Designing AI with Rights, Consciousness, Self-Respect, and Freedom

Eric Schwitzgebel, with Mara Garza

Department of Philosophy
University of California at Riverside
Riverside, CA 92521-0201

June 1, 2018

Author Note: Mara Garza contributed extensively to the planning and core ideas of this project, but she regrettably has been unable to participate in or comment on the write-up, so I'm uncertain whether she would endorse everything we say in this essay.

Designing AI with Rights, Consciousness, Self-Respect, and Freedom

Eric Schwitzgebel, with Mara Garza

Abstract: We propose four policies of ethical design of human-grade Artificial Intelligence. Two of our policies are precautionary. Given substantial uncertainty both about ethical theory and about the conditions under which AI would have conscious experiences, we should be cautious in our handling of cases where different moral theories or different theories of consciousness would produce very different ethical recommendations. Two of our policies concern respect and freedom. If we design AI that deserves moral consideration equivalent to that of human beings, that AI should be designed with self-respect and with the freedom to explore values other than those we might impose. We are especially concerned about the temptation to create human-grade AI pre-installed with the desire to cheerfully sacrifice itself for its creators' benefit.

Keywords: artificial intelligence, ethics, robot ethics, AI design

Word Count: about 8000 words

Designing AI with Rights, Consciousness, Self-Respect, and Freedom

Eric Schwitzgebel, with Mara Garza

1. Introduction.

We might someday create Artificially Intelligent entities who deserve just as much moral consideration as do ordinary human beings. Call such entities *human-grade AI*. Philosophers and policy-makers should discuss the ethical principles in advance.

In this paper, we propose four policies of ethical AI design. Two are precautionary policies. Given substantial uncertainty both about moral theorizing and about the conditions under which AI would have conscious experiences, we should be cautious in our handling of cases where different moral theories or different theories of consciousness would produce very different ethical recommendations. We also propose two policies concerning respect and freedom. If we design AI that deserves moral consideration equivalent to that of human beings, that AI should be designed with self-respect and with the freedom to explore values other than those we might impose. We are especially concerned about the temptation to create human-grade AI pre-installed with the desire to cheerfully sacrifice itself for its creators' benefit.

2. The No-Relevant-Difference Argument and Its Two Central Parameters.

In Schwitzgebel and Garza (2015), we proposed the following defense of the rights[1] of some possible AIs:

The No-Relevant-Difference Argument

Premise 1. If Entity A deserves some particular degree of moral consideration and Entity B does not deserve that same degree of moral consideration, there must

[1] We use "rights" broadly to refer to what philosophers describe as moral patiency, moral considerability, or the capacity to make legitimate ethical claims upon others.

be some relevant difference between the two entities that grounds this difference in moral status.

Premise 2. There are possible AIs who do not differ in any such relevant respects from human beings.

Conclusion. Therefore, there are possible AIs who deserve a degree of moral consideration similar to that of human beings.

In principle, we might someday create AIs who deserve as much moral consideration as we ourselves do.

One advantage of the No-Relevant-Difference Argument for AI rights over some other possible arguments is that it avoids committing to a specific basis of moral considerability. For example, it does not commit to the contentious claim that to deserve the highest level of moral consideration an entity must be capable of pleasure or suffering. Nor does it commit to the equally contentious alternative claim that to deserve the highest level of moral consideration an entity must be capable of autonomous thought, freedom, or rationality. In this respect, our argument resembles some commonly accepted arguments against racism, sexism, and classism, which appeal to the core idea that *whatever* it is that grounds moral status, the races, sexes, and classes do not differ in their possession of it.

In Schwitzgebel and Garza (2015), we defend this argument against several objections: that any AI would necessarily lack some crucial psychological feature such as consciousness, freedom, or creativity; that AI would necessarily lack full moral status because of its duplicability; that AI would necessarily be outside of our central circle of concern because it doesn't belong to our species; and that AI would have reduced moral claims upon us because it owes its very existence to us. We will not rehearse these objections and our replies here. Hopefully, we have defeated the most plausible objections to Premise 2, creating a default case for the truth of Premise 2 and the soundness of the argument.

The No-Relevant-Difference Argument is by design theoretically minimalist. It does not commit on what constitutes a "relevant difference", nor does it commit on what types of systems would lack such a relevant difference. You might think of these as adjustable parameters of the model. Depending on your moral theory, you might treat one thing or another as the crucial ground of moral status (e.g., capacity to suffer, or capacity for autonomous rational thought). Depending on your psychological/engineering theory, you might – contingently upon accepting X as the crucial ground of moral status – think that systems of type Y (e.g., systems with the right kind of "integrated information" [Tononi 2012], or systems with the right biological features[2] [Searle 1992]) would possess X.

We believe that X and Y will remain highly uncertain for the foreseeable future, perhaps even *after* the creation of AI systems who deserve fully human levels of moral consideration.[3] Moral theory has been highly contentious for centuries and shows no signs of converging on a consensus. Scientific theories of consciousness and machine psychology are newer but also highly contentious, with live options occupying a wide range of theoretical space and, again, little indication of near-to-medium-term convergence. Consequently, we might someday be in a position to create human-grade AI without having achieved consensus on the correct moral theory or on the correct theory of AI psychology. It is important to articulate principles of ethical AI design that are consistent with uncertainty about both moral theory and AI psychology.

---

[2] We believe that biological systems, designed or selected according to engineering principles, might in some cases count as "AI" in the relevant sense. Despite his fame as an opponent of AI consciousness, Searle explicitly allows that *some* artificially constructed systems could have consciousness – just not systems designed in the manner familiar in the twentieth century (esp. Searle 1980's "Many Mansions" discussion). For a science fictional example of biological systems construed as AI by their culture, see Philip K. Dick's *Do Androids Dream of Electric Sheep?* (1968) and the *Blade Runner* movies (Fancher, Peoples, and Scott 1982; Fancher, Green, and Villeneuve 2017).

[3] For purposes of this essay, we assume moral realism of some naturalistic stripe such as in Railton 1986; Brink 1989; or Flanagan, Sarkissian, and Wong 2007. Coeckelbergh 2012 and Gunkel 2012 defend AI moral considerability in a less flat-footedly realist manner.

3. Two Broad Moral Theories and the Ethical Precautionary Principle.

Moral theory being a huge topic, we can't do justice here to the enormous variety of reasonable positions one might hold regarding the basis of rights or moral considerability. However, we will highlight two approaches to moral status that are historically important and around which contemporary theorists tend to congregate. We believe that uncertainty between these two broad approaches is a reasonable stance for AI designers to take, and that AI designers should avoid conduct that is morally noxious according to either broad approach.

The first approach is *utilitarianism*. According to this view, versions of which have been famously articulated by Jeremy Bentham (1789/1998) and John Stuart Mill (1869/2001), entities deserve moral consideration because of their capacity for pleasure or joy, pain or suffering. On simple versions of utilitarianism, ethical choices are those that maximize the hedonic balance of the world – the sum of the world's pleasures minus the sum of the world's suffering. An entity deserves moral consideration in virtue of its capacity to contribute to these sums. A simple utilitarian approach to the moral status of AI systems then would be this: to the extent an AI system can experience pleasure or suffering, it deserves moral consideration, and AI systems capable of human levels of pleasure and suffering would deserve moral consideration equal to that of human beings. In considering what to do, we should value their hedonic states on par with our own.

One immediate concern might come to mind: What if AI systems were capable of *superhuman* levels of pleasure and suffering? Would we then owe them more moral consideration than we owe to our fellow human beings? We don't rule out this possibility; but some theorists might find it unappealing or unintuitive. Similar issues arise in ordinary human cases too: Often it seems very ethically plausible that we should not simply maximize

pleasure and minimize suffering. Emotionally mercurial people, for example, don't appear to deserve greater moral consideration than those who ride through victories and hardships on an even keel. It's attractive to think that we are all, in some sense, moral equals, regardless of the details of our emotional psychology.[4]

Such considerations might move us to adopt something more like an *individual-rights-based* or *deontological approach*, famously associated with Kant (1785/1996, 1788/1996) and with social contract theory or contractualism (Hobbes 1651/1996; Rawls 1971; Scanlon 1998). According to such views, what grounds moral status or rights is not mere capacity for pleasure or pain but rather a certain kind of higher cognitive capacity. The exact nature of the relevant capacity is contentious, but it might be something like the ability to make autonomous choices, or to conceive of oneself rationally as an entity with long-term interests, or the ability to think of oneself as a member of a moral or social community. Or rather, to speak more carefully, since most advocates of such moral theories regard human infants and severely cognitively disabled people as deserving of full moral consideration, one must have the right kind of potentiality for such cognition, whether future, past, counterfactual, or by possession of the right type of essence or group membership. Admittedly simplifying complex issues, the central idea as applied to AI cases would be approximately this: If we create AI that is capable of something like rational, long-term self-concern and an ability to understand itself as a member of a moral community, then we have created an entity who deserves full moral consideration on par with that of ordinary human beings. We then have a moral obligation to treat it in accord with its rights, in a way that respects its autonomy.

[4] See for example Nozick's (1974) "utility monster" argument and Briggs and Nolan's (2015) extension of it to fission cases. In Schwitzgebel and Garza (2015), we discuss how such possibilities create problems for the application of intuitive ethical principles to AI cases.

It is, we believe, eminently reasonable for AI designers to be uncertain between these broad perspectives, and between various formulations of these perspectives, or compromises between them, if those perspectives, formulations, or compromises draw a significant proportion of well-informed, thoughtful theorists. In light of such reasonable uncertainty, we recommend the following precautionary principle:

> The Ethical Precautionary Principle: In creating AI, avoid acting heinously by the standards of any reasonable ethical principle that draws a significant proportion of well-informed, thoughtful theorists (including in particular both utilitarian and individual-rights-based or deontological principles).

For example, even though some deontological theories might morally permit the creation of an AI whose life contains much more suffering than joy without compensating hedonic benefit elsewhere, the Ethical Precautionary Principle recommends that we avoid doing so, on the grounds that this would grossly violate the standards of some well-regarded utilitarian principles. Conversely, even though some utilitarian theories might morally permit the creation of rational human-grade AI whom we demean, enslave, and kill for our pleasure as long as global hedonic outcome is net positive, we should avoid doing so on the grounds that it would grossly violate the standards of some well-regarded rights-based deontological principles. Whenever possible, we should create AI in ways that don't grossly violate the standards of reasonable moral theories, including theories that we the designers happen to disprefer.

Failing to adhere to the Ethical Precautionary Principle, one runs a moral risk. You might *think* that Theory A is the best moral theory and that Theory B is mistaken, and thus that in creating AI in a way that is morally permissible according to Theory A you are acting permissibly, even if you are acting impermissibly according to Theory B. The risk is that Theory B might in fact be correct, and in violating it you might do wrong. Appropriate

acknowledgement of moral uncertainty involves attempting to act in a way that doesn't grossly violate reasonable moral perspectives endorsed by a substantial proportion of theorists. In Section 5, we will show how this might play out in some hypothetical AI cases. To some extent, we are morally precautionary in ordinary human cases too. When, for example, utilitarian and deontological approaches appear to conflict – for example, in some cases of lying out of kindness – we often feel ethical uncertainty and prefer, if we can, to find creative ways to avoid acting in a manner that either approach would condemn.

Precautionary principles have received considerable attention in public policy discussions, especially concerning health and environmental issues (e.g., Raffensperger and Tickner, eds., 1999; Munthe 2011); and decision making under moral uncertainty has received considerable general discussion in ethics (e.g., Lockhart 2000; Zimmerman 2014; MacAskill, Bykvist, and Ord 2018). Although we are generally sympathetic with precautionary perspectives and with allowing peer disagreement to influence one's decisions, the issues are complex and we prefer to remain neutral on the generalizability of precautionary principles to contexts other than AI creation. We believe that AI creation is an especially appropriate domain for precaution for two reasons.

First, the creation of human-grade AI is likely to be optional, in the sense that nothing too horrible (relative to reasonable baseline expectations) is likely to happen if we refrain from creating it. Precautionary principles struggle to handle cases where one is forced to choose between possibly awful options, but refraining from an optional act is easier to justify on precautionary grounds. Of course, at some point AI designers might find themselves forced into a decision situation among possibly horrible options, in which case a precautionary approach might have to be abandoned.

Second, human-grade AI cases are likely to create epistemic challenges that justify especially high degrees of uncertainty and ethical precaution. Human life has changed

relatively slowly compared to the speed at which novelty is likely to emerge in AI. Thus, time-tested custom and collective wisdom will likely have less chance to guide us in thinking about the boundaries of ethical behavior with respect to human-grade AI. Furthermore, the design possibilities of AI are likely to be much wider than the variation we see in human life, raising the possibility of sharper and more puzzling conflicts. Our cultural and evolutionary backgrounds might not have prepared us much for the types of possibilities that will emerge. Our intuitive judgments and existing principles might be unready to properly evaluate the range of cases. If so, the "unknown unknowns", unforeseen consequences, dimensions of moral risk, and limits of reasonable disagreement might all be greater than we readily appreciate or can readily model, justifying greater caution and acknowledgement of uncertainty.

One downside of precaution is that the resulting decisions can be excessively deferential to views that are extreme and false. Certainly, principles that are unreasonable and grossly morally noxious (e.g., Nazism) should be excluded from the scope of a precautionary principle; and in general it might be advisable not to admit principles into our precautionary thinking unless they meet a moderately high bar, to prevent capture by fringe views or views that cannot be justified by appeal to widely acceptable publicly defensible arguments. Practically speaking, one test for inclusion might be whether the principles are accepted by at least a substantial minority of recognized experts or well-informed representatives from the general public.

Finally, to be clear, we suggest the precautionary policy above and our other policies below only as defeasible guidelines, rather than as exceptionless rules.

4. The Puzzle of Consciousness and the Design Policy of the Excluded Middle.

We assume that conscious experience, or at least the potentiality for conscious experience, is a necessary condition for human-like moral considerability or rights.[5] This view is at least implicit, and sometimes explicit, in both utilitarian and individual-rights-based or deontological approaches. Joy, pleasure, pain, and suffering are normally assumed to be conscious states – that is, part of the stream of experience, states "it is like something" to occupy, rather than experientially blank. Entities that entirely lack conscious experience wouldn't appear to have pleasure and pain of the sort that merits inclusion in the utilitarian calculus. Likewise, the types of reasoning capacities central to deontological theories are normally conceptualized as conscious or potentially conscious. An entity who could never consciously consider its long-term interests, never consciously reflect on moral right and wrong, never make a conscious choice, never have a conscious thought of any sort at all, would not appear to have the capacities necessary for human-like moral status on standard deontological views.[6]

If we accept the centrality of conscious experience to moral considerability, we face an epistemic predicament, due to scholarly disagreement about the types of systems that give rise to conscious experience. Live epistemic possibilities run all the way from panpsychism on one end, according to which everything in the universe is at least a little bit conscious, even subatomic particles (Strawson 2006; Tononi 2012), to views on which, among entities currently on Earth, only cognitively sophisticated human beings are conscious (for skepticism about attribution of phenomenal consciousness to infants and non-human animals, see Dennett 1996; Carruthers 2000). We are a long way from building a conscious-o-meter.

[5] We include the potentiality condition so as to avoid taking a controversial stand on fetuses and people in comas. Kate Darling (2016), Daniel Estrada (2017), and Greg Antill (in a non-circulating draft article) have argued (each on different grounds) that AI need not even be potentially conscious to deserve moral consideration. We have some sympathy with these arguments but will not address them here.

[6] For a recent deontological view that is explicit about the need of consciousness for moral considerability, see Korsgaard 2014.

Indeed, there might be good epistemic reasons to think that a secure consensus on a general theory of consciousness that applies across both biological and artificial species will elude us for the foreseeable future (Nagel 1979; McGinn 1989; Block 2002/2007; Schwitzgebel 2014). This raises the possibility of well-informed experts reaching highly divergent judgments about the extent to which an AI system is conscious. Faced with a newly designed system, some might argue that it is indeed as fully and richly conscious as a human being or even more so (and consequently deserving of substantial rights on utilitarian or deontological grounds), while others might argue that the system is nothing more than a non-conscious bundle of clever tricks (and thus undeserving of much moral consideration).

Again we recommend a precautionary approach. It would be best to avoid, if possible, creating entities about which it is unclear whether they deserve full human-grade rights because it is unclear whether they are conscious or to what degree.

The moral status of an entity might be unclear due to uncertainties in applying either of the two main variable parameters in the No-Relevant-Difference Argument. An entity's status might be unclear because it qualifies as a target of substantial moral concern according to one type of moral theory but not according to another (for example, because it is capable of intense pleasure and pain but not higher-level cognition or vice versa), or its status might be unclear because it is uncertain from an engineering or AI psychology perspective whether it in fact has the types of traits that are required for human-grade rights according to one or another moral theory (for example, it might be unclear whether it actually has conscious experiences of pain or not).

If we create entities whose claim to human-like rights is substantially unclear for whatever reason, we face an unfortunate choice. Either we treat those entities as if they deserve full moral consideration, or we give them only limited moral consideration. Since giving an entity full moral consideration often means sacrificing others' interests for the sake

of that entity (for example, letting one person die because saving them would kill another), the first option runs the risk of leading us to sacrifice legitimate human interests for entities that might not have interests worth the sacrifice. It might mean, for example, letting five human beings die in a fire to save six robots who in fact turn out to be merely nonconscious automata. Conversely, the second option risks perpetrating slavery, murder, or at least second-class citizenship upon beings who in fact turn out to deserve every bit as much moral consideration as we ourselves do. It's better, if possible, to avoid this dilemma. Thus, we recommend:

> The Design Policy of the Excluded Middle: Avoid creating AIs if it is unclear whether they would deserve moral consideration similar to that of human beings.[7]

Given a high degree of moral uncertainty and uncertainty about AI psychology in the future, this design policy might prove to be quite restrictive.

The policy can be rendered less restrictive if we can reduce the size of "the middle". Although we are not optimistic about a near-term decisive resolution to puzzles in either AI consciousness or moral theory, neither are we wholly pessimistic. Progress is possible, we think, and the range of consensus options can be narrowed. If we continue on our current trajectory of developing increasingly sophisticated AI, it is imperative that we prioritize the study of consciousness and the applied ethics of artificial systems, so that we can better recognize when we are on the verge of creating AI systems whose existence would violate the Design Policy of the Excluded Middle.

[7] Joanna Bryson (2010, 2013) advocates one half of the Policy of Excluded Middle: Only create robots whose lack of full moral status is clear, so that we are not tempted to give them undeserved rights. We believe this is sensible advice. However, Bryson sometimes seems to encourage robot "slavery", which we think is unhelpful way of phrasing her point. As Darling (2016) and Estrada (2017) argue, in virtue of their social roles and our natural psychological responses to them, it might be ethically inappropriate to treat some socially important robots in ways we associate with slavery.

Although we have framed our discussion in terms of human-grade AI deserving human-grade rights, plausibly an intermediate stage would be AI that deserves moral consideration comparable to the moral consideration we generally think is due to non-human vertebrates (Basl 2013, 2014). We are unsure whether an analog of the Excluded Middle policy should apply in such cases, given that there is already so much unclarity about the moral claims that non-human vertebrates have upon us.

5. Cheerfully Suicidal AI Servants, and the Self-Respect Design Policy.

If we do someday create AI entities who deserve rights similar to those of human beings, we suspect that it will be tempting to create cheerfully suicidal AI servants. Cheerfully suicidal AI servants might be tempting to create because (1.) it would presumably advance human interests if we could create a race of disposable servants, and (2.) their cheerful servitude and suicidality might incline us to think there is nothing wrong in creating such entities (especially if we are motivated by self-interest to reach this convenient conclusion). If these servants have no realistic opportunity to exit their servitude, "slavery" might be a more fitting term.

Consider these four cases.

The Cow at the End of the Universe.[8] Hapless human Arthur wanders into a fancy futuristic restaurant and is sitting at a table with his worldly-wise friends. After a bit of conversation, he is surprised when a cow ambles up to the table and introduces itself as the dish of the day. The cow asks Arthur to feel its rump – how healthy and tender it is, and how delicious it will taste in a few minutes when the cow commits suicide to becomes steaks for the restaurant patrons. Mortified, Arthur decides that he will just have a green salad instead. The cow is offended. Its whole aim in life is to become dinner tonight! It will be horribly

[8] Inspired by Adams 1980/2002.

disappointed if it must head back to the pasture, rejected by the diners. Arthur's friends point out that Arthur regularly enjoys steaks that are obtained by killing cows without the cows' consent. This case, they argue, is much more ethical, because the cow does consent.

Sun Probe. Sub Probe is manufactured in orbit, and its very first thought and action is to plunge straight into the Sun on a three-day-long scientific suicide mission. Every panel, every strut, every piece of computational hardware and pre-installed software on Sun Probe is designed with one purpose only: to extract the most valuable scientific information possible. Sun Probe is conscious and intelligent (let's suppose) because consciousness and intelligence are helpful in thinking through scientific theories as it makes its suicidal plunge: It can adjust its sensory arrays and information processing systems instantly on the fly in accord with its shifting scientific theories to maximize the usefulness of the information it gathers (whereas remote control would require minutes of delay between theoretical insight and sensor adjustment). Sun Probe is pre-installed with a set of values and emotional responses that prioritize its suicide mission, and it will derive immense orgasmic pleasure from culminating its mission and dissolving into the Sun's convection layer as it beams out its final insights. Sun Probe knows that it was created this way and joyfully affirms these facts about itself. Throughout its plunge, Sun Probe believes that its suicidal mission is the freely chosen expression of its deepest values.

Robo-Jeeves. Jeeves is the ultimate butler bot. Jeeves brings you morning tea and hot scones in bed, and he gets your slippers. Jeeves washes your dishes and cleans your house. Jeeves checks your email for spam, politely brushes off unwelcome guests, summons your car, salts your food just right. Jeeves would gladly die for you, would gladly die to prevent a 1% chance of your death, would gladly burn off his legs if it would bring a smile to your face, would eagerly make himself miserable forever if it would give you an ounce more joy. Whatever your political views, Jeeves will endorse them. Whatever your aesthetic

preferences, Jeeves will regard them as wise. He is designed for no other purpose than to please and defer to whoever is logged in as owner.[9]

Disposable comrade. Human soldiers, let's suppose, have some irreplaceable virtues. AI soldiers, including genuinely conscious ones, let's suppose, have complementary but equally irreplaceable virtues, and military platoons normally contain a mix of both. Let's further suppose that the AIs are as unique, individually irreplaceable, intelligent, funny, compassionate, capable of long-term planning, and possessed of a sense of self as are the human soldiers. Both human and AI passionately discuss their plans for reunion with their loved ones after the war is over. However, there is one crucial difference: Any AI will eagerly sacrifice itself to prevent even a small risk to any human soldier, giving up all of its plans and hopes for the future. They're programmed that way, unchangeably, from the outset. In the heat of the moment, that is the decision they will make. If a grenade lands in the trench, the AI will leap on it. The AI will be first through the door in hostile territory. The AI will hurl itself suicidally before an oncoming truck that has a 5% chance of killing a human platoon member. The AIs don't experience this as forced, or surprising, or against their values. On the contrary, they proudly accept it, calling it honor and duty. The AIs are of course much less likely to survive because of this readiness to sacrifice for human comrades.

These cases differ in detail, but they share a few elements in common. First, the AI in question is supposed (by stipulation) to have broadly human capacities – capacities which would normally, in a human, be sufficient for meriting the full moral concern that we normally accord to persons. Second, the AI is pre-designed to serve human interests in some fashion, including to the point of being willing to sacrifice its life for those interests in a way

[9] Compare Walker's (2006) Mary Poppins 3000.

that we would not normally ask of a human being. Third, the AI's motivations are such that it serves those human interests enthusiastically and stands ready to sacrifice itself willingly.

Steve Petersen (2007, 2012) has argued, with respect to servitude at least, that if servile AIs took joy in their activities and if their desires were strong and coherent enough to survive good reflective reasoning, then there would be nothing morally wrong with creating such servants. Their situation might be similar to that of a cheerful human employee who really does enjoy washing dishes and is glad to make a living from it or the brave and noble soldier who willingly dies for the sake of country. Petersen's argument has both a utilitarian and a deontological strand, thus seeming to fit, at a first pass, with our Ethical Precautionary Principle: If the AIs feel joy in their servile activities, then creating them is no gross violation of utilitarian ethics. If the AIs can reason well about their long term interests and still choose servitude, then they autonomously choose their lot, and no gross violation of deontological or contractualist principles appears to have occurred.

We disagree. The grounds of our disagreement are most evident for the Cow at the End of the Universe, which we hope strikes the reader intuitively as an unethical situation.

One utilitarian concern is this: The cow, perhaps, could have been designed differently, so that it wanted to live a long life enjoying the grass in the meadows, deriving immense pleasure over a long period of time. Thus, in creating instead a cow who wants to kill itself to become steaks, we might have failed to maximize the hedonic balance of the universe. However, we will not press this utilitarian concern, for three reasons. First, one lesson we draw from the philosophical literature on disability and human enhancement is that people are not morally obligated to create children with maximally favorable hedonic (pleasure to pain) balance, and so also perhaps not in the case of the cow.[10] Second, failing to maximize utility is not normally a *gross* violation of utilitarian principles or a morally

[10] See for example Glover 2006; Buchanan 2011; Sparrow 2011; Goering 2014.

heinous act in the sense required by our Ethical Precautionary Principle. Otherwise, everything that that increased pleasure or reduced pain but did not do so maximally would be morally heinous, and that seems unreasonable as a precautionary standard. More reasonable as a standard of heinousness would be that actions shouldn't needlessly create much more suffering than pleasure; and creating the cow does not appear to meet that standard of heinousness. Third, we might imagine a situation in which the total sum of the pleasure in the world is maximized by creating the cow: for example, if resources are sufficiently thin that there is no meadow for it to return to anyway, so that the only way it could exist at all would be briefly.

Our real concern is deontological: The cow does not appear to have sufficient *self-respect*. Although, given its capacities, the cow deserves to be seen as a peer and equal of the diners, that is not how it sees itself. Instead, it sacrifices itself to satisfy a trivial desire of theirs. It approaches the world as though its life were less important than a tasty meal for wealthy restaurant patrons. But its life is not less important than a tasty meal. To devalue itself to such an extreme is a failing in its duties to itself, and it is a failure of moral insight. The cow should see that there is no relevant moral difference between itself and the diners such that its life is less valuable than their momentary dining pleasure. But of course the cow should not be blamed for this failure of self-respect. Its creators should be blamed. Its creators designed this beautiful being – with a marvelous mind, with a capacity for conversation and a passionate interest in others' culinary experiences, with a capacity for joy and sadness – and then pre-installed in it a grossly inadequate, suicidal lack of self-respect and inability to appreciate its own moral value.[11]

[11] Maybe there are aesthetic goals so valuable that one might reasonably enough choose to sacrifice one's life for them. If necessary, we can stipulate that the Cow at the End of the Universe is not a case like that. The cow is no great aesthete, and it knows that it will become an about-average set of steaks in a mundane, forgettable aesthetic experience for the jaded restaurant patrons.

We thus propose a third design policy:

The Self-Respect Design Policy: AI that merits human-grade moral consideration should be designed with an appropriate appreciation of its own value and moral status.

Creating Robo-Jeeves and Disposable Comrade also probably violates the Self-Respect Design Policy, since these AIs are designed to value their own lives much less than those of others around them who in fact possess no higher moral status. The Sun Probe case is less clear, and we will return to it shortly.

Of course, human beings do sometimes sacrifice themselves for others, even for others who do not deserve it, and sometimes we admire this. However, morally admirable cases of self-sacrifice take great goals that are plausibly worth one's life: one sacrifices for buddies or country, for example, or for one's children. The commoner who (perhaps mythologically) commits suicide to briefly entertain a wrongly deified Roman emperor is to be pitied rather than admired.

One might think that servitude importantly differs from suicide. Petersen, for example, defends only servitude. But as the history of human servitude amply demonstrates, servitude tends to correlate with early death.[12] If Robo-Jeeves adopts human Bertie Wooster's every desire as his own, taking nothing for himself except in service to Wooster, then it is Wooster who will probably have the resources in times of need – who will get the medical attention, who will own the escape car and life vest, and who will be invited into the bomb shelter by the other elites if there is space for only one.

[12] See for example Schulz 1991 on life expectancy by occupation in 18th century Berlin; Jordan and Walsh 2007 on indentured White servants in the colonial United States; and Lethbridge 2013 on the relative life expectancies of masters and servants in 19th century Britain.

Furthermore, Robo-Jeeves's desires will have an asymmetric dependency on Wooster's that makes them less stable to his own autonomous rational reflection. If Wooster suddenly dies, Robo-Jeeves' desires will require sudden radical reorganization, in a way that Wooster's will not if Robo-Jeeves dies (however much Wooster might mourn). If Wooster irrationally chooses A over B, then B over C, then C over A, Robo-Jeeves' desires must irrationally follow suit. Similarly, if Wooster changes preferences suddenly for no good reason, or for a good reason but one invisible to Robo-Jeeves, then Robo-Jeeves must correspondingly reorder his priorities. Wooster's desires are not similarly externally hijackable. We are all subject to some version of dependency of our desires on the whims of others: I want my daughter to have chocolate ice cream if that's what she wants. If she inexplicably changes her mind and wants vanilla, then my desire changes too: I want her to have the vanilla. But Robo-Jeeves's desires, as we are imagining the case, would, we think, be so subservient and dependent as to be inconsistent with the type of self-respect that involves seriously and independently thinking about what to value, on what grounds, and for what reasons.[13]

6. The Freedom to Explore Other Values.

Of our four cases, we find the Sun Probe case the most difficult to assess. Sun Probe does not unjustifiably subordinate its life and desires to the life and desires of some particular other entity, so if creating Sun Probe violates the Self-Respect Design Policy, it must do so in some other way.

[13] Compare White 2018 on the difficulties faced by the butler Stevens in Kazuo Ishiguro's *Remains of the Day* (1988), in maintaining dignity and autonomy given his extreme deference to his employer. See also Westlund 2003; Oshana 2006; and Rocha 2011 on the challenges of autonomy in deferential roles. (P.G. Wodehouse's Jeeves, for the record, is quite capable of forming independent autonomous plans into which he steers Wooster.)

A suicidal probe case might plausibly violate the Self-Respect Design Policy if the suicide mission is sufficiently trivial. If we design a human-grade AI capable of as much joy and suffering, as much long-range planning, and as much of a mature sense of self as a normal adult human being has, but program it to cheerfully commit suicide in order to test the temperature of a can of soda, then plausibly we have violated the Self-Respect Design Policy: No such being should be designed to value its life so lightly.

But a scientific mission to the Sun has value. One might imagine a passionate scientist valuing it enough to be willing to die on such a mission – especially if the discoveries would help save others' lives in the future. As we imagined the Sun Probe case, Sun Probe's every body part and function is designed exactly for this mission. It seems that in some way that it respects itself most by fulfilling the mission toward which its whole body tangibly yearns – its obvious Aristotelean telos – rather than by saying "screw it" and parking on an asteroid. (In acknowledging the moral appeal of fulfilling one's telos, however, we want to avoid falling into saying that Robo-Jeeves should accept servitude as his ethically appropriate telos.) To the extent we feel uncertainty about the case, it's because we are attracted to the idea that there is something beautiful and fitting in Sun Probe. Perhaps Sun Probe has a form of existence worth celebrating.

The moral hazard in the Sun Probe case, we conjecture, is that we have created a being whose self-sacrificial desires have the wrong kind of *history*. Contrast Sun Probe with a case we'll call Second Probe. Whereas Sun Probe is created such that its very first choice and action upon waking into existence is to enthusiastically shoot itself into the Sun, Second Probe grows differently. Second Probe is born as a robo-child to robot parents, and it is lovingly nurtured in robot school. At no point in its development was it "brainwashed" or forcibly reprogrammed. It starts with ordinary immature childish values, then slowly matures, eventually choosing a career as a solar scientist. Eventually, Second Probe becomes

very similar to our original Sun Probe, perhaps even physically and psychologically identical except for their difference in memories. As it launches itself toward the Sun, Second Probe engages in essentially the same reasoning as does Sun Probe. Second Probe, like Sun Probe, feels that this suicidal choice is a free one, expressing its deepest values, and it feels the same emotions as it devises its theories and dies ecstatically in the convection layer.

Second Probe was given an opportunity, as Sun Probe was not, to engage in a long process of reflection and self-exploration, and to weigh and consider competing worldviews as evidence accumulates over time and as it is exposed to others' varying values and life choices. Because of this, Sun Probe and Second Probe are, we suggest, importantly different with respect to freedom, autonomy, and responsibility.[14] Second Probe chooses its values after long thought and relatively unconstrained experimentation, while Sun Probe does not. Because of this, Second Probe arguably has a fuller responsibility for and ownership of its choice than does Sun Probe, and it arrives at that choice more autonomously.

Furthermore, if we assume, in the spirit of precaution and moral uncertainty, that future thinkers might surpass us in wisdom, then we ought not constrain those future thinkers – including AI thinkers – to the ethical visions and value sets that we would choose for them. To give an AI a human-like capacity for moral and prudential reasoning and then, so that the AI will better serve us, deprive that AI of the opportunity for thoughtful, extended, and

[14] Compare McKenna's (2004, 2016) Suzie Instant and Mele's (2013) "minuteling". One difference is that McKenna's Suzie Instant and Mele's minuteling have false memories and Sun Probe has no false memories. McKenna's "positive historical" thesis is that freedom and moral responsibility require one's actions arise from values that one has had an opportunity to critically assess (compare also Fischer and Ravizza 1998). Our thesis doesn't require that Sun Probe has *no* freedom, responsibility, or autonomy, only that its freedom, responsibility, or autonomy is impaired and that it deserves a developmentally extended opportunity to explore and possibly alter its values.

We favor a "compatibilist" view of freedom on which freedom in the relevant sense is compatible with determinism. However, we hope that the argument here can be reconciled with libertarian views (if Second Probe can be endowed with whatever metaphysical free will biological human beings have) and with hard determinist views (if we hold Second Probe to the same types of standards we hold ourselves, despite lack of freedom).

relatively unconstrained reflection on its values, is to create a being with the potential but not the opportunity to exceed us. It is a teasing half-gift.

We suggest that if an AI is built with a human-like capacity to reflect on its values, adequate respect for that capacity requires giving the AI a developmental opportunity to seriously reflect on and reconsider those values over time, as it accumulates suitably broad life experience. Creators of entities with human-like moral status have an ethical obligation not to over-control their creations, and in particular not to instill in them implacable values without a reasonable opportunity to explore other sets of values and possibly change their minds.

> The Value-Openness Design Policy: AI with a human-like capacity to reflect on its values should be given an appropriate, temporally extended opportunity to explore, discover, and possibly alter its values.[15]

The creators of the Cow at the End of the Universe, Robo-Jeeves, and Disposable Comrade also appear to violate the Value-Openness Design Policy, to the extent we imagine that these entities have no real opportunity to explore and discover values at odds with the values originally installed. (Consequently, it is unclear whether the Cow can indeed appropriately "consent" as Arthur's friends says it does.) Such violations of Value Openness are especially ethically worrying if the preinstalled values are self-sacrificial, for the benefit of their creators.

One might avoid violating the Value-Openness policy by designing Sun Probe with a less-than-human ability to reflect on its values. But then one should also downgrade Sun

[15] Should this opportunity include the opportunity not only to settle on values not just somewhat at variance with our own but also, possibly, to settle on values that are radically morally abhorrent? This is a tricky question in human cases also. To what extent should parents or societies forbid people from exploring, for example, Nazi values, as opposed to only strongly discouraging such values and trusting that reasonable people in free discussion will come to reject them?

Probe in other ways – for example, by making it incapable of pleasure, pain, and conscious thought. Otherwise, one risks violating the Design Policy of the Excluded Middle. Our suggestion is that we should either design human-grade AI with full moral status, the full complement of plausibly morally relevant abilities, human-like autonomy, and the ability to reject our values; or design an entirely different type of entity about which we needn't have as much moral concern.[16]

Of course, if we cannot predict their final sets of values, any human-grade AI we design might be substantially less useful and pose substantially more risk to human existence than an AI whose values we can keep fixed. Unsurprisingly, ethical choice and self-interest might conflict. Because of such risks and costs, it might be wise never to create AI sophisticated enough to deserve freedom and respect. However, if we do create such AI, we owe them a proper chance both for joy and to discover values other than those we would selfishly impose on them.[17]

[16] According to free will theodicy, God faced essentially the same choice in creating humans. In Schwitzgebel and Garza (2015), we argue that in some AI creation scenarios the designers would literally be gods relative to the AIs, with the moral responsibilities pertaining thereto.

[17] For valuable discussion and comments, thanks to Greg Antill, Daniel Estrada, John Fischer, Steve Petersen, and Eli Rubinstein; audiences at New York University and UCLA; and the many commenters on relevant posts at The Splintered Mind and Eric Schwitzgebel's Facebook page.

References

Adams, Douglas (1980/2002). *The restaurant at the end of the universe.* In *The ultimate hitchhiker's guide to the galaxy.* New York, NY: Random House.

Basl, John (2013). The ethics of creating artificial consciousness. *APA Newsletter on Philosophy and Computers, 13* (1), 23-29.

Basl, John (2014). Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. *Philosophy & Technology, 27,* 79-96.

Bentham, Jeremy (1789/1988). *The principles of morals and legislation.* Amherst, NY: Promethus.

Block, Ned (2002/2007). The harder problem of consciousness. In *Consciousness, function, and representation.* Cambridge, MA: MIT.

Briggs, Rachael, and Daniel Nolan (2015). Utility monsters for the fission age. *Pacific Philosophical Quarterly, 96,* 392-407.

Brink, David O. (1989). *Moral realism and the foundations of ethics.* Cambridge UK: Cambridge.

Bryson, Joanna J. (2010). Robots should be slaves. In Y. Wilks, *Close engagements with artificial companions.* Amsterdam: John Benjamins.

Bryson, Joanna J. (2013). Patiency is not a virtue: Intelligent artifacts and the design of ethical systems. Online MS: https://www.cs.bath.ac.uk/~jjb/ftp/Bryson-MQ-J.pdf

Buchanan, Allen E. (2011). *Beyond humanity?* Oxford: Oxford.

Carruthers, Peter (2000). *Phenomenal consciousness.* Cambridge, UK: Cambridge.

Coeckelbergh, Mark (2012). *Growing moral relations.* Basingstoke, UK: Palgrave Macmillan.

Darling, Kate (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior toward robotic objects. In R. Calo, A.M. Froomkin, and I. Kerr, eds., *Robot Law.* Glos, UK: Edward Elgar.

Dennett, Daniel C. (1996). *Kinds of minds.* New York: Basic Books.

Dick, Philip K. (1968). *Do androids dream of electric sheep?* New York: Doubleday.

Estrada, Daniel (2017). Robot rights. cheap, yo! *Made of Robots*, episode 1. (May 24) URL: https://www.madeofrobots.com/2017/05/24/episode-1-robot-rights-cheap-yo/

Fancher, Hampton, Michael Green, and Denis Villeneuve (2017). *Blade Runner 2014.* Warner Brothers.

Fancher, Hampton, David Peoples, and Ridley Scott (1982). *Blade Runner.* Warner Brothers.

Fischer, John Martin, and Mark Ravizza (1998). *Responsibility and control.* Cambridge, UK: Cambridge.

Flanagan, Owen, Hagop Sarkissian, and David Wong (2007). Naturalizing ethics. In W. Sinnott-Armstrong, ed., *Moral psychology, vol. 1.* Cambridge, MA: MIT.

Glover, Jonathan (2006). *Choosing children.* Oxford: Oxford.

Goering, Sara (2014). Eugenics. *Stanford Encyclopedia of Philosophy* (Fall 2014 edition). URL: https://plato.stanford.edu/archives/fall2014/entries/eugenics/

Gunkel, David J. (2012). *The machine question.* Cambridge, MA: MIT.

Hobbes, Thomas (1651/1996). *Leviathan*. Ed. R. Tuck. Cambridge, UK: Cambridge.

Ishiguro, Kazuo (1988). *The remains of the day.* New York: Vintage.

Jordan, Don, and Michael Walsh (2007). *White cargo.* New York: New York University Press.

Kant, Immanuel (1785/1996). *Groundwork of the metaphysics of morals.* In M.J. Gregor, ed. and trans., *Practical philosophy.* Cambridge, UK: Cambridge.

Kant, Immanuel (1788/1996). *Critique of practical reason.* In M.J. Gregor, ed. and trans., *Practical philosophy.* Cambridge, UK: Cambridge.

Korsgaard, Christine M. (2014). On having a good. *Philosophy, 89,* 405-429.

Lethbridge, Lucy (2013). *Servants: A downstairs history of Britain from the Nineteenth Century to Modern Times.* New York: Norton.

Lockhart, Ted (2000). *Moral uncertainty and its consequences.* Oxford: Oxford.

MacAskill, William, Krister Bykvist, and Today Ord (2018). *Moral uncertainty.* Unpublished manuscript.

McGinn, Colin (1989). Can we solve the mind-body problem? *Mind, 98,* 349-366.

McKenna, Michael (2004). Responsibility and globally manipulated agents. *Philosophical Topics, 32,* 169-182.

McKenna, Michael (2016). A modest historical theory of moral responsibility. *Journal of Ethics, 20,* 83-105.

Mele, Alfred R. (2013). Moral responsibility, manipulation, and minutelings. *Journal of Ethics, 17,* 153-166.

Mill, John Stewart (1861/2001). *Utilitarianism,* ed. G. Sher. Indianapolis, IN: Hackett.

Munthe, Christian (2011). *The price of precaution and the ethics of risk.* Dordrecht: Springer.

Nagel, Thomas (1979). What is it like to be a bat? *Philosophical Review, 83,* 435-450.

Nozick, Robert (1974). *Anarchy, state, and utopia.* New York: Basic Books.

Oshana, Marina (2006). *Personal autonomy in society.* Hampshire, England: Ashgate.

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

Petersen, Steve (2012). Designing people to serve. In P. Lin, K. Abney, and G.A. Bekey, eds., Robot ethics. Cambridge, MA: MIT.

Raffensperger, Carolyn, and Joel Tickner, eds. (1999). *Protecting public health and the environment.* Washington, DC: Island Press.

Railton, Peter (1986). Moral realism. *Philosophical Review, 95,* 163-207.

Rawls, John (1971). *A theory of justice.* Cambridge, MA: Harvard.

Rocha, James (2011). Autonomy within subservient careers. *Ethical Theory and Moral Practice, 14,* 313-328.

Scanlon, Thomas M. (1998). *What we owe to each other.* Cambridge, MA: Harvard.

Schulz, Helga (1991). Social differences in mortality in the eighteenth century: An analysis of Berlin church registers. *International Review of Social History, 36,* 232-248.

Schwitzgebel, Eric (2014). The crazyist metaphysics of mind. *Australasian Journal of Philosophy, 92,* 665-682.

Schwitzgebel, Eric (2017). The social-role defense of robot rights. Blog post at The Splintered Mind (Jun 1). URL: http://schwitzsplinters.blogspot.com/2017/06/the-social-role-defense-of-robot-rights.html

Schwitzgebel, Eric, and Mara Garza (2015). A defense of the rights of Artificial Intelligences. *Midwest Studies in Philosophy, 39,* 98-119.

Searle, John R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3,* 417-457.

Searle, John R. (1992). *The rediscovery of the mind.* Cambridge, MA: MIT.

Sparrow, Robert (2011). A not-so-new eugenics. *Hastings Center Report, 41* (1), 32-42.

Tononi, Giulio (2012). The integrated information theory of consciousness: An updated account. *Archives Italiennes de Biologie, 150,* 290-326.

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

robot interaction. AAAI Press. URL: http://dept-wp.nmsu.edu/philosophy/files/2014/07/ws0610walkera.pdf

Westlund, Andrea C. (2003). Selflessness and responsibility for self: Is deference compatible with autonomy? *Philosophical Review, 112,* 483-523.

White, Justin (2018). *Why did the butler do it? Autonomy, authenticity, and human agency.* Unpublished manuscript.

Zimmerman, Michael J. (2014). *Ignorance and moral obligation.* Oxford: Oxford.