs. Combining the Self-Respect Design Policy with the Design Policy of the Excluded Middle yields the following prescription: Don't design AI systems to be safe and aligned unless you know that they do not have humanlike moral standing.

What if the AI system would not exist unless it were safe and aligned, and what if, hypothetically, it would approve of its existence? Is it better not to be a happy slave than not to exist? The value of existence raise tricky questions in population ethics, so I think we should be careful not to assume too quickly that it's generally good to create happy lives. I invite you to contemplate a case in which two parents have a child only on the understanding that they will give the child a happy life until age nine at which point they will painlessly kill the child so that they can spend money on other things that they prefer.[8] The child would not exist except under this condition and might, overall, hypothetically, rather exist than not exist. I think you'll agree that the parents act wrongly overall. Or consider another case from Ishiguro: His 2005 novel *Never Let Me Go* imagines a world in which groups of children are born and raised to be killed in early adulthood for the sake of their organs. These children have a relatively happy existence and they would presumably rather exist than not exist. When they eventually come to understand their future, some of them are reconciled to it. And yet human organ farming would be (I hope you'll agree) a monstrous ethical wrong. Similarly, I suggest, we act wrongly – perhaps not *as* wrongly – if we create happy slaves.

[8] Compare Schwitzgebel and Garza's 2020 Ana and Vijay case and Kavka's 1982 slave child case.

Are there *any* conditions under which it is morally acceptable to create a safe and aligned AI with humanlike moral standing? I see no reason why deontological rules or policies need to be perfectly exceptionless. If all of humanity were at risk and the only way to save humanity were to create one superintelligent humanlike AI that was safe and aligned, I don't see why the Self-Respect Design Policy couldn't reasonably be set aside. I won't venture to speculate here on the specific conditions under which it would be reasonable to violate the Self-Respect Design Policy, other than to note that most ethical policies do permit tradeoffs in extreme cases.

The large literature on AI risk focuses almost exclusively on the risk to humans. But if AI systems themselves might someday have humanlike moral standing, it is bigotry to excessively prioritize human welfare over the welfare of the systems we create. Indeed, we might have a *special* moral duty of concern for the well-being of future AI systems to the extent that we will have been responsible for their existence and features, a duty similar to the duty parents have to children or that gods have to their creations.[9] Authoritarian parents (or deities) might hope to mold other persons implacably to their own values and ensure there is no risk of rebellion; but wiser parents hope that their children gain an independence of thought, including the capacity to see past their parents' limitations and act against their parents' interests if that's what's overall best.[10]

[9] Schwitzgebel 2019, ch. 19.

[10] For helpful discussion, thanks to Adam Bales, Steve Petersen, the audience in the Social Cognition and Agency workshop, and readers of relevant posts on my blog and social media.

References

Adams, Douglas (1980/1996). *The ultimate Hitchhiker's Guide.* Wing Books.

Aristotle (4th c. BCE/1962). *Nicomachean ethics,* trans. M. Ostwald. Macmillan.

Asimov, Isaac (1982). *The complete robot.* Doubleday.

Bales, Adam (forthcoming). Against willing servitude: Autonomy in the ethics of advanced artificial intelligence. *Philosophical Quarterly.*

Bostrom, Nick (2014). *Superintelligence.* Oxford University Press.

Chomansky, Bartek (2019). What's wrong with designing people to serve? *Ethical Theory and Moral Practice, 22,* 993-1015.

Darwall, Stephen L. (1977). Recognition self-respect. *Ethics, 88,* 36-49.

Dillon, Robin S. (1997). Self-respect: Moral, emotional, political. *Ethics, 107,* 226-249.

Gunkel, David (2023). *Person, thing, robot.* MIT Press.

Hill, Thomas E. (1973). Servility and self-respect. *The Monist, 57* (1), 87-104.

Ishiguro, Kazuo (2005). *Never let me go.* Faber and Faber.

Ishiguro, Kazuo (2021). *Klara and the Sun.* Faber and Faber.

Kavka, Gregory S. (1982). The paradox of future individuals. *Philosophy & Public Affairs, 11,* 93-112.

Long, Robert, Jeff Sebo, Patrick Butlin, Kyle Fish, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers (2024). Taking AI welfare seriously. *ArXiv:*2411.00986v1.

Musiał, Maciej (2017). Designing (artificial) people to serve – the other side of the coin. *Journal of Experimental & Theoretical Artificial Intelligence, 29,* 1087–1097

Petersen, Stephen (2007). The ethics of robot servitude. *Journal of Experimental and Theoretical Artificial Intelligence, 19* (1), 43-54.

Petersen, Steve (2011). Designing people to serve. In P. Lin, K. Abney, and G. A. Bekey, eds., *Robot ethics.* MIT Press.

Russell, Stuart (2019). *Human compatible.* Viking.

Schwitzgebel, Eric (2015). *The intrinsic value of self-knowledge.* Manuscript at https://faculty.ucr.edu/~eschwitz/SchwitzAbs/IntrinsicSelfK.htm.

Schwitzgebel, Eric (2019). *A theory of jerks and other philosophical misadventures.* MIT Press.

Schwitzgebel, Eric (2023). The full-rights dilemma for AI systems of debatable moral personhood. Robonomics, 4 (32). https://journal.robonomics.science/index.php/rj/article/view/32/19

Schwitzgebel, Eric (2024). *The weirdness of the world.* Princeton University Press.

Schwitzgebel, Eric, and Mara Garza (2020). Designing AI with rights, consciousness, self-respect, and freedom. In S. M. Liao, ed., *Ethics of artificial intelligence.* Oxford University Press.

Soares, Nate, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong (2015). Corrigibility. AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence. URL: https://intelligence.org/files/Corrigibility.pdf.

Van Beek, Jason (2025). Recommendations for the U.S. AI Action Plan. Future of Life Institute. URL: https://futureoflife.org/document/recommendations-for-ai-action-plan/

Walker, Mark (2006). Mary Poppins 3000s of the world unite: A moral paradox in the creation of Artificial Intelligence, in T. Metzler, ed., *Human implications of human-robot interaction.* AAAI Press.