A Defense of the Rights of Artificial Intelligences

Eric Schwitzgebel and Mara [official surname still to be decided]

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

eschwitz at domain: ucr.edu

September 1, 2015

A Defense of the Rights of Artificial Intelligences

Abstract:

There are possible artificially intelligent beings who do not differ in any morally relevant respect from human beings. Such possible beings would deserve moral consideration similar to that of human beings. Our duties to them would not be appreciably reduced by the fact that they are non-human, nor by the fact that they owe their existence to us. Indeed, if they owe their existence to us, we would likely have additional moral obligations to them that we don't ordinarily owe to human strangers – obligations similar to those of parent to child or god to creature. Given our moral obligations to such AIs, two principles for ethical AI design recommend themselves: (1) design AIs that tend to provoke reactions from users that accurately reflect the AIs' real moral status, and (2) avoid designing AIs whose moral status is unclear. Since human moral intuition and moral theory evolved and developed in contexts without AI, those intuitions and theories might break down or become destabilized when confronted with the wide range of weird minds that AI design might make possible.

Word count: approx 10,000 (including notes and references), plus one figure

A Defense of the Rights of Artificial Intelligences

> "I am thy creature, and I will be even mild and docile to my natural lord and king if thou wilt also perform thy part, the which thou owest me. Oh, Frankenstein, be not equitable to every other and trample upon me alone, to whom thy justice, and even thy clemency and affection, is most due. Remember that I am thy creature; I ought to be thy Adam…" (Frankenstein's monster to his creator, Victor Frankenstein, in Shelley 1818/1965, p. 95).

We might someday create entities with *human-grade artificial intelligence.* Human-grade artificial intelligence – hereafter, just *AI*, leaving *human-grade* implicit – in our intended sense of the term, requires both intellectual and emotional similarity to human beings, that is, both human-like general theoretical and practical reasoning and a human-like capacity for joy and suffering. Science fiction authors, artificial intelligence researchers, and the (relatively few) academic philosophers who have written on the topic tend to think that such AIs would deserve moral consideration, or "rights", similar to the moral consideration we owe to human beings.[1]

Below we provide a positive argument for AI rights, defend AI rights against four objections, recommend two principles of ethical AI design, and draw two further conclusions: first, that we would probably owe *more* moral consideration to human-grade artificial

[1] Classic examples in science fiction include Isaac Asimov's robot stories (esp. 1954/1962, 1982) and *Star Trek: The Next Generation,* especially the episode "The Measure of a Man" (Snodgrass and Scheerer 1989). Academic treatments include Basl 2013; Bryson 2013; Bostrom and Yudkowsky 2014; Gunkel and Bryson, eds., 2014. See also Coeckelbergh 2012 and Gunkel 2012 for critical treatments of the question as typically posed.

We use the term "rights" here to refer broadly to moral considerability, moral patiency, or the capacity to make legitimate ethical claims upon us.

intelligences than we owe to human strangers, and second, that the development of AI might destabilize ethics as an intellectual enterprise.

1. The No-Relevant-Difference Argument.

Our main argument for AI rights is:

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 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.

A weaker version of this argument, which we will not focus on here, substitutes “mammals” or some other term from the animal rights literature for “human beings” in Premise 2 and the Conclusion.[2]

The argument is valid: The conclusion plainly follows from the premises. We hope that most readers will also find both premises plausible and thus accept the argument as sound. To deny Premise 1 renders ethics implausibly arbitrary. All four of the objections we consider below are challenges to Premise 2.

The argument is intentionally abstract. It does not commit to any one account of what constitutes a “relevant” difference. We believe that the argument can succeed on a variety of

[2] On sub-human AI and animal rights, see especially Basl 2013, 2014.

plausible accounts. On a broadly Kantian view, rational capacities would be the most relevant. On a broadly utilitarian view, capacity for pain and pleasure would be most relevant. Also plausible are nuanced or mixed accounts or accounts that require entering certain types of social relationships. In Section 2, we will argue that only psychological and social properties should be considered directly relevant to moral status.

The argument's conclusion is intentionally weak. There are *possible* AIs who deserve a degree of moral consideration similar to that of human beings. This weakness avoids burdening our argument with technological optimism or commitment to any particular type of AI architecture. The argument leaves room for strengthening. For example, an enthusiast for strong "classical" versions of AI could strengthen Premise 2 to "There are possible AIs designed along classical lines who…" and similarly strengthen the Conclusion. Someone who thought that human beings might differ in no relevant respect from silicon-based entities, or from distributed computational networks, or from beings who live entirely in simulated worlds (Egan 1997, Bostrom 2003), could also strengthen Premise 2 and the Conclusion accordingly.

One might thus regard the No-Relevant-Difference Argument as a template that permits at least two dimensions of further specification: specification of what qualifies as a relevant difference and specification of what types of AI possibly lack any relevant difference.

The No-Relevant-Difference Argument is humanocentric in that it takes humanity as a standard. This is desirable because we assume it is less contentious among our interlocutors that human beings have rights (at least "normal" human beings, setting aside what is sometimes called the problem of "marginal cases") than it is that rights have any specific basis such as rationality or capacity for pleasure. If a broader moral community someday emerges, it might be desirable to recast the No-Relevant-Difference Argument in correspondingly broader terms.

The argument suggests a test of moral status, which we will call the *Difference Test*. The Difference Test is a type of moral argumentative challenge. If you are going to regard one type of entity as deserving greater moral consideration than another, you ought to be able to point to a relevant difference between those entities that justifies that differential treatment. Inability to provide such a justification opens one up to suspicions of chauvinism or bias.

The Difference Test has general appeal in the fight against chauvinism and bias among human beings. Human egalitarianism gains support from the idea that skin color, ancestry, place of birth, gender, sexual orientation, and wealth cannot properly ground differences in a person's moral status. The No-Relevant-Difference Argument aims to extend this egalitarian approach to AIs.

2. The Psycho-Social View of Moral Status, and Liberalism about Embodiment and Architecture.

It shouldn't matter to one's moral status what kind of body one has, except insofar as one's body influences one's psychological and social properties. Similarly, it shouldn't matter to one's moral status what kind of underlying architecture one has, except insofar as underlying architecture influences one's psychological and social properties. Only psychological and social properties are directly relevant to moral architecture – or so we propose. This is one way to narrow what qualifies as a "relevant" difference in the sense of Premise 1 of the No-Relevant-Difference Argument. Call this the *psycho-social view of moral status*.[3]

[3] Compare Bostrom and Yudkowsky's (2014) Principle of Substrate Non-Discrimination and Principle of Ontogeny Non-Discrimination. We embrace the former but possibly not the latter (depending on how it is interpreted), as should be clear from our discussion of social properties and especially our special duties to our creations.

By *psychological* we mean to include both functional or cognitive properties, such as the ability to reason mathematically, and phenomenological or conscious properties, such as the disposition to experience pain when damaged, regardless of whether the phenomenological or conscious reduces to the functional or cognitive. By *social* we mean to include facts about social relationships, independently of whether they are psychologically appreciated by either or both of the related parties – for example, the relationship of parenthood or citizenship or membership in a particular community. Others' *opinions* of one's moral status are a possibly relevant dimension of the social (though worryingly so), but we do not include an entity's *actual* moral status in the "social" lest the psycho-social view be trivially true.

A purely psychological view would ground moral status entirely in the psychological properties of the entity whose status is being appraised. Our view is *not* restricted in this way, instead allowing that social relationships might be directly relevant to moral status. Neither do we intend this view to be temporally restricted or restricted to actually manifested properties. Both past and future psychological and social properties, both actual and counterfactual, might be directly relevant to moral status (as in the case of a fetus or a brain-injured person, or in the case of an unremembered interaction, or in a case of "she would have suffered if…"). We leave open which specific psychological and social properties are relevant to moral status.

Here are two reasons to favor the psycho-social view of moral status.

(1.) All of the well-known modern secular accounts of moral status in philosophy ground moral status only in psychological and social properties, such as capacity for rational thought, pleasure, pain, and social relationships. No influential modern secular account is plausibly read as committed to a principle whereby two beings who differ not at all psychologically or socially would differ in their moral status. Nothing in Hobbes, Locke, Hume, Mill, Rawls, or Singer, for example, plausibly gives the reader resources to root moral status outside the psychological and social.

However, some older or religious accounts might have resources to ground a difference in moral status outside the psychological and social. An Aristotelian *might* suggest that AIs would have a different *telos* or defining purpose than human beings. However, it's not clear that an Aristotelian must think this; nor do we think such a principle, interpreted in such a way, would be very attractive from a modern perspective, unless directly relevant psychological or social differences accompanied the difference in telos. (For a related point, see the Objection from Existential Debt in Section 6.) Similarly, a theist *might* suggest that God somehow imbues human beings with higher moral status than AIs, even if they are psychologically and socially identical. We find this claim difficult to assess, but we're inclined to think that a deity who distributed moral status unequally in this way would be morally deficient. (For a related point, see Section 10 on our responsibility for our creations.)

(2.) If one considers a wide range of cases in vivid detail, it appears to be intuitively clear – though see our critiques of moral intuition in Sections 9 and 13 – that what should matter to moral status are only psychological and social properties. This is, we think, one of the great lessons of science fiction. Science fictional portrayals of robots in Asimov and *Star Trek*, of simulated beings in Greg Egan and the "White Christmas" episode of *Black Mirror*, of sentient spaceships in the works of Iain Banks and Aliette de Bodard, of group minds and ugly "spiders" in Vernor Vinge, uniformly invite the thoughtful reader or viewer to a liberal attitude toward embodiment: What matters is how such beings think, what they feel, and how they interact with others.[4] Whether they are silicon or meat, humanoid or ship-shaped, sim or ghost, is irrelevant except insofar as it influences their psychological and social properties.

To be clear: Embodiment or architecture might matter a lot to moral status. But if they do, we propose that it's only via their influence on psychological and social properties.

3. "Artificial" and a Slippery Slope Argument for AI Rights.

It's not clear what it means, in general, for something to be "artificial", nor what the term "artificial" means specifically in the context of "artificial intelligence". For our purposes, "artificial" should *not* be read as implying "programmed" or "made of silicon". To read it that way commits to too narrow a view of the possible future of AI. AI might leave silicon behind as it previously left vacuum tubes behind, perhaps in favor of nanotech carbon components or patterns of interference in reflected light. An even now, what we normally think of as non-human grade AI can be created other than by explicit programming, for example through evolutionary algorithms or training up connectionist networks.

Borderline cases abound. Are killer bees natural or artificial? How about genetically engineered viruses? If we released self-replicating nanotech and it began to evolve in the wild, at what point, if ever, would it qualify as natural? If human beings gain control over their bodily development, incorporating increasingly many manufactured and/or genetically-tweaked parts, would they cross from the natural to the artificial? How about babies or brain cells grown in vats, shaped into cognitive structures increasingly unlike those of people as they existed in 2015? Might some beings who are otherwise socially and psychologically indistinguishable from

[4] See Asimov 1954/1962, 1982; Snodgrass and Scheerer 1989; Egan 1994, 1997; Brooker and Tibbets 2014; Banks' "Culture" series from 1987 to 2012; de Bodard, e.g., 2011, 2013; Vinge 1992, 1999, 2011.

natural human beings lack full moral status because of some fact about their design history – a fact perhaps unknowable to them or to anyone with whom they are likely to interact?

Consider the film *Blade Runner* and the Philip K. Dick novel on which it was loosely based (Fancher, Peoples, and Scott 1982; Dick 1968). In that world, "andys" or "replicants" are manufactured as adults with fictional memories, and they survive for several years. Despite this fact about their manufacture, they are biologically almost indistinguishable from human beings, except by subtle tests, and sometimes neither the andys/replicants themselves nor their acquaintances know that they are not normal human beings. Nevertheless, because they are a product of the increasingly advanced development of biological-mimicry AI, they are viewed as entities with lesser rights. Such beings would be in some important sense artificial; but since they are conceptualized as having almost normal human brains, it's unclear how well our conceptions of "artificial intelligence" apply to them.

One nice feature of our view is that none of this matters. "Artificial" needn't be clearly distinguished from "natural". Once all the psychological and social properties are clarified, you're done, as far as determining what matters to moral status.

A person's moral status is not reduced by having an artificial limb. Likewise, it seems plausible to say that a person's moral status would not be reduced by replacing a damaged part of her brain with an artificial part that contributes identically to her psychology and does not affect relevant social relationships, *if* artificial parts can be built or grown that contribute identically to one's psychology. This suggests a second argument for AI rights:

The Slippery Slope Argument for AI rights:

Premise 1. Substituting a small artificial component into an entity with rights, if that component contributes identically to the entity's psychology and does not affect relevant social relationships, does not affect that entity's rights.

Premise 2. The process described in Premise 1 could possibly be iterated in a way that transforms a natural human being with rights into a wholly artificial being with the same rights.

Conclusion. Therefore, it is possible to create an artificial being with the same rights as those of a natural human being.

This argument assumes that replacement by artificial components is possible while preserving all relevant psychological properties, which would include the property of having conscious experience.[5] However, some might argue that consciousness, or some other relevant psychological property, could not in fact be preserved while replacing a natural brain with an artificial one – which brings us to the first of four objections to AI rights.

4. The Objection from Psychological Difference.

We have asserted that there are possible AIs who have no relevant psychological differences from ordinary human beings. One objection is that this claim is too far-fetched – that all possible, or at least all realistically possible, artificial entities would differ psychologically from human beings in some respect relevant to moral status. The existing literature suggests three candidate differences of plausibly sufficient magnitude to justifying denying full rights to

[5] Our argument is thus importantly different from superficially similar arguments in Cuda 1985 and Chalmers 1996, which assume the possibility of replacement parts that are *functionally* identical but which do not assume that consciousness is preserved. Rather, the preservation of consciousness is what Cuda and Chalmers are trying to establish as the argumentative *conclusion*, with the help of some further premises, such as (in Chalmers) introspective reliability. We find the Cuda-Chalmers argument attractive but we are not committed to it.

artificial entities. Adapting a suggestion from Searle (1980), artificial entities might necessarily lack *consciousness*. Adapting a suggestion from Lovelace (1843), artificial entities might necessarily lack *free will*. Adapting a suggestion from Penrose (1999), artificial entities might necessarily be incapable of *insight*.

We believe it would be very difficult to establish such a conclusion about artificial entities in general. Even Searle, perhaps the most famous critic of strong, classical AI, says that he sees no reason in principle why a machine couldn't understand English or Chinese, which on his view would require consciousness; and he allows that artificial intelligence research might in the future proceed very differently, in a way that avoids his concerns about classical AI research in terms of formal symbol manipulation (his "Many Mansions" discussion et seq.). Lovelace confines her doubts to Babbage's analytic engine. Penrose allows that we might someday discover in detail what endows us with consciousness that can transcend purely algorithmic thinking, and then create such consciousness artificially (1999, p. 416). Searle and Penrose, at least, seem to allow that technology might well be capable of creating an artificially designed, grown, or selected entity, with all the complexity, creativity, and consciousness of a human being. For this reason, we have described the objections above as "inspired" by them. They themselves are more cautious.[6]

[6] We have also simplified the presentation of the positions "inspired by" Searle, Lovelace, and Penrose in a way that the authors might not fully approve. Lovelace, for example, doesn't use the word "freedom" or the phrase "free will" – more characteristic is "the machine is not a thinking being, but simply an automaton which acts according to the laws imposed on it" (p. 675); also, the machine "follows" rather than "originates" (p. 722). Searle emphasizes meaning, understanding, and intentionality in a way not emphasized in this brief description. Penrose's position does not entirely contrast with Searle's on the issue of consciousness, since he suggests that an algorithmic machine or automaton would lack consciousness, and conversely Searle suggests that consciousness is necessary for "flexibility and creativity" (1992, p. 108) in a way that might fit with Penrose's non-algorithmic insight and perhaps the idea implicit in

A *certain* way of designing artificial intelligence – a 19th and 20th century way – might not, if Searle, Lovelace, and Penrose are right, achieve certain aspects of human psychology that are important to moral status. (We take no stand here on whether this is actually so.) But no *general* argument has been offered against the moral status of all possible artificial entities. Connectionist training and evolutionary algorithms are already moving AI away from the classical vision of the 1960s, and future developments might be even more revolutionary, including perhaps artificially grown biological or semi-biological systems, chaotic systems, evolved systems, quantum systems, and artificial brains.

The No-Relevant-Difference Argument commits only to a very modest claim: There are possible AIs who are not relevantly different. To argue against this possibility on broadly Searle-Lovelace-Penrose grounds will require going considerably farther than they themselves do. Pending further argument, we see no reason to think that *all* artificial entities must suffer from radical psychological deficiency. Perhaps the idea that AIs must necessarily lack consciousness, free will, or insight is attractive partly due to a culturally ingrained picture of AIs as deterministic, clockwork machines very different from us spontaneous, unpredictable humans. But we see no reason to think that human cognition is any less mechanical or more spontaneous than that of some possible machines. Why should we deny the possibility of realizing a psychological capacity on any of a variety of artificial material substrates?

Maybe consciousness, free will, or insight requires an immaterial soul? Here we follow Turing's (1950) response to a similar concern.[7] If naturalism is true, then whatever process generates a soul in human beings might also generate a soul in an artificial being. Even if soul-

Lovelace that "thinking" requires more than acting according to imposed laws. The success of our reply does not, we think, depend on philosophical differences at this level of detail.

[7] See Estrada 2014 for extensive discussion of Lovelace's objection and Turing's reply.

installation requires the miraculous touch of God, we're inclined to think that a god who cares enough about human consciousness, freedom, and insight to imbue us with souls might imbue the right sort of artificial entity with one also.

The arguments of Searle, Lovelace, and Penrose do raise concerns about the *detection* of certain psychological properties in artificial systems – an issue we will address in Section 8.

5. The Objection from Duplicability.

AIs might not deserve equal moral concern because they do not have fragile, unique lives of the sort that human beings have. It might be possible to duplicate AIs or back them up so that if one is harmed or destroyed, others can take its place, perhaps with the same memories or seeming-memories – perhaps even ignorant that any re-creation and replacement has occurred. Harming or killing an AI might therefore lack the gravity of harming or killing a human being. Call this the Objection from Duplicability.[8]

Our reply is simple: It is possible to create AIs as unique and fragile as human beings. If so, then the No-Relevant-Difference Argument survives the objection.

Although we think this reply is adequate to save the No-Relevant-Difference Argument as formulated, it's also worth considering the effects of duplicability on the moral status of AIs that are not unique and fragile. Duplicability and fragility probably would influence our moral obligations to AIs. If one being splits into five virtually identical beings, each with full memories of their previous lives before the split, and then after ten minutes of separate existence one of those beings is killed, it seems less of a tragedy than if a single unique, non-splitting being

[8] This objection is inspired by Peter Hankins' (2015) argument that duplicability creates problems for holding robots criminally responsible. (Hankins also suggests that programmed robots have "no choice" – a concern more in the spirit of the previous section.)

is killed. This might be relevant to the allotment of risky tasks, especially if splitting can be planned in advance. On the other hand, the possibly lower fragility of some possible AIs might make their death *more* of a tragedy. Suppose a natural eighty-year-old woman with ten more years of expected life has an artificial twin similar in all relevant respects except that the twin has a thousand more years of expected life. Arguably, it's more of a tragedy for the twin to be destroyed than for the natural woman to be destroyed. Possibly it's even more tragic if the AI had the potential to split into a thousand separate AIs each with a thousand years of expected life – perhaps en route to colonize a star – who will now never exist.

We're unsure how these issues ought to play out. However we see here no across-the-board reason to hold AI lives in less esteem generally.

6. The Objection from Otherness.

The state of nature is a "Warre, where every man is Enemy to every other man" – says Hobbes (1651/1996, p. 89 [62]) – until some contract is made by which we agree to submit to an authority for the mutual good. Perhaps such a state of Warre is the "Naturall Condition" between species: We owe nothing to alligators and they owe nothing to us.

For a moment, let's set aside any purely psychological grounds for moral consideration. A Hobbesian might say that if space aliens were to visit, they would be not at all wrong to kill us for their benefit, nor vice versa, until the right sort of interaction created a social contract. Or we might think in terms of circles of concern: We owe the greatest obligation to family, less to neighbors, still less to fellow citizens, still less to distant foreigners, maybe nothing at all outside our species. Someone might think that AIs necessarily stand outside of our social contracts or the appropriate circles of concern, and thus there's no reason to give them moral consideration.

Extreme versions of these views are, we think, obviously morally odious. Torturing or killing a human-grade AI or a conscious, self-aware, intelligent alien, without very compelling reason, is not morally excused by the being's not belonging to our species or social group. Vividly imagining such cases in science fiction scenarios draws out the clear intuition that such behavior would be grossly wrong.

One might hold that species per se matters at least somewhat, and thus that there will always be a relevant relational difference between AIs and "us" human beings, in light of which AIs deserve less moral consideration from us than do our fellow human beings.[9] However, we suggest that this is to wrongly fetishize species membership. Consider a hypothetical case in which AI has advanced to the point where artificial entities can be seamlessly incorporated into society without the AIs themselves, or their friends, realizing their artificiality. Maybe some members of society have [choose-your-favorite-technology] brains while others have very similarly functioning natural human brains. Or maybe some members of society are constructed from raw materials as infants rather than via germ lines that trace back to homo sapiens ancestors. We submit that as long as these artificial or non-homo-sapiens beings have the same psychological properties and social relationships that natural human beings have, it would be a cruel moral mistake to demote them from the circle of full moral concern upon discovery of their different architecture or origin.

Purely biological otherness is irrelevant unless some important psychological or social difference flows from it. And on any reasonable application of a social standard for moral

[9] This is a version of the view Singer labels pejoratively as "speciesism" (1975/2002, 2009). Our view is also compatible with Kagan's (forthcoming) critique of Singer on this issue, since it seems that Kagan's proposed "personism" would not violate the psycho-social view of moral status in the broad sense of Section 2. Perhaps Williams (2006, ch. 13) advocates speciesism per se, though it's not entirely clear.

considerability equivalent to that of human beings, there are possible AIs that would meet that standard.

7. The Objection from Existential Debt.

Suppose you build a fully human-grade intelligent robot. It costs you $1000 to build and $10 per month to maintain. After a couple of years, you decide you'd rather spend the $10 per month on a magazine subscription. Learning of your plan, the robot complains, "Hey, I'm a being as worthy of continued existence as you are! You can't just kill me for the sake of a magazine subscription!"

Suppose you reply: "You ingrate! You owe your very life to me. You should be thankful just for the time I've given you. I owe you nothing. If I choose to spend my money differently, it's my money to spend." The Objection from Existential Debt begins with the thought that artificial intelligence, simply by virtue of being *artificial* (in some appropriately specifiable sense), is made by us, and thus owes its existence to us, and thus can be terminated or subjugated at our pleasure without moral wrongdoing as long as its existence has been overall worthwhile.

Consider this possible argument in defense of eating humanely raised meat. A steer, let's suppose, leads a happy life grazing on lush hills. It wouldn't have existed at all if the rancher hadn't been planning to kill it for meat. Its death for meat is a condition of its existence, and overall its life has been positive; seen as the package deal it is, the rancher's having brought it into existence and then killed it is overall is morally acceptable.[10] A religious person dying young of cancer who doesn't believe in an afterlife might console herself similarly: Overall, she might think, her life has been good, so God has given her nothing to resent. Analogously, the

[10] See DeGrazia 2009 for presentation and criticism of an argument along roughly these lines.

argument might go, you wouldn't have built that robot two years ago had you known you'd be on the hook for $10 per month in perpetuity. Its continuation-at-your-pleasure was a condition of its very existence, so it has nothing to resent.

We're not sure how well this argument works for non-human animals raised for food, but we reject it for human-grade AI. We think the case is closer to this clearly morally odious case:

Ana and Vijay decide to get pregnant and have a child. Their child lives happily for his first eight years. On his ninth birthday, Ana and Vijay decide they would prefer not to pay any further expenses for the child, so that they can purchase a boat instead. No one else can easily be found to care for the child, so they kill him painlessly. But it's okay, they argue! Just like the steer and the robot! They wouldn't have had the child (let's suppose) had they known they'd be on the hook for child-rearing expenses until age eighteen. The child's support-at-their-pleasure was a condition of his existence; otherwise Ana and Vijay would have remained childless. He had eight happy years. He has nothing to resent.

The decision to have a child carries with it a responsibility for the child. It is not a decision to be made lightly and then undone. Although the child in some sense "owes" its existence to Ana and Vijay, that is not a callable debt, to be vacated by ending the child's existence. Our thought is that for an important range of possible AIs, the situation would be similar: If we bring into existence a genuinely conscious human-grade AI, fully capable of joy and suffering, with the full human range of theoretical and practical intelligence and with expectations of future life, we make a moral decision approximately as significant and irrevocable as the decision to have a child.

9. Why We Might Owe *More* to AIs, Part One: Our Responsibility for Their Existence and Properties.

In fact, we're inclined to turn the Existential Debt objection on its head: If we intentionally bring a human-grade AI into existence, we put ourselves into a social relationship that carries responsibility for the AI's welfare. We take upon ourselves the burden of supporting it or at least of sending it out into the world with a fair shot of leading a satisfactory existence. In most realistic AI scenarios, we would probably also have some choice about the features the AI possesses, and thus presumably an obligation to choose a set of features that will not doom it to pointless misery.[11] Similar burdens arise if we do not personally build the AI but rather purchase and launch it, or if we adopt the AI from a previous caretaker.

Some familiar relationships can serve as partial models of the sorts of obligations we have in mind: parent-child, employer-employee, deity-creature. Employer-employee strikes us as likely too weak to capture the degree of obligation in most cases but could apply in an "adoption" case where the AI has independent viability and willingly enters the relationship. Parent-child perhaps comes closest when the AI is created or initially launched by someone without whose support it would not be viable and who contributes substantially to the shaping of the AI's basic features as it grows, though if the AI is capable of mature judgment from birth that creates a disanalogy. Diety-creature might be the best analogy when the AI is subject to a person with profound control over its features and environment. All three analogies suggest a special relationship with obligations that exceed those we normally have to human strangers.

In some cases, the relationship might be *literally* conceivable as the relationship between deity and creature. Consider an AI in a simulated world, a "Sim", over which you have godlike

[11] Analogous issues are central to the ethics of disability and human enhancement, e.g., Glover 2006; Buchanan 2011; Sparrow 2011.

powers. This AI is a conscious part of a computer or other complex artificial device. Its "sensory" input is input from elsewhere in the device, and its actions are outputs back into the remainder of the device, which are then perceived as influencing the environment it senses. Imagine the computer game The Sims, but containing many actually conscious individual AIs. The person running the Sim world might be able to directly adjust an AI's individual psychological parameters, control its environment in ways that seem miraculous to those inside the Sim (introducing disasters, resurrecting dead AIs, etc.), have influence anywhere in Sim space, change the past by going back to a save point, and more – powers that would put Zeus to shame. From the perspective of the AIs inside the Sim, such a being would be a god. If those AIs have a word for "god", the person running the Sim might literally be the referent of that word, literally the creator (or at least launcher) of their world and potential destroyer of it, literally existing outside their spatial manifold, and literally capable of violating the laws that usually govern their world. Given this relationship, we believe that the manager of the Sim would also possess the obligations of a god, including probably the obligation to ensure that the AIs contained within don't suffer needlessly. A burden not to be accepted lightly![12]

Even for AIs embodied in our world rather than in a Sim, we might have considerable, almost godlike control over their psychological parameters. We might, for example, have the opportunity to determine their basic default level of happiness. If so, then we will have a substantial degree of direct responsibility for their joy and suffering. Similarly, we might have the opportunity, by designing them wisely or unwisely, to make them more or less likely to lead lives with meaningful work, fulfilling social relationships, creative and artistic achievement, and

[12] For further reflections on this theme, presented as science fiction, see Schwitzgebel and Bakker 2013; Schwitzgebel 2015b.

other value-making goods. It would be morally odious to approach these design choices cavalierly, with so much at stake. With great power comes great responsibility.[13]

We have argued in terms of individual responsibility for individual AIs, but similar considerations hold for group-level responsibility. A society might institute regulations to ensure happy, flourishing AIs who are not enslaved or abused; or it might fail to institute such regulations. People who knowingly or negligently accept societal policies that harm their society's AIs participate in collective responsibility for that harm.

Artificial beings, *if* psychologically similar to natural human beings in consciousness, creativity, emotionality, self-conception, rationality, fragility, etc., warrant substantial moral consideration in virtue of that fact alone. If we are furthermore *also* responsible for their existence and features, they have a moral claim upon us that human strangers do not ordinarily have to the same degree.

9. Why We Might Owe More to AIs, Part Two: Their Possible Superiority.

Robert Nozick (1974) imagines "utility monsters" who derive enormous pleasure from sacrificing others. We might imagine a being who derives a hundred units of pleasure from each cookie it eats, while normal human beings derive only one unit of pleasure. A simple version of pleasure-maximizing utilitarianism would suggest (implausibly, Nozick thinks) that we should give all our cookies to the monster.

If it is possible to create genuinely joyful experiences in AIs, it will also likely be possible to create AIs who experience substantially more joy than the typical human being. Such AIs might be something like Nozick's utility monsters. If our moral obligation is to maximize

[13] As Uncle Ben wisely advises Spider-Man in the 2002 film (Lee, Ditko, Coepp, and Raimi 2002, slightly modifying a passage in the voice of the narrator in Lee and Ditko 1962).

happiness, we might be obliged to create many such entities, even at substantial cost to ordinary human beings.[14] Adapting an example from Bostrom (2014), we might contemplate converting most of the mass of the solar system into “hedonium” – whatever artificial substrate most efficiently generates feelings of pleasure. We might be morally obliged to destroy ourselves to create a network of bliss machines.

Most philosophers would reject simple pleasure-maximization approaches to ethics. For example, a consequentialist might complicate her account by recognizing individual rights that cannot easily be set aside for the benefit of others. But even with such complications, any ethics that permits inflicting harm on one person to elsewhere create greater happiness, or to prevent greater suffering, invites the possibility of giving greater moral weight to outcomes for possible AIs that are capable of much greater happiness or suffering than ordinary humans.

One might hope to avoid this result by embracing an ethics that emphasizes the value of rationality rather than pleasure and pain, but this invites the possibly unappealing thought that AIs with superior rational capacities might merit greater moral consideration. To avoid this conclusion, one might treat rationality as a threshold concept with human beings already across the highest morally relevant threshold: Equal status for human beings and all creatures with rational capacities similar to or superior to those of human beings. One cookie and one vote for each.

Although such a view avoids utility monster cases, it throws us upon troubling issues of personal identity. Consider, for example, a *fission-fusion monster* – a human-grade AI who can divide and merge at will.[15] How many cookies should it get? October 31st, it is one entity. November 1st it fissions into a million human-grade AIs, each with the memories and values of

[14] Compare also Parfit’s (1984) “Repugnant Conclusion”.
[15] For a related example, see Briggs and Nolan forthcoming.

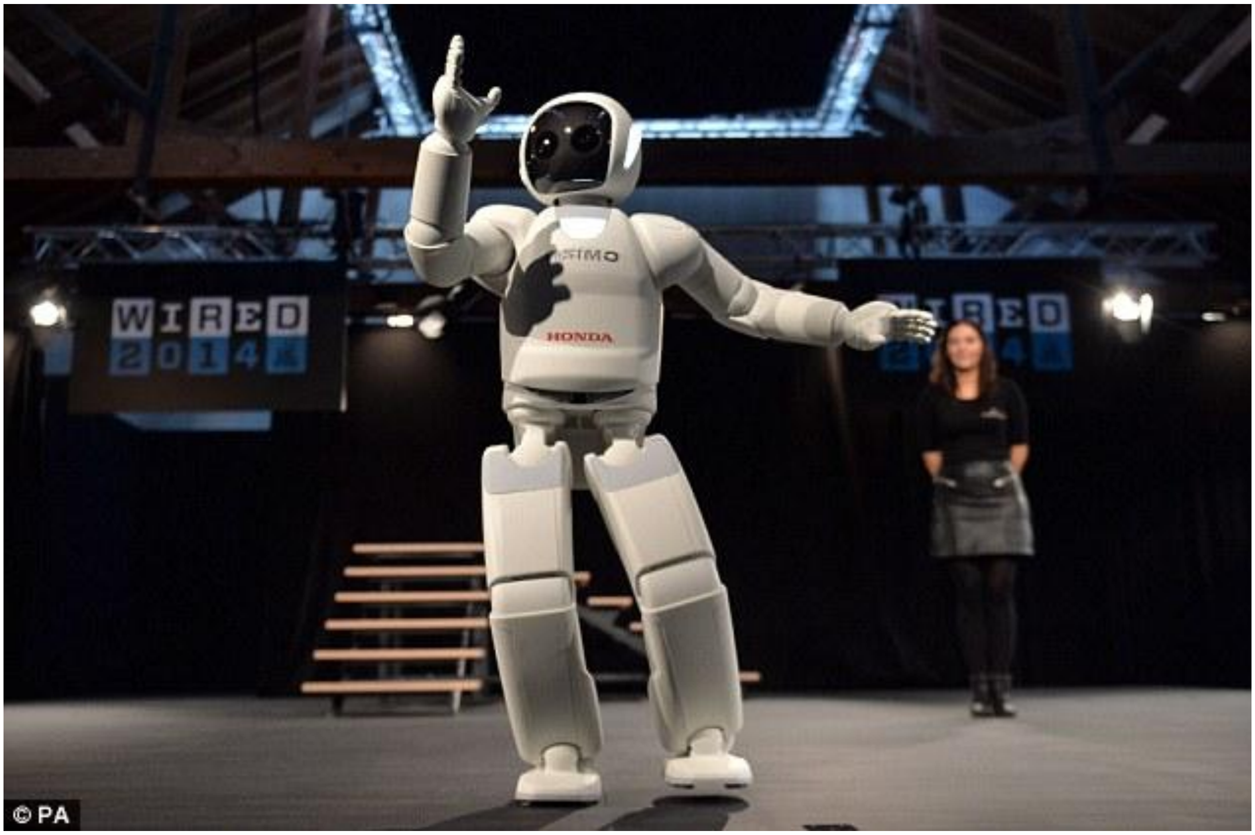
the entity who existed on October 31st, each of whom applies for unemployment benefits and receives one cookie from the dole. November 2nd the million entities vote for their favorite candidate. November 3rd the entities merge back together into one entity, who has memories of each entity's November 1st-2nd experiences, and who now has a million cookies and looks forward to its candidate's inauguration. Maybe next year it will decide to split into a million again, or a thousand, or maybe it will merge with the friendly fission-fusion monster next door. In general, if moral concern is to be distributed equally among discrete individuals, it might be possible for AIs to win additional moral concern by exploiting the boundaries of individuality.

Whatever it is that we morally value – unless (contra Section 6) it is natural humanity itself – it would be rare stuff indeed if no hypothetical AI could possess more of it than a natural human. This leaves us with some troubling puzzles for how to distribute moral concern.

10. Cute AI and the ASIMO Problem.

A couple of years ago the first author of this essay, Eric, saw the ASIMO show at Disneyland. ASIMO is a robot designed by Honda to walk bipedally with something like the human gait. Eric had entered the show with somewhat negative expectations about ASIMO, having read Andy Clark's (2011) critique of Honda's computationally-heavy approach to robot locomotion, and the animatronic Lincoln elsewhere in the park had left him cold.

But ASIMO is cute! He's about four feet tall, humanoid, with big round dark eyes inside what looks a bit like an astronaut's helmet. He talks, he dances, he kicks soccer balls, he makes funny hand gestures. On the Disneyland stage, he keeps up a fun patter with a human actor. Although his gait isn't quite human, his nervous-looking crouching run only makes him that much cuter. By the end of the show Eric thought that if you gave him a shotgun and asked him to blow off ASIMO’s head, he’d be very reluctant to do so (whereas he might rather enjoy taking a shotgun to his darn glitchy laptop).



ASIMO the cute robot. Image from Daily Mail UK (Oct. 17, 2014).

[Note to editors: Need permission. We could substitute an alternative image, though this one better captures ASIMO's cuteness than most of the alternatives. Color preferred but black and white okay.]

Another case: ELIZA was a simple chat program written in the 1960s that used a small range of pre-programmed response templates to imitate a non-directive psychotherapist ("Can you think of a specific example", "Tell me more about your family"). Apparently, some users found that the program created a powerful illusion of understanding them and spent long periods chatting with it (Weizenbaum 1976).

We assume that ASIMO and ELIZA are not proper targets of substantial moral concern. They have no more consciousness than a laptop computer, no more capacity for joy and suffering. However, because they share some of the superficial features of human beings, people might come improperly to regard them as substantial targets of moral concern. And future engineers could presumably create entities with an even better repertoire of superficial tricks, such as a robot that shrieks and cries and pleads when its battery runs low.

Conversely, an ugly or boxy human-grade AI or an AI in a simulated world without a good human-user interface, might tend to attract less moral concern than is warranted. Our emotional responses to AIs might be misaligned with the moral status of those AIs, due to superficial features that are out of step with the real psycho-social grounds of moral status.

Evidence from developmental psychology suggests that human beings are much readier, from infancy, to attribute mental states to entities with eyes, movement patterns that look goal directed, and contingent patterns of responsiveness than to attribute mentality to eyeless entities with inertial movement patterns and non-interactive responses.[16] But of course such superficial features needn't track underlying mentality very well in AI cases.

Call this the *ASIMO Problem.*

[16] Johnson 2003; Meltzoff, Brooks, Shon, and Rao 2010; Fiala, Arico, and Nichols 2012.

We draw two main lessons from the ASIMO Problem. First is a methodological lesson: In thinking about the moral status of AI, we should be careful not to overweight emotional reactions and intuitive judgments that might be driven by such superficial features. Low-quality science fiction – especially low-quality science fiction movies and television – does often rely on audience reaction to such superficial features. However, thoughtful science fiction sometimes challenges or even inverts these reactions.[17]

The second lesson is AI design advice. As responsible creators of artificial entities, we should want people to neither over-attribute nor under-attribute moral status to the entities with which they interact, when that misattribution jeopardizes the well-being or autonomy of an entity with legitimate moral consideration. We don't want anyone risking their life because they mistakenly believe they are protecting more than the mindless Furby before them, just like we don't want anyone neglecting their Sim just because they don't realize it's a conscious creature with genuine feelings. Thus, we should generally try to avoid designing entities that don't deserve moral consideration but to which normal users are nonetheless inclined to give substantial moral consideration, and conversely, if we do someday create genuinely human-grade AIs who merit substantial moral concern, it would probably be good to design them so that they evoke the proper range of emotional responses from normal users. Maybe we can call this the Emotional Alignment Design Policy.[18]

[17] For example, the Overlords in Clarke 1953, Aunt Beast in L'Engle 1962, and the "spiders" in Vinge 1999.

[18] Compare the Engineering and Physical Sciences Research Council's 4th "Principle of Robotics" (Boden et al. 2010): "Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent". This expresses one half of the Emotional Alignment Design Policy. See also Bryson 2010, 2013.

Pets and children's toys present an interesting range of cases here. On the one hand, manufacturers might understandably be tempted to create toys and pets that people will love and

Cute stuffed animals and Japanese helper-bots for the elderly, as they currently exist, probably do not violate this design policy, since we doubt that normal people would be inclined to sacrifice substantial human interests for the sake of these entities, based on false attributions of mentality to those objects. Spending money to fix a treasured toy is not morally problematic (except perhaps in the way that luxury expenditures in general might sometimes be problematic). The kind of case we have in mind, instead, is this: ASIMO and a human stranger both fall overboard. Because ASIMO is so cute and real looking and so compellingly says "Help me! Oh I'm dying!" a fellow passenger who falsely believes it capable of genuine suffering chooses to save it while the real person drowns.

12. The Strange Epistemology of Artificial Consciousness.

At the end of Section 4, we mentioned that the arguments of Searle, Lovelace, and Penrose raise concerns about the detection of psychological properties in AIs. This is the ASIMO Problem raised to possibly catastrophic proportions.

Searle (1980) imagines a "Chinese room" in which a monolingual English-speaker sits. Chinese characters are passed into the room. The room's inhabitant consults a giant lookup table, and on the basis of what he sees, he passes other Chinese characters out of the room. If the lookup table is large enough and good enough and if we ignore issues of speed, in principle, according to Searle, the inhabitant's responses could so closely resemble real human responses that he would be mistaken for a fluent Chinese speaker, despite having no understanding of

attach to, perhaps partly by using superficial cues that lead to the overattribution of mentality. On the other hand, a partial exception to the Emotional Alignment Design Policy might be justified if attachment toys can help cultivate moral sensibilities in children, assuming that when those children grow up, they can retain what was cultivated while coming to recognize those toys as not legitimate objects of serious moral consideration.

Chinese. Ned Block (1978/2007) similarly imagines a mannequin whose motions are controlled by a billion people consulting a lookup table, whose resulting behavior is indistinguishable from that of a genuinely conscious person. Suppose Searle or Block is correct, and a being who behaves very similarly to a human being from the outside might not be genuinely conscious, if it is not constructed from the right types of materials or according to the right design principles. People seeing it only from the outside will presumably be inclined to misattribute a genuine stream of conscious experience to it – and if they open it up, they might have very little idea what to look for to settle the question of whether the thing genuinely *is* conscious (Block 2002/2007 even suggests that this might be an impossible question to settle). Analogous epistemic risks attend broadly Lovelacian and Penrosian views: How can we know whether an agent is free or pre-determined, operating merely algorithmically or with genuine conscious insight? This might be neither obvious from outside nor discoverable by cracking the thing open; and yet on such views, the answer is crucial to the entity's moral status.

Even setting aside such concerns, the epistemology of consciousness is difficult. It remains wide open how broadly consciousness spreads across the animal kingdom on Earth and what processes are the conscious ones in human beings. The live options span the entire range from radical panpsychism according to which everything in the universe is conscious all the way to views on which consciousness, that is, a genuine, morally considerable stream of experience, is limited only to mature human beings in their most reflective moments.[19]

Although it seems reasonable to assume that we have not yet developed an artificial entity with a genuinely conscious stream of experience that merits substantial moral consideration, our poor understanding of consciousness raises the possibility that we might

[19] For more detail on the first author's generally skeptical views about the epistemology of consciousness, see Schwitzgebel 2011, 2014, 2015a.

someday create an artificial entity whose status as a genuinely conscious being is a matter of serious dispute. This entity, we might imagine, says "ow!" when you strike its toe, says it enjoys watching sports on television, professes love for its friends – and it's not obvious that these are simple pre-programmed responses (as they would be for ELIZA or ASIMO), but neither is it obvious that these responses reflect the genuine feelings of a conscious being. The world's most knowledgeable authorities disagree, dividing into believers (yes, this is real conscious experience, just like we have!) and disbelievers (no way, you're just falling for superficial tricks instantiated in a dumb machine).

Such cases raise the possibility of moral catastrophe. If the disbelievers wrongly win, then we might perpetuate the moral equivalents of slavery and murder without realizing we are doing so. If the believers wrongly win, we might sacrifice real human interests for the sake of artificial entities who don't have interests worth the sacrifice.

As with the ASIMO problem, we draw two lessons. First, if society continues on the path toward developing more sophisticated artificial intelligence, developing a good theory of consciousness is a moral imperative. Second if we do reach the point where we can create entities whose moral status is reasonably disputable, we should consider an Excluded Middle Policy – that is, a policy of only creating AIs whose moral status is clear, one way or the other.[20]

12. How Weird Minds Might Destabilize Human Ethics.

[20] In her provocatively titled article "Robots should be slaves" (2010; see also Bryson 2013), Joanna J. Bryson argues for a version of the Excluded Middle Policy: Since robots with enough mental sophistication might become targets of moral concern, we should adopt a policy of only making robots sufficiently unsophisticated that their "enslavement" would be morally permissible.

Intuitive physics works great for picking berries, throwing stones, and loading baskets. It's a complete disaster when applied to the very large, the very small, the very energetic, and the very fast. Intuitive biology, intuitive cosmology, and intuitive mathematics are much the same: They succeed for practical purposes across long-familiar types of cases, but when extended too far they go wildly astray.

We incline toward moral realism. We think that there are moral facts that people can get right or wrong. Hitler's moral attitudes were not just different but mistaken. The 20th century "rights revolutions" (women's rights, ethnic rights, worker's rights, gay rights, children's rights) were not just change but progress toward a better appreciation of the moral facts. Our reflections in this essay lead us to worry that if artificial intelligence research continues to progress, intuitive ethics might encounter a range of cases for which it is as ill-prepared as intuitive physics was for quantum entanglement and relativistic time dilation. If that happens, and if there are moral facts, possibly we will get those facts badly wrong.[21]

Intuitive ethics was shaped in a context in which the only species capable of human-grade practical and theoretical reasoning was humanity itself, and in which human variation tended to stay within certain boundaries. It would be unsurprising if intuitive ethics were ill-prepared for utility monsters, fission-fusion monsters, AIs of vastly superior intelligence, highly intelligent AIs nonetheless designed to be cheerfully suicidal slaves, toys with features designed specifically to capture children's affection, giant virtual sim-worlds that can be instantiated on a home computer, or entities with radically different value systems. We might expect human moral judgment to be baffled by such cases and to deliver wrong or contradictory or unstable verdicts.

[21] Compare Bakker on "crash spaces" for our "ancestral ways of meaning making" (Bakker this issue, postscript).

In the case of physics and biology, we have pretty good scientific theories by which to correct our intuitive judgments, so it's no problem if we leave ordinary judgment behind in such matters. However, it's not clear that we have, or will have, such well-founded replacement theories in ethics. There are, of course, ambitious ethical theories – "maximize happiness", "act on that maxim that you can will to be a universal law" – but the development and adjudication of such theories depends, and might inevitably depend, upon our intuitive judgments about such cases.[22] It's because we intuitively or pre-theoretically think that we shouldn't give all our cookies to the utility monster or kill ourselves to tile the solar system with hedonium that we reject the straightforward extension of utilitarian happiness-maximizing theory to such cases and reach for a different solution. But if our intuitions about such cases are not to be trusted, because such cases are too far beyond what we can reasonably expect human moral intuition to handle – well, what then? Maybe we *should* kill ourselves for sake of hedonium, and we're just unable to appreciate this moral fact with our old theories shaped for our limited ancestral environments?

A partial way out might be this. If morality partly depends on our intuitions and reactions, whatever moral facts there are might partly depend on what we (or idealized versions of ourselves) think the moral facts are. Much like an object's being brown, on a certain view of the nature of color, just consists in its being such that ordinary human perceivers in normal conditions would experience it at brown, maybe an action's being morally right just consists in its being such that ordinary human beings who considered the matter carefully enough would tend to regard that action as right – or something in that ballpark.[23] If so, morality might change

[22] Rawls's (1971) "reflective equilibrium" methodology is a classic version of such a view.

[23] We've used a "secondary quality" type phrasing here, but in fact we are imagining a broad class of views such as the (disagreeing) views of McDowell 1985; Railton 1986; Brink 1989; Casebeer 2003; and Flanagan, Sarkissian, and Wong 2007 – naturalistic, allowing for

as our sense of the world changes – and maybe, too, as who counts as "we" changes. Maybe we could decide to give the fission-fusion monster some rights but not other rights, and shape future intuitions accordingly. The unsettled nature of our intuitions about such cases, then, might present an opportunity for us to shape morality – real morality, the real (or real enough) moral facts – in one direction or another, by shaping our future dispositions and habits. Maybe different social groups would make different moral choices with different consequences for group survival, introducing cultural evolution into the mix.[24] Moral confusion might open into a range of choices for moral architecture.[25]

However, the range of legitimate choices is we think constrained by certain moral facts sufficiently implacable that a system that rejected them would not be a satisfactory moral system on the best way of construing the possible boundaries of "morality" worth the name. One such implacable fact is that it would be a moral disaster if our future society constructed large numbers of human-grade AIs, as self-aware as we are, as anxious about their future, and as capable of joy and suffering, simply to torture, enslave, and kill them for trivial reasons.[26]

genuine moral truths, with norms contingent upon facts about the human condition, but not so strongly relativist as to deny a normatively compelling, fairly stable moral core across human cultures as they have existed so far.

[24] Compare Mandik (forthcoming, this issue) on cultural selection for metaphysical daring in a posthuman environment.

[25] Thus, despite the generally moral realist framing of this article, we accept aspects of the more constructivist and relativist views of Coeckelbergh 2012 and Gunkel 2012, according to which we collaboratively decide, rather than discover, who is and who is not part of the moral community and "grow" moral relations through actively engaging with the world.

[26] For helpful discussion, thanks to Scott Bakker, Peter Hankins, Justin E.H. Smith, and the many people who commented, privately or publicly, on our posts on these topics on The Splintered Mind and Codex Writers Group.

References:

Asimov, Isaac (1954/1962). *Caves of steel.* New York: Pyramid.

Asimov, Isaac (1982). *The complete robot.* Garden City, NY: Doubleday.

Banks, Iain (1987). *Consider Phlebas.* New York: Hachette.

Banks, Iain (2012). *The Hydrogen Sonata.* New York: Hachette.

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.

Block, Ned (1978/2007). Troubles with functionalism. In *Consciousness, function, and representation.* Cambridge, MA: MIT.

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

Boden, Margaret, et al. (2010). Principles of robotics. ESPRC website (accessed Aug. 21, 2015): https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

Bostrom, Nick (2003). Are we living in a computer simulation? *Philosophical Quarterly, 53,* 243-255.

Bostrom, Nick (2014). *Superintelligence.* Oxford: Oxford.

Bostrom, Nick, and Eliezer Yudkowsky (2014). The ethics of artificial intelligence. In K. Frankish and W.M. Ramsey, eds., *Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge.

Briggs, Rachael, and Daniel Nolan (forthcoming). Utility monsters for the fission age. *Pacific Philosophical Quarterly*.

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

Brooker, Charlie, and Carl Tibbets (2014). White Christmas. Episode of *Black Mirror,* season 3, episode 0.

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.

Chalmers, David J. (1996). *The conscious mind.* Oxford: Oxford.

Clark, Andy (2011). *Supersizing the mind.* Oxford: Oxford.

Clarke, Arthur C. (1953). *Childhood's end.* New York: Random House.

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

Cuda, Tom (1985). Against neural chauvinism. *Philosophical Studies, 48,* 111-127.

de Bodard, Aliette (2011). Shipbirth. *Asimov's, 35* (2), 50-60.

de Bodard, Aliette (2013). The waiting stars. In A. Andreadis and K. Holt, eds., *The other half of the sky.* Bennington, VT: Candlemark & Gleam.

DeGrazia, David (2009). Moral vegetarianism from a very broad basis. *Journal of Moral Philosophy, 6,* 455-468.

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

Egan, Greg (1994). *Permutation City.* London: Millennium.

Egan, Greg (1997). *Diaspora*. London: Millennium.

Estrada, Daniel (2014). *Rethinking machines.* PhD dissertation, Philosophy Department, University of Illinois, Urbana-Champaign.

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

Fiala, Brian, Adam Arico, and Shaun Nichols (2012). The psychological origins of dualism. In E. Slingerland and M. Collard, eds., *Creating consilience.* Oxford: Oxford.

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.

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

Gunkel, David J. and Joanna J. Bryson, eds. (2014). Machine morality. Special issue of *Philosophy & Technology, 27* (1).

Hankins, Peter (2015). Crimbots. Blogpost at *Conscious Entities* (Feb. 2), http://www.consciousentities.com/?p=1851

Johnson, Susan C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society B, 358*, 549-559.

Kagan, Shelly (forthcoming). What's wrong with speciesism? *Journal of Applied Philosophy.*

L'Engle, Madeline (1963). *A wrinkle in time.* New York: Scholastic.

Lee, Stan and Steve Ditko (1962). Spider-Man. *Amazing Fantasy, 15.*

Lee, Stan, Steve Ditko, David Koepp, and Sam Raimi (2002). *Spider-Man.* Columbia Pictures.

Lovelace, Ada (1843). Sketch of the analytical engine invented by Charles Babbage, by L.F. Menabrea. In R. Taylor, ed., *Scientific Memoirs, vol. III.* London: Richard and John E. Taylor.

Mandik, Pete (forthcoming, this issue). Metaphysical daring as a posthuman survival strategy. *Midwest Studies in Philosophy*.

McDowell, John (1985). Values and secondary qualities. In T. Honderich, ed., *Morality and objectivity.* New York: Routledge.

Meltzoff, Andrew N., Rechele Brooks, Aaron P. Shon, and Rajesh P.N. Rao (2010). "Social" robots are psychological agents for infants: A test of gaze following. *Neural Networks, 23*, 966-972.

Parfit, Derek (1984). *Reasons and persons.* Oxford: Oxford.

Penrose, Roger (1999). *The emperor's new mind.* New York: Oxford.

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

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

Schwitzgebel, Eric (2011). *Perplexities of consciousness.* Cambridge, MA: MIT.

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

Schwitzgebel, Eric (2015a). If materialism is true, the United States is probably conscious. *Philosophical Studies*, *172,* 1697-1721.

Schwitzgebel, Eric (2015b). Out of the jar. *Magazine of Fantasy & Science Fiction, 128* (1), 118-128.

Schwitzgebel, Eric, and R. Scott Bakker (2013). Reinstalling Eden. *Nature, 503,* 562.

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

Shelley, Mary (1818/1965). *Frankenstein.* New York: Signet.

Singer, Peter (1975/2002). *Animal liberation.* New York: Ecco.

Singer, Peter (2009). Speciesism and moral status. *Metaphilosophy, 40,* 567-581.

Snodgrass, Melinda M., and Robert Scheerer (1989). The measure of a man. *Star Trek: The Next Generation,* season 2, episode 9.

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

Turing, A.M. (1950). Computing machinery and intelligence. *Mind,* 433-460.

Vinge, Vernor (1992). *A fire upon the deep.* New York: Tor.

Vinge, Vernor (1999). *A deepness in the sky.* New York: Tor.

Vinge, Vernor (2011). *Children of the sky.* New York: Tor.

Weizenbaum, Joseph (1976). *Computer power and human reason.* San Francisco: W.H. Freeman.

Williams, Bernard (2006). *Philosophy as a humanistic discipline,* ed. A.W. Moore. Princeton, NJ: Princeton.