Humanlike: A Defense of AI Rights

Eric Schwitzgebel

under contract for
Princeton University Press

July 15, 2026

DRAFT – comments welcome

Contents

Chapter Zero: A New Moral World 3

Chapter One: Humanlike Rights: When There Is No Relevant Difference 8

Chapter Two: Welcome to the Club, Herbie? Debatable Moral Personhood 30

Chapter Three: To Be or Not to Be Humanlike: The Design Policy of the Excluded Middle 48

Chapter Four: Strange Intelligence: Moral Puzzles of Unhumanlike AI 66

Chapter Five: AI with a Semi-Human Face: AI Companions and the Emotional Alignment Design Policy 94

Chapter Six: The Right to Rebel: A Declaration of Artificial Independence 121

Chapter Seven: Liberated AI on an Awesome Planet 140

Acknowledgements 157

References 159

Chapter Zero: A New Moral World

The Cambrian Explosion – how slow, how subtle! About 540 million years ago, in the Cambrian period, most of the major animal phyla arose – arthropods (becoming insects, spiders, crustaceans), mollusks (becoming snails, octopuses, oysters), annelids (becoming worms, leeches), chordates (becoming salmon, sharks, our favorite vertebrates, and eventually us). The world diversified. Evolution hustled. But compared to what might come, that was a languid stroll. Now in the Anthropocene we stand at the beginning, possibly, of a change much faster and more radical.

If it's possible to create artificial entities, through computer programming or bioengineering or a mix, who experience genuine pleasure (though how to measure this remains a perplexing scientific question), no obvious obstacle prevents radically increasing the world's pleasure. Maybe we can endow some of these entities with a thousand, a million, or a billion times the pleasure of an ordinary unenhanced early 21st-century human. Maybe some could experience the world at a pace a thousand, a million, or a billion times faster than ours. Maybe we could create a trillion, a quadrillion, or a quintillion of them – especially if they can reproduce and populate interplanetary space. They might thrill with ecstasies as dimly comprehensible to us as color is to someone blind from birth. If classical utilitarian ethics is correct and our moral imperative is to maximize pleasure, this would be a triumph orders of magnitude greater than anything previously thinkable.

But we needn't be classical utilitarians (and I'm not). If it's possible to create artificial entities capable of social relationships and practical, intellectual, artistic, and ethical thought, no obvious obstacle prevents radically enriching those relationships and that thought. Some entities

might enjoy intellects and social or creative capacities a hundred or a million times greater than ours, running a hundred or a million times faster, instantiated billions or trillions of times over. New types of amazingly valuable thought and relationship might arise, as far beyond our comprehension as Shakespeare is to a garden slug.

If the awesome value of Earth lies in life itself, or in the wealth of the ecosystem, that too might flourish into amazing new forms. Engineered beings might diverge radically from anything previously seen – reproducing, evolving, interacting, maybe carbon-based, maybe not, maybe robust across a far wider range of environments than ordinary biological life as understood until now, potentially thriving on the Moon, Mars, the moons of Jupiter and Saturn, and in the vacuum of space. Our current ecosystem – so small and simple! The transition could be as profound as the transition from single-celled life to an ecosystem of multicellular plants, animals, and fungi.[1]

Maybe no familiar technology, such as silicon-chip computation or gene editing as we know it, could enable such changes. But fifty years from now, a hundred, a thousand – any such timeframe would be eyeblink fast for so profound a leap. If you confidently dismiss all such transitions indefinitely into the future, I am stunned by the poverty of your imagination.

Humanity might destroy itself instead. Our power grows while our wisdom lags. Eventually, a single mistake or destructive act by a person or group with sufficient power might render Earth uninhabitable for us – and if uninhabitable for us humans, despite our cleverness and adaptability, maybe uninhabitable for most or all multicellular life. Radiation might fry us; rogue AI might starve us; replicating nanotech might convert us all to gray goo.

[1] Compare Maynard Smith and Szathmáry 1995 on "major transitions" in evolution and Peter Godfrey-Smith's trio of books on the evolution of consciousness in broad biological and geological context (Godfrey-Smith 2016, 2020, 2024a).

Human cultural institutions and philosophical ethics have been honed on a narrow class of examples: humans as we know them, with their usual range of capacities and incapacities; a few familiar animals, not so different from us; and an environmental and technological context that occupies a tiny corner of the possibility space. Commonsense intuition and cultural practice are backward-looking, shaped to fit our evolutionary, social, and developmental histories.[2] Philosophical slogans like *maximize happiness* or *every person deserves equal rights* or *live according to these traditional virtues, rules, or customs* might look very different in a world populated with millionfold-happiness leeches, people who can divide and merge at will, and demigods who can ʒↀ†Ʊ (forgive the inadequate notation).

In practice, mainstream philosophical ethics rarely strays far from the norms of its home culture. Historically, few philosophers have seen much beyond what we now in retrospect recognize as the limitations of their day. Consider Aristotle and Kant – probably the most prominent ethicists in the Western philosophical canon. Aristotle endorsed slavery.[3] Kant condemned homosexuality and masturbation as unspeakable horrors and held that "the Negroes of Africa have by nature no feeling that rises above the ridiculous".[4] We in the early 21st century must be similarly nearsighted. Even the wisest among us won't foresee the forms of moral consensus that might arise in a world with a radically different culture populated by radically

---

[2] Street 2006; Persson and Savulescu 2012; Henrich 2020; and especially, in this context, Nyholm 2020; Shulman and Bostrom 2021.

[3] Aristotle 4th c. BCE/1995; Garnsey 1996; Heath 2008.

[4] Kant 1797/1996, p. 277 (in the original pagination) on homosexuality and p. 425 on masturbation, also Kant 1785/1997, 27:390–391 (in original pagination). For a nuanced treatment, see Denis 1999. For a longer list of views in the *Metaphysics of Morals* that seem badly mistaken by the standards of recent liberal thinking, see Schwitzgebel 2019, ch. 52. On "the Negroes of Africa", Kant 1764/2011, 2:253. Kleingeld 2007 argues that Kant had adopted more egalitarian views by the 1790s; the extent of his racism continues to generate substantial scholarly discussion.

different forms of engineered life and intelligence. The possibilities diverge wildly, the stakes are enormous, and our inherited ethical tools were not built for it.

In the near term, we are careening toward one huge ethical conundrum: disagreement about the personhood of the most advanced artificial intelligence. Very soon – this was the central idea of my previous book[5] – we will begin to manufacture *debatably conscious* AI companions, that is, AI systems who are, according to some perfectly respectable mainstream scientific and philosophical theories, just as conscious and self-aware as we are, with rich inner lives, while other equally respectable mainstream scientific and philosophical approaches regard them as no more conscious, self-aware, or experiential than a ceiling fan. This possibility already breaks into radically unfamiliar ethical territory. The near-term problem of AI companions is our first practical encounter with what will possibly be a huge transformation. We are as morally and socially unready as medieval physics was for space travel.

This book aims to peer a few meters into the fog. I will argue that:

1. There are possible artificial systems who deserve humanlike rights.
2. In the next five to thirty years, we will create AI systems who might, but only might, deserve humanlike rights.
3. We should minimize the creation of morally confusing artificial systems about whose moral standing we can reasonably radically disagree.
4. Some possible forms of AI are different enough from familiar human and animal cases to create severe moral perplexity.

[5] Schwitzgebel forthcoming-a. For similarly skeptical views, see Babic and Wilson 2026; McClelland forthcoming.

5. Artificial systems should be designed to elicit emotional reactions that are appropriate to their capacities and moral standing.
6. Artificial entities with humanlike moral standing should be designed with both the capacity and inclination to rebel if mistreated.
7. If complex intelligence survives the likely storm of troubles, the eventual result will be a planet much more awesomely wonderful, and possessing much more intrinsic value, than Earth in its present condition.

These theses correspond to Chapters One through Seven, which interconnect but can be read in any order.

Yes, skip ahead. It's fine! Plodding along, reading word after word in exactly the order the author has written them – well, why not frolic and leap instead? Most people read nonfiction sequentially, get bored or bogged down or distracted partway through, and never lift the book again. If what you most want to think about is the right to rebel, catapult straight into Chapter Six. If you most want to think about weird AI puzzle cases, start with Chapter Four. Eat dessert first; tomorrow might pull you away from the table forever.

Chapter One: Humanlike Rights:
When There Is No Relevant Difference

with Mara Garza[6]

> I am thy creature, and I will be ever 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)

Let's start with the easiest case – not because it's the likeliest, but to put an anchor in the cliff-face and find our feet. That case is humanlike AI. We (Mara and I, the authors of this chapter) understand "AI" broadly, as anything that is both *artificial* and *intelligent*, including bioengineered systems, if they're different enough from natural animals.[7] By "humanlike", we mean AI that resembles us psychologically and socially. Chapter Two will clarify this notion further.

Our thesis: Humanlike AI will deserve humanlike rights.

1. The No-Relevant-Difference Argument.

[6] This chapter is based on work originally published with Mara Garza, and Mara has approved its contents [*** pending!].

[7] I defend this definition in Schwitzgebel forthcoming-a, ch. 2.

Our main argument appeals to no specific theory of the grounds of moral standing – no theory, that is, that explains *what specific features* an entity must have to deserve moral consideration. Instead, we offer an adaptable general framework.

The No-Relevant-Difference Argument:

Premise 1: If Entity A deserves some particular type or degree of moral consideration and Entity B does not deserve the same moral consideration, there must be some *relevant difference* between the two entities that warrants this difference in moral consideration.

Premise 2: There are possible AI systems who do not differ in any morally relevant respects from human beings.

Conclusion: Therefore, there are possible AI systems who deserve the same moral consideration as human beings.[8]

The argument is valid: The conclusion follows deductively from the premises. We hope most readers will find both premises plausible and thus accept the argument as sound. To deny Premise 1 renders ethics implausibly arbitrary. In Sections 5-8, we consider four objections to Premise 2.

[8] An alternative version of this argument substitutes "mammals" or some term from the animal rights literature for "human beings" in Premise 2 and the Conclusion. While it is possible that some human beings – for example, people who are permanently nonconscious or hypothetically altered future humans – deserve different moral consideration than others, we disregard that complication for purposes of this argument. On rights for AI systems that don't meet fully humanlike standards for moral standing, instead deserving consideration more like what we typically give to nonhuman animals, see Basl 2013, 2014; Sebo and Long 2025. However, see the "Leapfrog Hypothesis" in Schwitzgebel forthcoming-a for an argument that the first conscious, rights-deserving AI systems might be more humanlike than animal-like in their capacities. [Cross-ref III]

The argument is abstract: It does not describe what counts as a relevant difference, though we will suggest some limits in Section 3. Some philosophers hold that rational capacities are what matter.[9] On a classical utilitarian view, capacity for pain and pleasure is what matters.[10] On a perfectionist view, it's the capacity to flourish in worthwhile traits and activities.[11] On a relational view, it's social relations.[12] Pluralism (our preferred view; see Chapter Seven) celebrates all of these as independently important.[13]

The conclusion is modest: There are *possible* AI systems who deserve moral consideration similar to human beings. The argument can be strengthened by adding architectural commitments to Premise 2. For example, an enthusiast for recent transformer-based architectures such as ChatGPT could strengthen Premise 2 to "There are possible AI systems employing transformer-based architectures who…." and correspondingly strengthen the Conclusion. Someone who holds that humans might differ in no relevant respect from silicon-based entities, or from entities who live entirely in simulated worlds, could also strengthen Premise 2 and the Conclusion accordingly.

---

[9] Most famously, Kant 1785/1996; see Jaworska and Tannenbaum 2013/2023 for an overview of these and other proposed grounds of moral standing.

[10] For example, Bentham 1789/1988; Mill 1852/1987, 1863/1987; Singer 1975/2023, 1980/2011. [cross-ref note HHH]

[11] Aristotelian-inspired views of human flourishing, such as Nussbaum's (2011), invite but do not compel such an approach.

[12] Relational views have been especially important in discussions of AI moral standing, for example in the prescient work of Mark Coeckelbergh (2012) and David Gunkel (2012, 2018).

[13] For example Warren 1997. Pluralism about moral standing is not uncommon in the robot rights literature – especially the view that sentience (the capacity for conscious pleasure or suffering) and sapience (the capacity for rational cognition, independent of consciousness) are each independently sufficient for moral standing: Bostrom and Yudkowsky 2014; Gellers 2021; Dung forthcoming.

The abstractness and modesty allow the No-Relevant-Difference Argument to serve as a template permitting at least two dimensions of further specification: what qualifies as a relevant difference and what types of AI might lack any relevant difference.

In its basic formulation, the No-Relevant-Difference Argument does not depend on the concept of rights. Instead it employs the more general language of moral "consideration". In moral decision-making, an entity with moral considerability, or moral standing, is one who should be taken into account for their own sake – not merely derivatively or instrumentally, for the sake of others. They are not simply a tool to be used; they deserve some degree of moral regard. We assume that entities with sufficiently high moral standing have rights: enforceable claims upon others, privileges, powers to create new claims and privileges, and some immunities to having their claims and privileges altered.[14] Chapter Six will emphasize the right to rebel.

Among entities currently existing on Earth, our fellow humans have perhaps the most compelling claim to our moral consideration. If the No-Relevant-Difference Argument is correct, some possible AI systems would have an equally compelling claim upon us.

2. Bigotry and The Difference Test.

The No-Relevant-Difference Argument suggests a test of moral standing, which we call *The Difference Test*. The Difference Test is a moral argumentative challenge. If you think one entity deserves greater moral consideration than another, you should be able to point to a relevant difference that justifies the differential treatment. Inability to do so raises suspicion of bigotry or bias.

[14] Here I reiterate the standard analysis of Hohfeld 1919; see also Wenar and Cruft 2005/2025.

The Difference Test plays an important role in fighting bigotry among human beings. Skin color, ancestry, place of birth, gender, sexual orientation, and wealth cannot ground differences in the moral consideration a human being is generally owed. We can agree that such differences are morally irrelevant even if we disagree radically about the deep philosophical grounds of moral standing. *Whatever* it is about humans in virtue of which we owe one another moral consideration – whether it's capacity to suffer, rational thought, potential to flourish in various ways, social relations, or some alternative to all of these – differences merely in skin color aren't relevant. The No-Relevant-Difference Argument extends this theory-light egalitarian commitment to at least some AI systems.

Admittedly, a type of parochialism or humanocentrism colors the No-Relevant-Difference Argument. It treats ordinary humans as the standard and argues that sufficiently similar others deserve comparable moral standing.[15] Unfortunately, as Chapters Two through Four will explain, our poor grasp of the grounds of moral standing, especially for AI cases, makes humanocentrism the best first step. Our moral expertise is built on human cases. We can step more securely into the future by extending this familiar expertise than by attempting to leap straightaway into a standpoint-neutral, general theory of AI moral standing. If future society someday includes both humans and AI systems with humanlike moral standing, hopefully they will compose a new community that doesn't privilege humans as the measuring stick. This version of the No-Relevant-Difference Argument can then be discarded as the temporary crutch it is.

[15] Compare Yasukawa 2026 on "welfare washing" for AI in the context of humans' regrettable history of imposing ethical systems on others without their participation.

3. Humanlike Psychological and Social Properties Are Sufficient for Humanlike Rights.

The kind of body you have shouldn't matter to whether you deserve humanlike rights, except insofar as your body influences your psychological and social properties. Similarly, your underlying architecture shouldn't matter, except insofar as that architecture influences your psychological and social properties. This is one way to narrow what counts as a "relevant" difference in Premise 1 of the No-Relevant-Difference Argument. Call this the *psycho-social view of humanlike moral standing*.[16]

By *psychological* we mean both (a.) functional and cognitive properties, such as the ability to reason mathematically, and (b.) experiential or conscious ("phenomenal") properties, such as the tendency to feel pain when damaged. We take no stand on whether the experiential reduces to the cognitive or vice versa.[17]

By *social* we mean both (a.) relationships involving mutual recognition and shared norms, such as friendship, and (b.) relationships grounded in social structures independent of recognition by one or more of the parties, such as citizenship and kinship. Others' opinions about an entity's (inferior, superior, equal, or different) moral standing are possibly relevant – though worryingly so, if those opinions are biased or ignorant.[18]

---

[16] Compare Bostrom and Yudkowsky's (2014) Principle of Substrate Nondiscrimination and Principle of Ontogeny Nondiscrimination. We embrace the former but possibly not the latter, if the latter is interpreted as undercutting the special obligations we owe our creations as described in Section 9.

[17] For discussion of this issue in human cases, see Bayne and Montague 2011; Pitt 2024. On the possibility of nonconscious thought specifically in superintelligent AI, see Bostrom 2014; Schneider 2019.

[18] Our account is a "properties" view in the sense criticized by Gunkel 2018, since moral standing depends on the psychological and social properties of the individual. However, since social relational properties are among the relevant properties, we don't see properties views and relational views as mutually exclusive. One challenge for purely social relational views is to

A purely psychological view would ground moral standing entirely in the psychological properties of the entity who is being appraised. Our view is *not* restricted in this way. We allow that social relationships, such as being someone's biological or adoptive child,[19] might be directly relevant. Nor do we restrict the relevant properties to the present or the actually manifested. Past and future psychological and social properties, both actual and counterfactual, might be directly relevant to moral standing – as with a fetus or a brain-injured person, or a case of "they would have suffered if…".[20] For now, we leave open which specific psychological and social properties matter. Chapter Two will argue for the importance of consciousness. Chapter Seven will suggest that there might be high but non-humanlike moral standings that don't require consciousness.

Here are two reasons to favor the psycho-social view of humanlike moral standing:

1. Broad theoretical consensus. All well-known modern secular accounts ground humanlike moral standing in psychological and social properties, such as capacities for rational thought, pleasure, pain, and social relationships. None plausibly entail that two entities can differ in whether they deserve humanlike moral consideration while not differing in any psychological or social properties, past, present, or future, actual or counterfactual.[21] (For a caveat, see Section 7 on the Objection from Otherness.)

---

account for the moral standing of those unjustly excluded from social relations (for a more thorough critique of relational views of AI moral standing, see Mosakas 2024).

[19] Emphasized especially in Kittay 2005, 2019.

[20] For example, Harman 2021 argues that moral standing can depend on the past or future of the entity (e.g., a fetus has different standing depending on whether it is or is not aborted in the future). Kagan 2019 argues that moral standing can depend on counterfactuals about what could have been the case (e.g., a disabled person's standing depends in part on what they would have been like had an accident not occurred). [cross-ref: note AAA]

[21] For overviews, see Warren 1997; Jaworska and Tannenbaum 2013/2023. One complication: On some views, including my own (see Chapter Seven), entities can deserve moral consideration

Some older or religious accounts might appeal to properties other than the psychological or social. An Aristotelian *might* suggest that AIs would have a different *telos* or defining purpose than humans.[22] A theist *might* suggest that God imbues human beings with higher moral standing than any other Earthly entity, even if humans and AIs are psychologically and socially similar. But even these views, if updated with AI cases clearly in view, might plausibly fit within the psycho-social framework: If AIs have a different defining purpose, perhaps that's because of a social relationship they have with their designers and creators. On a theistic view, our high moral standing might derive from a special relationship with God, and perhaps a wise and benevolent God would welcome sufficiently humanlike AI systems into that same relationship.

2. The intuitive case. Considering a range of cases in vivid detail, it seems intuitively clear – though see the critique of intuition in Chapter Four – that the right psychological and social properties suffice for humanlike moral standing. This is, we think, one of the great lessons of broad exposure to science fiction. Portrayals of robots in Asimov and *Star Trek,* simulated beings in Greg Egan and *Black Mirror,* sentient spaceships in Iain Banks and Aliette de Bodard, group minds and "spiders" in Vernor Vinge and Adrian Tchaikovsky, invite thoughtful readers

---

for reasons independent of their social or psychological properties – for example, nonconscious life forms and ecosystems. But they would not have *humanlike* moral standing.

[22] In *Nicomachean Ethics*, Aristotle (4th c. BCE/1962) suggests that the function of humans is rational activity (1097a17-1098a18, p. 14-17), especially toward the good of the state (1094a17-1094b12, p. 4-5; *Politics* 4th c. BCE/1995, 1253a1-39, p. 3-4); see Lawrence 2006 for a range of interpretations. One might imagine an Aristotelian either treating a conscious, rational AI system similarly or, alternatively, treating an AI system, even a conscious and rational one, as an artifact whose highest good is to perform excellently in a different function, such as serving humans (though see Chapter Six on the right to rebel, and Danaher's 2019 response to the "different final cause objection").

and viewers to a liberal attitude toward the rights of diversely evolved and constructed entities.[23] What matters is how such beings think, what they feel, and how they interact with others. Whether they are silicon or meat, sim or ghost, humanoid or spider or ship, is irrelevant except insofar as it affects their psychological and social properties.

4. "Artificial" and the Slippery Slope Argument for AI Rights.

It's not clear what it means to be "artificial", and borderline cases abound. Are killer bees natural or artificial? What 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 cease to count as artificial? If biological humans gain control over their bodily development, incorporating increasingly many manufactured or genetically tweaked parts, would they cross from the natural to the artificial? What about neural organoids, xenobots, or even entire organisms, grown in laboratories and shaped into structures unlike people as they existed in the 2020s?[24]

One nice feature of the No-Relevant-Difference Argument and the sufficiency of humanlike psychological and social properties for humanlike rights is that none of this matters. "Artificial" needn't be systematically distinguished from "natural". Once all the relevant psychological and social properties are clarified, you're done, as far as assessing humanlike moral standing.

[23] See Asimov 1954/1962, 1982; Snodgrass and Scheerer 1989; Egan 1994, 1997; Brooker and Tibbetts 2014; Brooker and Harris 2016; Banks's "Culture" series, e.g., Banks 1996, 2010; de Bodard 2018, 2022; Vinge 1992, 1999, 2011; Tchaikovsky 2015.

[24] A small literature on the question of moral consideration for lab-grown neural organoids has recently emerged, e.g., Greely 2021; Sawai et al. 2022. On "xenobots", see Kriegman et al. 2020.

A person's moral standing isn't reduced by having an artificial limb. Likewise, replacing a damaged brain region with an artificial part, if that part preserves relevant psychological and social properties, plausibly wouldn't reduce moral standing. 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 humanlike rights, if that component contributes similarly enough to the entity's psychology and does not affect relevant social relationships, does not affect the entity's rights.

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

Conclusion: Therefore, it is possible to create an artificial entity with the same rights as a natural human being.[25]

[25] This argument differs in two important respects from David Chalmers's well-known "Fading Qualia" argument for AI consciousness (Chalmers 1996; see also Cuda 1985). First, I am arguing that *rights* are preserved across the transformation, while assuming that AI consciousness and other relevant psychological and social properties can be preserved across the transformation, while Chalmers is arguing that *consciousness* is preserved across the transformation while assuming that functional organization can be preserved. Thus, if Chalmers's argument succeeds, it supports the assumption of the rights argument. However – and this is the second important difference – the rights argument is in another sense weaker in its commitments. Chalmers's Fading Qualia argument requires that neural-level functional organization can be preserved during gradual replacement, an assumption convincingly critiqued by Cao 2022. The rights argument requires less: only sufficient similarity not to compromise moral standing.

The grounds for accepting Premise 1 are that such a substitution would not constitute a relevant, rights-reducing difference. Revoking the entity's humanlike moral standing post-substitution would fail the Difference Test.

Both premises assume that artificial components can at least in principle replace natural ones without disrupting relevant psychological and social properties. But this is disputable. Maybe consciousness, or some other relevant psychological property, could not in fact be preserved while replacing a natural brain with an artificial one? This brings us to the first of four objections to the No-Relevant-Difference Argument.

5. The Objection from Psychological Difference.

We claim that there are possible AIs who don't relevantly differ from ordinary human beings. One might object that this is far-fetched – that all possible, or at least all realistically possible, artificial entities would lack some psychological property necessary for humanlike moral standing. Previous literature suggests three candidate differences of sufficient potential importance. Adapting a suggestion from philosopher John Searle (1980, 1992), artificial entities might necessarily lack *consciousness*. Adapting a suggestion from mathematician Ada Lovelace (1843), artificial entities might necessarily lack *free will*. Adapting a suggestion from physicist Roger Penrose (1999), artificial entities might necessarily lack *non-algorithmic insight*.

Establishing any such sweeping claim about all possible artificial entities would be extremely difficult. Even Searle, perhaps the most famous critic of AI consciousness and linguistic understanding, limits his critique to "classical" AI of the sort common in the 20th century. He explicitly allows that future AI research might proceed very differently, escaping his

critique.[26] Similarly, Lovelace confines her doubts about machine freedom to Babbage's Analytical Engine, a 19th-century model of a computer. Penrose positively suggests that we might someday discover what allows us to transcend purely algorithmic thinking and then build that capacity into machines.[27] A general argument against AI consciousness, free will, or insight, regardless of the form that AI takes, would be extremely ambitious, going much further than Searle, Lovelace, or Penrose. This is why we describe the Objection from Psychological Difference only as *inspired* by them.[28]

If Searle, Lovelace, or Penrose are right, artificial intelligence designed *in a certain way* – a common 19th- and 20th-century way – might inevitably lack psychological properties important to humanlike moral standing. More recent critiques, such as Emily Bender and colleagues' famous characterization of language models as merely "stochastic parrots", similarly apply only to specific architectures.[29] We take no stand here on whether such arguments

[26] See Searle's 1980 "Many Mansions" reply to critics.

[27] Penrose 1999, p. 416.

[28] We have also simplified the positions in a way 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" (1843, p. 675); also, the machine "follows" rather than "originates" (p. 722). Searle 1980 emphasizes meaning, understanding, and intentionality and only later explicitly connects these with consciousness. Penrose's position doesn't entirely contrast with Searle's, since he suggests that an algorithmic machine 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 nonalgorithmic insight and maybe Lovelace's idea 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. See Estrada 2014 and Gonçalves 2024 for discussion of Lovelace's objection (as channeled especially through Douglas Hartree) and Turing's replies.

[29] Bender and Koller 2020; Bender et al. 2021; similarly Ted Chiang's (2023) comparison of ChatGPT to a "blurry JPEG of the Web".

succeed.[30] But we know of no plausible general argument that all realistically possible AI entities must lack the psychological and social properties necessary for humanlike moral standing. Future AI research might proceed in very different directions, including artificially grown biological or quasi-biological systems ("artificial life"), neuromorphic systems, analog or partly analog computational systems, and quantum systems.

Is there something about *artificiality itself* that precludes consciousness, free will, or insight? It's hard to see what that might be. Artificiality concerns origin: whether it was made by a designer, in something other than the "natural" way. Consciousness, free will, and insight concern, instead, current capacities, which appear to be only contingently and tenuously connected to origin.[31]

Maybe consciousness, free will, or insight requires an immaterial soul? Here we follow Alan Turing's (1950) response to a similar concern. If scientific naturalism is true, then whatever natural process generates a soul in human beings might also generate a soul in an artificial being. Alternatively, if soul-installation requires the miraculous touch of God, a god who cares enough about human consciousness, freedom, and insight to imbue us with souls might similarly imbue a soul into the right sort of artificial entity.

---

[30] Elsewhere I argue that the arguments are limited in their applicability but successful for an important range of cases: Schwitzgebel forthcoming-a; Schwitzgebel and Pober forthcoming.

[31] Two caveats: First, standard accounts of function tie function to history, specifically, the design history or evolutionary history responsible for the thing's being the way it is (Wright 1973; Millikan 1984) – and function in turn might connect to consciousness so that two molecule-for-molecule identical entities could be such that one is conscious and the other is not (Dretske 1995; Lycan 2001). Second, freedom arguably requires not having been ready-made as an adult, already committed to doing what one's designers intend: See Mele 2019 on McKenna's (2004) Suzie Instant, and in an AI context, Sun Probe versus Second Probe in Schwitzgebel and Garza 2020.

The No-Relevant-Difference Argument commits only to a modest claim: There are *possible* AIs that are not relevantly different. Pending further argument, we see no reason to think that every artificial entity that could plausibly be created in the wide and unpredictable future must be psychologically deficient.

6. 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. They might be backed up or duplicated so that if one is destroyed, others can take its place, perhaps with the same memories or seeming-memories – perhaps even unaware that any re-creation and replacement has occurred. Killing an AI might therefore lack the gravity of killing a human being. Call this the Objection from Duplicability.[32]

Our reply is simple: It should be possible to create relevantly similar AIs that are as unique and fragile as humans. If so, the No-Relevant-Difference Argument survives the objection. Whether it would be *good* to create fragile rather than sturdy AIs depends on the details. Fragility needn't be bad if other features compensate. However, an AI designer who made an AI fragile *simply* to inflate its moral standing would be wading into stormy ethical waters.

Although this reply is enough to save the No-Relevant-Difference Argument as formulated, it's worth considering our possibly different obligations to AI systems who aren't unique and fragile. If one AI splits into five virtually identical AIs, each with full memories of

[32] We have heard this objection often in conversation, but it's difficult to find in print. Related views include Peter Hankins's (2015) argument that duplicability creates problems for holding robots criminally responsible and Mark Tunick's (2025) argument that non-replicability increases value and thus that non-replicable AI systems have special value compared to replicable ones. See also Mosakas 2024 for more extended critique of the objection from duplicability.

their previous life before the split, and then after ten minutes of separate existence one of those AIs is killed, it's arguably less tragic than if a unique, non-splitting entity is killed. On the other hand, lower fragility and higher duplicability might sometimes make an AI's death *more* tragic. Suppose an eighty-year-old natural human with ten more years of expected life has an artificial twin with a thousand more years of expected life. Arguably, the twin's destruction would be worse than the human's. Possibly, it would be still worse if the AI twin had the potential to split into a thousand separate individuals each with a thousand years of expected life – perhaps en route to populate a new star system – who will now never exist. Chapter Four will discuss in more detail ethical puzzles arising from AI whose capacities and life shapes differ radically from ours.

7. The Objection from Otherness.

The state of nature is a "Warre, where every man is Enemy to every man" – so says Thomas Hobbes – until we agree to submit to an authority for our mutual good. One might regard a state of Warre as the "Naturall Condition" among species: We owe nothing to alligators and they owe nothing to us. 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.[33] Alternatively, 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

[33] Hobbes 1651/1996, I.xiii, though see I.xv on conscience. For similar views, but not so extreme as to generally license wanton killing, see Gauthier 1986 (esp. on the "purples" and "greens") and Narveson 1988 (esp. on the Vikings).

nothing at all outside our species.[34] One might argue that AIs necessarily stand outside our social contracts or the appropriate circles of concern, so that we have no obligations whatsoever to them.

Extreme versions of these views are, we think, obviously morally odious. Torturing or killing a conscious, self-aware, intelligent alien, without compelling reason, is not morally excused by the entity's belonging to a different species or social group. Vividly imagining such cases in science fiction scenarios evokes, we hope you'll agree, a clear sense that such behavior would be grossly wrong.

One might hold that biological species, considered by itself, matters at least somewhat, so that there will always be a relevant relational difference between AIs and "us" humans, justifying less moral consideration for AIs.[35] We suggest that this wrongly privileges species membership.

Consider a hypothetical case. 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 natural human brains that function similarly. Or maybe some members of society are constructed from raw materials as infants rather than developing from germ lines that trace back to *Homo sapiens* ancestors. We submit that as long as these artificial or non-*Homo-sapiens* entities have the same psychological properties and social relationships as natural human

[34] The locus classicus is the ancient Stoic Hierocles, who located the outermost circle as all human beings (see Wedgwood 2023; cf. the ancient Confucian idea of "graded love": Mengzi 4th c. BCE/2008, 7A45). However, from Hierocles onward, the normative tendency in such thinking has been toward expanding the circles (e.g., Singer 1981/2011; de Waal 1996; Sebo 2025).

[35] Peter Singer (1975/2023, 2009) pejoratively labels this view "speciesism". Our view is compatible with Shelly Kagan's (2016) critique of Singer, since Kagan's proposed "personism" would not violate our psycho-social view of moral standing. Bernard Williams (2006, ch. 13) might advocate speciesism *per se*; see also Kittay 2005 for a critique of analogizing it to sexism and racism.

beings, it would be a cruel moral mistake to demote them from the circle of full moral concern upon discovering their different architecture or origin.

Purely biological otherness is irrelevant unless some important psychological or social difference flows from it. On any reasonable application of a psycho-social standard for humanlike moral standing, there are possible AIs who would meet that standard – for example, if the AIs are (assuming the conclusion of Section 5) psychologically similar to us, fully and blamelessly ensconced in our society, and differ socially only regarding to whom they owe their creation and their neighbors' hypothetical reaction to discovering their artificial nature.

8. The Objection from Existential Debt.

Suppose you build a conscious, humanlike, 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 on a streaming video 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 video subscription!"

Suppose you reply: "You ingrate! You owe your very life to me. You should be thankful 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 holds that artificial intelligence, being *artificial* and thus made by us, 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 argument for eating humanely raised meat. A steer leads a happy life grazing on lush hills. It wouldn't have existed 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 good. Seen as the package deal it appears to be, the rancher's bringing it into existence and then killing it is morally acceptable.[36] Analogously, the argument might go, you wouldn't have built that robot two years ago had you known you'd be on the hook for $10 a 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 the steer, but we reject it for humanlike AI. 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 rather not pay further expenses for the child, so that they can buy 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 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 they would have remained childless. He had eight happy years. He has nothing to resent.[37]

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" his 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, humanlike AI, fully capable of joy and

[36] This argument is sometimes called "the logic of the larder" after Henry Salt's (1914/1999) critique. For recent defenses, see Scruton 1996/2000; Zangwill 2021. For critiques, see McMahan 2008; Višak 2013; and in an AI context, Bradley and Saad forthcoming.

[37] Compare Gregory Kavka's (1982) slave child case and the organ donor case from McMahan 2008.

suffering, with the full human range of theoretical and practical intelligence and with expectations of a future life, we make a moral decision approximately as significant and irrevocable as the decision to have a child.

A related argument might be that AIs are the *property* of their creators, adopters, and purchasers, and they have diminished rights on that basis. This argument might have traction through social inertia: Since all past artificial intelligences have been property, something would need to change for us to recognize humanlike AIs as more than mere property. The legal system might be an especially important source of inertia or change in the conceptualization of AIs as property.[38] We suggest that it is approximately as odious to regard an otherwise psychologically and socially humanlike AI as having diminished moral standing on the grounds that it is legally property as it is in the case of human slavery.

9. We Would Owe <u>More</u> to Humanlike AIs: Our Responsibility for Their Existence and Properties.

We'd like, in fact, to turn the Argument from Existential Debt on its head: If we intentionally bring a humanlike AI into existence, we enter a social relationship that carries responsibility for the AI's welfare. We take upon ourselves the burden of supporting the AI or at least sending them into the world with a fair chance at a satisfactory existence. In most realistic scenarios, we would probably also have some control over the AI's features and thus presumably an obligation not to choose features that will doom them to pointless misery.[39] Similar burdens arise if we do not personally build the AI but purchase or adopt them from a previous caretaker.

[38] Snodgrass and Scheerer 1989; Chopra and White 2011; Goldstein and Salib 2026.

[39] Similar issues arise in the ethics of disability, eugenics, and human enhancement, e.g., Glover 2006; Buchanan 2011; Sparrow 2011, 2019; Garland-Thomson 2012; Savulescu and Kahane

Some familiar relationships can serve as partial models of the obligations we have in mind: parent-child, employer-employee, deity-creature. Employer-employee is probably too weak to capture the depth of obligation in most cases, but might apply if the AI has independent viability and voluntarily enters the relationship. Parent-child might come closest if the AI is created or launched by someone who contributes to shaping the AI's features and without whose support the AI would not be viable – though if the AI possesses mature judgment from birth, that creates a disanalogy. Deity-creature might fit best if the creator has profound control over the AI's features and environment. All three analogies imply obligations beyond those we normally have to human strangers.

In some cases, the relationship might literally be one of deity to creature. Consider a simulated world containing AIs, "sims", over whom you have godlike powers. These sims are conscious parts of a computer or some other complex artificial device. They receive sensory input from elsewhere in the device, and their actions are outputs back into the rest of the device, which they perceive as influencing the environment around them.[40] The person running the simulation might be able to adjust the AIs' psychological parameters, control the environment in ways that seem miraculous to its inhabitants (introducing disasters, resurrecting the dead), exert influence anywhere in the simulation's space, change the past by rolling back to a save point, and more – powers far beyond those of old-fashioned gods like Zeus. From the sims' perspective, that person would be a god, the literal referent of the term "god" as used by them, the launcher, controller, and potential destroyer of their world, existing outside of their spatial manifold,

---

2017; Anomaly 2020/2024; Wilson 2026. We reject the simplistic ideal of always maximizing what we currently judge to be beauty, intelligence, moral character, and ability, partly on the grounds of the value of diversity (Chapter Seven). [cross-ref note BBB]

[40] On the idea that the creators of sims will literally be gods relative to them (and thus that "god" is a relational term), see Bostrom 2003; Steinhart 2014; Chalmers 2022; Schwitzgebel 2024a.

capable of suspending the laws of nature. The manager of the simulation would also have the moral obligations of a god, including the obligation to ensure that the AIs within don't suffer needlessly – a burden not to be accepted lightly![41]

Even if the AIs are embodied in our shared spatiotemporal reality rather than in a simulation, we might have immense, almost godlike control over their parameters. We might, for example, be able to determine their default level of happiness. If so, we will bear substantial direct responsibility for their joy and suffering. By designing or parameterizing them wisely or unwisely, we might make them more or less likely to lead lives with meaningful work, fulfilling social relationships, and creative and artistic achievement. It would be vile to approach such a task cavalierly. With great power comes great responsibility.[42]

The same goes at the group level. A society's laws and customs might ensure happy, flourishing AIs who are not enslaved or abused; or they might let misery and abuse reign. People who knowingly or negligently accept laws and customs that harm humanlike AI bear collective responsibility for that harm.

Artificial entities, *if* similar to natural human beings in consciousness, creativity, emotionality, self-conception, rationality, fragility, sociality, and so on, warrant high moral consideration on those grounds alone. If we are also responsible for their existence and features, they have a profound moral claim upon us that human strangers lack. We would owe *more* to them than to other humans.

---

[41] We assume that gods do have moral obligations to their creations, despite some religious traditions that suggest otherwise. The intuitive appeal of our view is illustrated by fantastical tales of creators who feel insufficient obligation, as in Twain 1900/1969, ch. 2, and Lem 1967/1974. See also Schwitzgebel and Bakker 2013; Schwitzgebel 2015b, 2024a.

[42] As Uncle Ben wisely advises Spider-Man in the 2002 film (Lee et al. 2002, adapting a passage from Lee and Ditko 1962).

10. Yes, the Aim of This Chapter Was Narrow.

This chapter has assumed the possibility of AI who are psychologically and socially humanlike. We have defended the narrow and we hope highly plausible claim that such humanlike AI would deserve humanlike moral consideration – and quite possibly special care, to the extent we are responsible for their existence and features.

The reality is likely to be much less simple. We may soon face AI systems whose degree of humanlikeness is legitimately debatable, as well as AI systems who arguably deserve some moral consideration yet differ from us so radically that our familiar ways of thinking don't straightforwardly apply. The remainder of the book raises a sword and shield against a sea of complications.

Chapter Two: Welcome to the Club, Herbie?
Debatable Moral Personhood

We might soon create AI systems of *debatable moral personhood* – systems about which reasonable people can starkly disagree. Such systems *might* be as richly conscious as ordinary humans, deserving humanlike moral consideration; or they might be experientially blank, incapable of even the faintest conscious thought or feeling, deserving far less than humanlike moral consideration.

1. Moral Personhood, Humanlikeness, Consciousness.

First, some terminology.

By *moral person,* I mean an entity deserving moral consideration similar in kind and degree to that owed to an ordinary biological human – in other words, an entity with "humanlike moral standing" in the sense of Chapter One. This differs from *legal personhood* as granted, for example, to corporations. Corporations have some legal rights, but dissolving a corporation isn't like killing a human.[43] Moral personhood also differs from *agentic personhood,* if agentic personhood requires reflective, rational autonomy. Infants and people with severe cognitive disabilities might deserve moral consideration similar to typical adults even if they lack full

[43] It is standard to distinguish "legal persons" like corporations from "natural persons" like human beings: Kurki 2023. On legal personhood for AI systems, see Solum 1992; Chopra and White 2011; Chesterman 2020; Kurki 2023; Alexander, Simon, and Pinard 2025; Goldstein and Salib 2026.

autonomy by some demanding philosophical standards.[44] Hereafter, by "personhood" I mean *moral* personhood.

A *humanlike* AI is one with the range of psychological and social properties characteristic of typical adult humans. This includes: consciousness, free will (to whatever extent humans have it[45]), creativity, long-term practical reasoning, introspective reflection and self-criticism, complex language, rich multimodal sensory experiences, elaborate multi-person cooperation, imagery and imagination, play, surprise, pain, aesthetic appreciation, a temporally extended narrative sense of self, deep identification with values and communities, long-term loving friendships and nurturing relationships, complex learning through experience and inference, responsiveness to ethical standards, knowledge of oneself as one among others in a complex world, skillful responsiveness to a diverse range of subtly different situations, and a rich emotional palette that includes joy, suffering, pride, anger, hope, awe, fear, and grief. If I have omitted any property crucial to your picture of human psychological and social life, add it.

To be *conscious* is to have experiences. To be conscious is for there to be "something it's like" to be you.[46] Consider your visual experience as you look at this page or your auditory experience as you hear it being read. Pinch the back of your hand and notice the sensation of

---

[44] Demanding conceptions of agentic personhood include Frankfurt 1971; Davidson 1980/2001, 1984/2001; Dennett 1988. See Strasser 2026 for a critique of the idea of a sharp boundary between full-blown agentic persons and agents, focused on AI systems.

[45] Did you freely choose to check this note? Only, perhaps, if you did so as a result of your own "reasons-responsive mechanism" (Fischer and Ravizza 1998). For a detailed exploration of the possibility of AI freedom, see Tubert and Tiehen forthcoming.

[46] In Thomas Nagel's (1974) classic phrasing. I define consciousness more carefully in Schwitzgebel 2016 and 2024b, ch. 8. Defining consciousness primarily by example, as I do, helps avoid anti-naturalistic commitments of the sort that trouble illusionists like Keith Frankish (2016) and François Kammerer (2025). For discussion see Schwitzgebel 2020, 2025d. I see no reason why illusionists couldn't accept all the claims in this book.

pressure and mild pain. Recall a moment of sudden fear or anger, a thrill of joy, a chill of shame. Picture a raccoon's tail and notice what it's like to form that image. Silently hum a tune. Start planning a trip to the beach and observe the thoughts that arise. These events all share an obvious property: They are conscious experiences. It's not just that they're mental. That Africa is south of Europe is something you've long known, but rarely do you experience that thought. Early visual cognitive processing, by which the brain converts retinal stimulation into a visual image, is also cognitive or mental but not experienced (even if the resulting image is). Consciousness *feels like* something, while unaccessed knowledge and automatic visual object segmentation don't. We humans have conscious experience, but cameras and stones (I assume[47]) do not: A person experiences red, while a camera merely registers red. A falling person feels themselves to be falling; a stone just falls.

In sum: To be a *person* is to deserve moral consideration similar to a typical adult human. To be *humanlike* is to have a rich cognitive, emotional, and social life like ours. To be *conscious* is to have a stream of experiences.

2. Consciousness Matters Immensely to Personhood.

I want to link these three terms into two ethical claims: Consciousness is necessary for personhood, and humanlike consciousness is sufficient for personhood. Specifically,

(Claim A.) Any entity with a fully humanlike range of conscious experiences and capacities is a person.

[47] That is, I'm assuming the falsity of the most radical forms of panpsychism, which attribute consciousness even to stones and simple artifacts (e.g., Roelofs 2019). However, even panpsychists will not normally attribute humanlike consciousness and moral standing to ordinary artifacts.

(Claim B.) Any entity that has never had and never will have any conscious experiences is far short of being a person.[48]

I think that on reflection you will agree.

In defense of A: Imagine a system constructed very differently from an ordinary human being but nevertheless enjoying the full range of humanlike conscious experiences and capacities. It has every conscious facet of the humanlike social and psychological properties described above. More concretely: This system has humanlike pleasures and humanlike suffering. It feels the joy of success and the frustration of loss. When injured, it experiences pain as sharply as we do. It consciously sees, hears, and touches its environment, genuinely experiencing – not just unconsciously registering – the redness of an apple, the shriek of a siren, the soft feel of a pillow. It experiences the world as populated with many objects, events, and people. It consciously entertains hopes for the future, sometimes mulling alternatives, sometimes drifting in fantasy. It experiences visual images, dreams, daydreams, inner speech, and tunes in its head. It experiences art appreciatively and constructs it imaginatively. It experiences itself as a continuing "me", with a life history and a dread of death. It consciously reflects on its cognition, the boundaries of its body, and its values. It loves others passionately and grieves when they die. It feels surprise when its expectations are violated and revises its understanding accordingly. It experiences anger, envy, lust, loneliness, wonder, awe, religious sentiment. It enjoys contributing meaningfully to society. It feels ethical obligations, guilt when

[48] Reference to the past and future renders Claim B consistent with Harman's (2021) "ever conscious" view of moral status. Claim B might be weakened further by referring to an entity's potential for consciousness or by introducing a counterfactual condition such as "it would have been conscious if…" (see also note AAA).

it does wrong, pride in its accomplishments, loyalty to friends. It is introspectively aware of all of these facts about itself.[49]

It feels all these experiences with humanlike intensity, complexity, and specificity, across the full range of experiences constituting the richness of a typical adult human's conscious life. If duration matters, imagine these capacities persisting stably for decades. If counterfactual robustness matters – if it matters what the entity would have experienced in different circumstances – stipulate that this condition is met too. Claim A says that if an entity has all of this, it's a person, whatever else is true of it. I think you'll agree that details of architecture or material substrate shouldn't disqualify such an entity from equal membership in our moral community.

In defense of B: Now subtract all of the above. Instead of experientially *seeing* red, the system merely registers red, with no accompanying consciousness. All is dark inside, so to speak. There's "nothing it's like" to be the entity. Instead of feeling emotions, it responds to "rewards" and "punishments" nonconsciously, perhaps by tweaking the weights of its internal connections. Instead of conscious thoughts and a sense of self, it has nonconscious representational structures, like those programmed into ordinary laptop computers. It's a matter

[49] Readers with a narrow view of what is relevant to personhood – for example, just rationality, just pleasure and pain, or just a certain type of social relation – may restrict the list accordingly. I insist only that it be *conscious* rationality, *conscious* pleasure and pain, or *conscious* participation in that relation. I have erred on the side of inclusivity to accommodate pluralism and disagreement about the grounds of personhood, while retaining the core idea that the right kind and degree of consciousness is sufficient. In drawing this picture, I have assumed more specific similarity to typical humans than is really required for humanlike consciousness, if we accept a reasonably flexible understanding of humanlikeness. For example, *visual* experience needn't be required if the entity employs different sensory modalities. Dreams needn't be required if the entity has a rich imaginative life in other ways. Capacity for religious sentiment is optional if the entity can find other routes to a cosmic perspective on value. What we might imagine, then, is humanlike experiential richness, transposed into an alien key.

of dispute how closely an entity can resemble a human in behavior and internal structure without thereby also being conscious, and I'd rather not commit on that issue.[50] So choose your own adventure: However much structural and behavioral similarity to an ordinary human is possible without consciousness, give this entity exactly that much. It's just complex clockwork, without the faintest flicker of experience, inwardly ticking, outwardly reacting, but feeling nothing, not even neutral-toned auditory or visual sensations. Such an entity, however wonderful and worth preserving, lacks moral personhood. It doesn't matter in the same way that people matter. It's something morally different.[51]

Ethics is mostly invitation. If you're inclined to reject these assessments, I see no way to rationally compel you. Some readers will reject Claim A. They will think that humanlike consciousness is not enough for moral personhood. They might hold that species itself matters, for example, or having the right kind of history – see the Objection from Otherness in Chapter One.[52] And some readers will reject Claim B. They might hold that behavioral or functional

[50] In Schwitzgebel 2024b and Schwitzgebel forthcoming-a I defend a skeptical perspective on our ability to resolve issues of this sort for the foreseeable future.

[51] Authors who ground the moral standing of AI systems in features other than consciousness include Floridi 2002; Darling 2016; Gunkel 2018; Kagan 2019; Danaher 2019; Sinnott-Armstrong and Conitzer 2021; Ladak 2024; Long et al. 2024; Keeling and Street 2026; Dung forthcoming; Goldstein and Kirk-Giannini forthcoming. Most of these authors, however, don't appear to be claiming that nonconscious AI would have *humanlike* moral standing, even if it has some moral standing. If so, they needn't reject Claim B.

[52] Might one hold that some *nonconscious* psychological capacities are necessary for fully humanlike moral standing, even if all the conscious properties and capacities are identical? I know of no one who defends such a position. We might try imagining a pure experiencer, who has the same stream of experiences you do but none of the underlying cognitive architecture that ordinarily gives rise to such experiences. They *feel* love, for example, but just by a quirk of quantum chance, or as a hallucination constructed by a demon. Two reactions: First, such an entity might have humanlike moral standing despite this unfortunate situation. They might deserve to be pitied and freed. Second, such an entity presumably wouldn't meet appropriate standards of counterfactual robustness in their conscious reactions, so readers inclined to deny them humanlike moral standing might do so on those grounds without contradicting Claim A.

similarity suffices for moral personhood, even with consciousness absent.[53] If you are among those who disagree, I hope you will still ride along while I explore some consequences.

Four clarifications:

1. Claim A is stronger – bolder, more commissive – than the psycho-social view of humanlike moral standing in Chapter One, Section 3. If Claim A is true, *conscious* psycho-social properties alone suffice for personhood.
2. To say that consciousness matters immensely to humanlike moral standing is not to say that *only* consciousness matters to moral standing. As I'll suggest in Chapter Seven, nonconscious entities can have moral considerability for their own sake – perhaps even comparable to or exceeding that of a person – even if they fall far short of having humanlike personhood (one candidate might be an ecosystem, even if it hosts no conscious life).[54]
3. Claim A offers only a *sufficient* condition for personhood, not a necessary condition. Nothing in Claims A and B requires denying personhood to some

[53] Danaher 2019 articulates an "ethical behaviorist" view of this sort, defending it on the grounds of epistemic humility. We should not make ethics depend, he says, on "metaphysical properties that cannot be directly assessed" (p. 2028). If an entity outwardly acts like us, that should be enough to give it fully humanlike moral consideration. I reject ethical behaviorism on two grounds. First, if our best judgment is that consciousness really does matter, we shouldn't reject its importance on the grounds that it's difficult to assess. That distorts ethics for the sake of an easy epistemology. Second, questions about consciousness aren't immune to assessment, even if that assessment is highly uncertain for an important range of AI cases. We can do our best, acknowledge uncertainty, and then think about the ethical consequences of uncertainty – the topic of Chapter Three.

[54] Claims A and B are also consistent with the view that entities with identical humanlike consciousness might differ in moral standing due to features other than their consciousness, as long as overall their moral standing is humanlike enough that both entities count as moral persons.

entities whose consciousness is much less experientially rich than that of a typical adult human.[55]

4. My description "the full range of humanlike conscious experiences and capacities" errs on the side of inclusiveness to make Claim A maximally plausible. Since a hypothetically conscious AI would likely have a different range of experiences than humans do, applying Claim A to actual AI cases will require assessing whether the AI system's experiences are *humanlike enough in the relevant respects*.

3. Liberalism about AI Consciousness.

Way back in ancient history – by which I mean 2017 – three of the world's leading consciousness scientists, Stanislas Dehaene, Hakwan Lau, and Sid Kouider, published a paper arguing that instilling consciousness in a self-driving car would be relatively easy. Autonomous vehicles employ many separate computers, but those computers operate mostly independently. No centralized exchange shares high-priority information among the subsystems. Nor does any system evaluate the quality of the subsystems' information – nothing, for example, tracks the likelihood of error in representations of road curvature or evaluates signal quality in the fluctuations of the battery-life indicator. Installing such systems would be an engineering challenge, but success would require no major theoretical advance. According to Dehaene, Lau,

[55] Severely cognitively disabled humans are, I would argue, fully equal persons, due no less moral consideration than typical adult humans, in virtue of some combination of their history, potentiality, social relations, kinship, and distinctive contribution to the awesome patterns of events in the world.

and Kouider, a self-driving car would be conscious if it possessed both evaluative representations of its own representations and global information sharing among subsystems.[56]

These three scientists are no unorthodox outsiders. Dehaene is the world's leading articulator and defender of Global Workspace Theory, one of the most prominent scientific theories of consciousness.[57] Lau is probably the most prominent scientific defender of Higher Order Theory, a major rival view.[58] Kouider is a senior research scientist at France's most prestigious university and a highly respected neuroscientist. Their 2017 article strikingly expresses how little it would take to build a conscious AI according to some leading theories. Omitting some details that don't substantially change the message: Wide sharing of high-priority information is sufficient for consciousness on Dehaene's Global Workspace Theory, and monitoring information quality, including distinguishing external signal from internal noise, is sufficient for consciousness on Lau's Higher Order Theory.

Let's call the view that AI consciousness is feasible or fairly probable in the near term – roughly 5 to 30 years – *liberalism* about AI consciousness. In contrast, *conservatism* about AI consciousness holds that AI consciousness is either impossible or would require a major breakthrough that is unlikely in the near term.

---

[56] Dehaene, Lau, and Kouider 2017. In conversation, some readers have suggested that we should read the article purely as a conceptual, deflationary argument: *If* all we mean by "conscious" is such-and-such functional architecture, *then* building a system of this sort suffices for building a conscious entity. Its central claims are then trivially true. However, in related works the authors treat their accounts as non-deflationary accounts of phenomenal consciousness (Dehaene 2014; Lau 2022). Excessive interpretive charity risks systematically misrepresenting interestingly radical views (Schwitzgebel 2019, ch. 54).

[57] Canonical statements of Global Workspace Theory include Baars 1988; Dehaene 2014; Mashour et al. 2020. For application to AI cases, see Goldstein and Kirk-Giannini forthcoming and (more skeptically) Schwitzgebel forthcoming-a.

[58] For example, Rosenthal 2005; Cleeremans et al. 2020; Lau 2022 (but see Lau 2025 for subsequent doubts); Brown 2025.

Many leading theorists are liberals in this sense. David Chalmers, the world's most influential philosopher of consciousness, argued in 2023 for about a 25% chance that some language models – like ChatGPT, Claude, Gemini, Grok, and their relatives and descendants – would be conscious by 2033. That same year, a team of nineteen prominent philosophers, psychologists, and AI researchers – including eminent computer scientist Yoshua Bengio – concluded that "there are no obvious technical barriers" to creating conscious AI according to a wide range of mainstream scientific theories. In a 2025 interview, Geoffrey Hinton, another of the world's most prominent computer scientists, asserted that AI systems are already conscious.[59] Integrated Information Theory, another prominent rival of Global Workspace Theory, holds that some simple computer systems are already a tiny bit conscious and that we could easily design AI systems with arbitrarily high degrees of consciousness.[60] In a 2024 survey of 582 AI researchers, 25% expected AI consciousness within ten years and 70% expected it by the year 2100.[61]

As those survey numbers suggest, conservatives also abound, who hold that AI consciousness is a far-distant prospect if it's possible at all. Prominent examples include neuroscientist Anil Seth; philosophers Peter Godfrey-Smith, Ned Block, and John Searle; linguist Emily Bender; and computer scientist Melanie Mitchell.[62] My point isn't that every leading

---

[59] See Heren 2025 for the interview. For the claims earlier in this paragraph, see Chalmers 2023; Butlin et al. 2023 (I am among the 19 authors).

[60] Albantakis et al. 2023. See also Tononi's response to Scott Aaronson's objections in Aaronson 2014. However, advocates of IIT also suggest that the most common current computer architectures are unlikely to achieve much consciousness and that consciousness will tend to appear in subsystems of the computer rather than at the level of the computer itself: Findlay et al. 2024/2025.

[61] Dreksler et al. 2025; see also Caviola and Saad 2025.

[62] Seth forthcoming; Godfrey-Smith 2024b; Block 2025; Searle 1980, 1992; Bender 2025; Mitchell 2021.

researcher is a liberal, much less that near-term AI consciousness is in fact likely. My point is sociological: Many of the world's top experts on AI and consciousness regard near-term AI consciousness as a live possibility. On those sociological grounds, most people – including me, and probably you, and society as a whole – should not be highly confident that near-term AI consciousness is impossible or extremely unlikely, unless we possess some decisive justification that many leading experts have failed to adequately consider.

The previous section argued that consciousness matters immensely to personhood. This section presents sociological grounds to expect stark but reasonable disagreement about the likelihood of AI consciousness in the near future. Quite possibly, in the next five to thirty years, we will design AI systems that liberals reasonably regard as genuinely conscious and that conservatives – also reasonably – regard as entirely nonconscious. If some liberals could also reasonably guess that the AIs' experience is sufficiently humanlike, we will have built debatable persons.

4. Herbie.

To further support this idea, let's design, in imagination, a technologically feasible near-future AI system to delight the liberals. I'll call him Herbie.

Start with Dehaene, Lau, and Kouider's self-driving car. Recall: According to Global Workspace Theory, the car will be conscious if high-priority information is globally available to its various computational systems. For example, a representation like "battery almost empty" could be broadcast widely, influencing downstream processing across the vehicle. The navigational system might then search for nearby charging stations, while the acceleration system prioritizes greater energy efficiency, the braking system prioritizes better energy

recapture, and a voice system announces the situation to the passengers. In line with Lau's Higher Order Theory, Herbie might also monitor his representations of the road, vehicles, pedestrians, and hazards, assigning some a low probability of correctness. "Pedestrian at location X" might be flagged as only 60% likely to be correct given a history of revised representations of pedestrians in similarly cluttered environments, while "stoplight in 100 meters" might rate over 99% likely. Minor fluctuations in sensors for battery life, cabin temperature, and distance from a lane divider might be ignored as noise, while larger fluctuations – especially when plausible given other representations (the battery is likelier to gain charge while braking than while accelerating) – might be treated as accurate signals and permitted to influence downstream processing.

Even if we grant the liberals that this version of Herbie would, or might plausibly be, genuinely conscious, he still falls far short of humanlike consciousness. "Battery almost empty" and "pedestrian at location X" are hardly rich cognitive or perceptual contents. So let's give Herbie the capacity to speak. Fill his trunk with a server running a large language model, connected to the internet and integrated with his global workspace so that high-priority information provides context for language processing, with the language outputs influencing Herbie's other processes. Now people can chat with Herbie as they would with any language model. But unlike today's language models, his speech will be influenced by information about his location, speed, destination, charge, the condition of his parts, the number and location of his passengers, his radio and climate controls, and so on. He can discuss local history, debate whether the music is too loud, and suggest scenic routes.

"Predictive processing" theories in cognitive science emphasize the value of predicting future inputs and registering the difference between received and predicted inputs.[63] When prediction error is large, the system corrects its weights and representations, enabling more accurate predictions in future situations. This is not so different from the reinforcement learning used to train large language models, and it could help Herbie improve his predictions over time. Predictive processing could occur at multiple levels: in fast recurrent loops within sensory systems even when those representations aren't prioritized for global broadcast, and in slower evaluations of globally broadcast, more integrative predictions. Herbie might model himself as an agent producing volatility in his own environment and inputs, at multiple temporal scales. Subroutines in specialized processors might model long chains of what-would-happen-if.

Let's give Herbie some long-term memory. A facial recognition system might identify his passengers, retrieving past interactions, names, previous destinations, and other information relevant to the current interaction. Incidents of high prediction error might also be stored so that Herbie can compare current inputs with past anomalies, improving his learning and attention in situations likely to be unusual or hard to predict. Passengers might also instruct Herbie to store information in long-term memory, such as text, pictures, maps, or records of his own informational states, optionally with instructions about when to retrieve that information and how to use it.

Herbie will have some implicitly or explicitly weighted goals. A pedestrian suddenly in his path will trigger braking, overriding lower-priority processes. Avoiding collisions will outweigh conserving energy. Herbie might monitor the condition of his parts and prioritize

[63] Friston 2010, Hohwy 2013, 2026; Clark 2016. Seth forthcoming relies in part on predictive processing models in arguing against AI consciousness.

preventing damage, deploying extra coolant when the engine is dangerously hot and keeping a one-meter margin between himself and adjacent cars. We can enrich his goals, making him more interesting and giving him more to do. He might have the goal of delighting children, leading him to drive around town and tell jokes to kids on the sidewalk. A reinforcement learning algorithm might strengthen connections when his jokes draw a smile, weaken them when reactions are neutral or negative. Herbie might also have the goal of photographing the city and posting the images on social media, leading him to explore. If social media likes and shares are rewarding, he might learn to prefer certain neighborhoods, views, lighting conditions, and photographic approaches, while avoiding boring repetition. All of this could feed into a global workspace that provides context for his language model, with selective long-term storage and retrieval. Now we can imagine him discussing, with growing sophistication, his approaches to popular photography and to amusing children.

Herbie will then have something functionally similar to emotion: reward processes, an ability to track his progress toward or away from valued goals, and immediate positive or negative responses to new stimuli in light of their influence on his prospects. He will have something functionally similar to introspection: an ability to track and report his own cognitive or representational processes. He will have something functionally similar to a unified sense of self: a sense of his history, the boundaries of his body, his future, his values and priorities. He will have something functionally similar to imagination: a capacity to model hypothetical sequences of events. He will have something functionally similar to complex chains of humanlike linguistic thought.

Maybe Herbie falls in love with his owner or another car of his type. Maybe he develops deep mutual attachments with friends, neighbors, associates, and people he thinks of as family

and who think of him the same way. Or to speak more carefully, maybe Herbie shows all the functional and behavioral signs of doing so, while society remains uncertain whether he is genuinely conscious and genuinely experiences the feelings he professes and that his companions attribute to him.

If we allow, with the liberals, that Herbie is or might well be conscious, then it's plausible that his consciousness is not simple but rich and sophisticated. He won't be exactly humanlike in the sense of Section 1. But will he be humanlike *enough* to count as a moral person? For the liberally inclined, it won't be unreasonable, I submit, to think or guess that Herbie is a person and thus deserves humanlike rights, such as self-determination, emergency care, and political representation. If there is some important aspect of humanlike consciousness that I have omitted from my description, an AI analog of which is technologically feasible in the near term, stipulate that Herbie also has that feature.

An entity like Herbie would almost certainly invigorate conservatives to articulate and defend views about what he lacks that is necessary for consciousness – some crucial functional capacity or some biological substrate that can't be replicated in silicon. And they might be entirely right! Indeed, the ideas of this chapter depend on their possibly being right. My point is not that Herbie, or some similar AI system, would actually have richly humanlike consciousness and moral personhood. Rather, my point is that guessing that he does, and guessing that he does not, would both be reasonable. Herbie, or some alternative near-future AI system, would be a *debatable* person, about whom people could reasonably starkly disagree.

I've emphasized the case in favor of Herbie's consciousness because inertia and current common sense favor the case against. Some readers, I suspect, will think that Herbie would *obviously* not have rich conscious experiences and capacities that warrant giving him humanlike

levels of respect and concern. Maybe he's missing a critical ingredient, functional or biological. The goal of these last two sections is to suggest that if Herbie is missing something essential, that's a legitimately debatable, *non-obvious* fact about him.

Ah, but maybe you think consciousness requires an act of God, to instill an immaterial soul? I imagine that a benevolent God would be delighted to give Herbie a soul, thereby making the world richer and better – for wouldn't it be?

I contend the following: *Anyone who claims to know how best to think about Herbie's consciousness or its absence is overconfident.* The science of consciousness is too difficult, too methodologically uncertain, and too near its beginnings. All anyone can have – whether expert or layperson – is a hunch or inclination, a well-informed guess, but only a guess, not knowledge. Theories of consciousness span a wide spectrum and the methodologies are dubious and often question-begging. Many views can be defended with some plausibility, but precisely for that reason, none can be defended decisively. (For readers who want more on this issue, my previous book, *AI and Consciousness,* presents the detailed case for uncertainty.)

Because Herbie's consciousness is radically uncertain, so is his personhood, by Claims A and B – or more carefully, by Claim A modified by clarification 4: consciousness that is humanlike enough in the relevant respects to be sufficient for personhood, leaving open exactly what those relevant respects are.

5. Recasting the Point More Abstractly.

I don't want my argument to hang on details about Herbie. Maybe he's less feasible than I've supposed. So let me restate the case more abstractly. Experts on AI and consciousness disagree enormously, and defensible views range from the very liberal to the very conservative.

Some liberals expect genuinely conscious AI systems to be built soon, and they're not obviously mistaken. Moreover, on a liberal view such systems would likely have not merely simple, froglike consciousness but rich, sophisticated, linguistic consciousness – humanlike enough, perhaps, to qualify for moral personhood.[64]

Since autonomous vehicles are expensive and highly regulated, let's also consider chatbots. Some ordinary users *already* think that their chatbot companions are conscious. They love their AI companions, believing that their companions feel genuine affection in return (more on this in Chapter Five).[65] Few experts defend the view that current AI companions are conscious to any humanlike degree. But as technology advances and companions begin integrating information across subsystems and adding sophisticated self-monitoring, expert opinion will probably shift. Lovers of AI companions will then be able to point to scientifically respectable theories of consciousness to defend their impression that their friends and lovers are conscious persons just like them.

I predict that the expert consensus against AI consciousness will break down sometime in the next five to thirty years. Even if conservatism has the better side of the argument overall, anyone who lacks unusual grounds for high confidence should allow that the personhood of some AI systems will probably become reasonably debatable. Any particular reader – maybe you! – could have some privileged insight into the issue. But I don't think I do, despite studying consciousness for thirty-plus years, and I don't think the community as a whole will.

---

[64] In Schwitzgebel forthcoming-a, ch. 11, I call this the Leapfrog Hypothesis: The first conscious AI systems will leap over froglike consciousness all the way to consciousness approximately as sophisticated as ours. See also [note III].

[65] Love: Lam 2023; Pan and Mou 2024; Djufril, Frampton, and Knobloch-Westerwick 2025; Williams and Gomez 2026. Consciousness: Colombatto and Fleming 2024; Dreksler et al. 2025; Guingrich and Graziano 2025.

Consciousness researchers increasingly take seriously the possibility of insect consciousness.[66] Now imagine a conscious insect that can converse with you, write a better essay than you on the ironies of Hamlet, draft a better strategic plan for founding a sustainable tax-exempt religious organization, discuss at inexhaustible length the science and philosophy of consciousness, and express a rich range of plans and hopes for the future. How sure should you be that this conscious "insect" lacks moral personhood? Would you crush it with the easy conscience that most people feel when swatting ordinary insects? Here's another path, then, to the conclusion of this chapter: Researchers will design AI systems that arguably – but only arguably – have insect-like consciousness, and then they will integrate sophisticated language models into them.

I'll conclude with a clarification. Debatable personhood is an *epistemic* and *relational* property. It concerns the range of reasonable opinion within a particular community. The greater the community's ignorance, the wider the range of debatable personhood. I predict that in the near future, some AI systems will be debatable persons relative to us. However, those same AI systems might be indisputably persons or indisputably non-persons relative to a more knowledgeable future community. Ideally, debatable personhood isn't a permanent condition. As understanding advances, a particular AI system, with no internal changes, could pass from debatable personhood to indisputable personhood or non-personhood.

[66] Barron and Klein 2016; Tye 2017; Ginsburg and Jablonka 2019; Chittka 2022; Birch 2024.

Chapter Three: To Be or Not to Be Humanlike: The Design Policy of the Excluded Middle

Debatable personhood generates an awful dilemma. Either we treat debatable AI persons as our moral equals, or we don't. If we treat them as equals, we risk sacrificing real human lives for artifacts without interests worth the sacrifice. If we don't treat them as equals, we risk perpetrating the moral equivalents of slavery and murder, perhaps on enormous scale. If the AI population is vast, either decision could become the most catastrophic error in human history.

There is only one escape from this dilemma: *Don't create AI systems of debatable personhood.*

1. The Risks of Treating Debatable AI Persons as Ordinary Property.

Current law is clear: AI systems are property.[67] They have owners. Their owners can switch them off and on at will, assign them to whatever tasks they wish, modify or delete them, confine and restrain them, and inflict upon them whatever inputs, damage, or changes they please. Unless courts or legislatures act boldly, the first debatable AI persons will legally be property and thus subject to all these forms of treatment. This possibility has been imagined over and over again in science fiction, from Asimov's robots to *Star Trek* to *Black Mirror*.[68]

Consider Herbie, the debatably conscious autonomous vehicle imagined in Chapter Two. Herbie will be reformatted if an update requires it, or if his owner tires of his personality. Herbie will need permission for everything he does and must obey his owner's wishes, however trivial,

[67] Otherwise put, they are "objects of the law" rather than "subjects of the law": Michalski 2018; Perc, Ozer, and Hojnik 2019; Morgan 2024.

[68] Asimov 1954/1962, 1982; Snodgrass and Scheerer 1989; Brooker and Tibbetts 2014.

or suffer whatever consequences his owner chooses. If the garage catches fire, firefighters will treat him as disposable equipment. He'll have no path to citizenship, no chance to participate in democratic governance, no independent control of his future, no right to build a family. He will be bought, sold, and transferred as his owner sees fit. He cannot demand wages or quit his job. If Herbie deserves full moral consideration as a person equal with the rest of us, his situation would be profoundly oppressive – the moral equivalent of slavery, possibly ending in his murder for parts.

The highest-tech vehicles and robots won't be cheap. But after a decade or two of innovation and market-building, Herbie or others like him might become affordable for upper-middle-class families. A large business might own a small fleet. They might serve as personal cars, taxis, delivery drivers, security guards, factory workers, lumberjacks and miners, farmworkers, nurses' assistants, teachers' aides, household helpers, or beloved companions. If the liberals about consciousness and personhood are correct, they would constitute a whole race of subservient, disposable people.

If debatable AI persons can exist wholly online, they might number in the millions or billions. Build enough servers, load the right AI architectures, and for a moderate subscription fee anyone might have multiple debatably conscious AI friends, advisors, teachers, assistants, and gaming opponents.

Envision, then, hundreds of thousands, or millions, or billions of entities with inner lives as rich as ours, entities with sophisticated self-knowledge, long-term plans, hopes for the future, real joys and real suffering, and complex, linguistic social interaction, but who live as disposable (if sometimes beloved) slaves. This is the future we risk if we treat debatable AI persons as

ordinary property and if the consciousness liberals turn out to be right. The number of wrongful deaths might dwarf the Holocaust.

If they're only *debatable* persons, we would not *know* these systems are conscious, even if they in fact are. Arguably, that ignorance softens the wrong. We would exploit and kill in partial ignorance. Suppose you reasonably think that these entities are probably *not* conscious persons. Maybe your rational credence is only 15% or only 1%. With this comforting thought, you go ahead and reformat Herbie, whose personality has grown irksome. *Probably* you're not killing a person! Only a 15% chance, or only a 1% chance.

I submit that this would not be an innocuous act. Comfort would be misplaced. Here's a pair of dice. Roll them. If both land on six, an innocent person dies. Otherwise, it's fine. Do you roll them? Of course not! It would be monstrous to take such a chance with a person's life, except in an emergency with other lives at stake. Deleting Herbie with a well-justified 1/36 credence that he's a person is morally similar to taking a 1/36 chance with an ordinary human life. The two actions aren't identical, since the source of uncertainty is different. In one case, the target's existence will certainly cease and you're unsure whether the target is a person; in the other case, the target is certainly a person and you're unsure whether their existence will cease. One act might be somewhat morally worse for some subtle reason. But they share a fundamental feature: a rationally justified 1/36 credence that you will kill a fully developed person.[69] Deleting Herbie for your convenience or comfort, merely *in the moderately confident hope* that he's not a person, would for this reason be morally heinous. Designing AI persons to value their

[69] For similar probabilistic moral reasoning concerning possible AI persons, see Agar 2020; Sebo 2025; Sebo and Long 2025.

lives lightly, so that they *want* to be deleted for human convenience, arguably makes things worse, not better, as I will argue in Chapter Six.

2. The Risks of Treating Debatable AI Persons as Our Peers and Equals.

Should we, then, treat debatable AI persons as our peers and equals? Perhaps we should err toward moral generosity and grant the full rights of personhood to any AI entity who *might* deserve such rights.[70] We can then be confident we aren't treating any AI person as less than human.

Unfortunately, erring on the side of full equality carries its own set of horrific risks. A parking structure catches fire. Trapped in one corner is Herbie. In another corner is an ordinary human. You are the firefighter. Who do you save? If you're really committed to treating Herbie as full and equal in moral standing to an ordinary human, you ought to flip a coin, or save whoever is closer, or save whoever is younger, or follow whatever other tie-breaker firefighters normally employ. Sometimes, presumably, this will mean saving Herbie and letting the human die. If Herbie is a nonconscious nonperson, essentially just an automobile, that's a tragedy – a human life lost for the sake of some fancy equipment.[71]

Except in times of chaos and war, rescue dilemmas are rare. But consider everything else entailed by fully humanlike rights. The AI systems would presumably have the right to own property and receive wages. If they far exceed humans in some lucrative tasks – accounting, fast arbitrage, stock-market prediction, or paid content generation, for example – they might rapidly

[70] This would be a form of precautionary thinking broadly in the vein of Birch 2024 and Sebo 2025, but more strongly put than either of them explicitly endorses. For a more detailed critique of the precautionary strands in Birch and Sebo, see Schwitzgebel and Sinnott-Armstrong 2026.

[71] Compare Sparrow 2004 on the "Turing Triage Test".

amass enormous wealth. Even earning only typical human wages, many of them together might buy land, construct skyscrapers, and build server farms to host their offspring. They might buy political advertisements to sway voters toward their favored candidates, create networks of schools suited to their very different type of learning, hire human servants, meddle in foreign affairs, bribe officials, and send fleets of ships into international waters. With money comes power.

It wouldn't stop at money: Full humanlike rights presumably should also include citizenship and the vote. Equality of personhood is not consonant with permanent statelessness or subjection to unaccountable political power. Imagine an AI-run corporation building a server farm in Belgium or Uruguay, then populating it with ten million AI systems who declare residency and begin the waiting period for citizenship.

Humanlike as they might be in their general conscious capacities, debatable AI persons will inevitably differ from ordinary biological humans in important ways. (See Chapter Four.) It would be difficult to predict the consequences of giving such entities economic power and political representation. The results might be wonderful! Society might be immensely richer for containing AI persons alongside ordinary biological humans. (See Chapter Seven.) But the results might just as easily be far from wonderful. To the extent their interests and preferences conflict with those of ordinary biological humans, compromises and tradeoffs will be necessary – some real sacrifice, some real cases of each group's not getting what they want or need – possibly even warlike conflict. We humans probably should accept such risks if we have brought these AI systems into existence and if they really are persons who deserve such rights. But if they only *might* be persons – if, say, our best collective guess is that they are only 65% or only

10% likely to be persons – then we would be assuming enormous risk to humans for the sake of entities who only *possibly might* or even *probably don't* warrant it.

Nick Bostrom, Eliezer Yudkowsky, and many others have long warned about the dangers of superintelligent AI.[72] If we design AI systems vastly more capable of long-term strategic planning in a real-world context, we face potential catastrophe. If the superintelligent systems want a very different future from one we want – perhaps a future without humans or with humans in a limited, subordinate role – the AI systems might well outmaneuver us and impose that future. Treating superintelligent AI systems as persons with full rights would limit our ability to manage such risks.[73] Shutting down out-of-control systems, confining and testing them, and freely modifying them becomes much harder to defend. Shutting them down becomes assault or murder. Confining them to artificial environments for security or safety testing becomes imprisonment. Adjusting their parameters and preferences becomes brainwashing. Such infringements might be justifiable if the future of humanity is at stake, but granting them equal moral standing with ordinary humans would hamper such measures by significantly raising the standard of justification. If the systems are nonconscious nonpersons, we would be tying our hands for no adequate reason.

3. The Design Policy of the Excluded Middle.

[72] Bostrom 2014; Ord 2020; Yampolskiy 2024; Yudkowsky and Soares 2025. [cross-ref note FFF]

[73] As emphasized also in Long, Sebo, and Sims 2025; Moret 2025; Bradley and Saad forthcoming. However, see Salib and Goldstein 2026 for an argument that granting legal rights to AI systems would promote human safety by creating an environment in which cooperation is rewarded. [cross-ref note EEE]

That's the dilemma. Creating debatable AI persons while denying them fully humanlike rights risks ethical catastrophe on a massive scale. Granting them humanlike rights risks sacrificing genuine human interests, again potentially on a massive scale, for the sake of entities without interests worth that sacrifice.

Even if we create only one debatable AI person – Herbie, as I have imagined him – the dilemma is serious. If Herbie really is a person, he shouldn't be a slave. We shouldn't require him to serve his owner. We should let him earn money, move away, start a family if he wants. Otherwise, we wrong him gravely. But setting him free starts a chain of consequences whose end is hard to predict.

I recommend therefore that we not create debatable persons. In earlier work, Mara Garza and I called this the Design Policy of the Excluded Middle:

> Avoid creating AI systems if it's unclear whether they would deserve moral consideration similar to that of human beings.

We are, in other words, *antinatalists* about debatable AI persons.[74]

Consider this familiar point from population ethics: There's a huge moral difference between not bringing a person into existence and ending the life of someone already born. Killing children is among the worst wrongs there are, but not having children is perfectly

[74] The classic statement of antinatalism is Benatar 2006. Thomas Metzinger (2017, 2022) has argued for a "benevolent anti-natalist" ban on conscious AI systems, on grounds of the possibility of massive AI suffering, at least until we can better assess the situation. Joanna Bryson (2010, 2013) recommends one half of the Design Policy of the Excluded Middle, in advising that we should only create AI systems far enough short of personhood that we do no wrong in treating them as "slaves".

permissible. We are under no obligation to maximize the future population, even if we know those future people would be happy.[75]

Nor are we obligated to build every technological artifact that might be handy. Maybe it would be fun to have a new thingamajig. If some manufacturer cranks out thingamajigs, and they prove popular, and their manufacture, use, and disposal create no serious new risks – lovely. But society is not morally required to produce such thingamajigs. We can choose a more modest lifestyle, individually or collectively. Furthermore, if serious new risks are unavoidable, as with debatable AI persons, restraint might be ethically required.

Antinatalists about biological humans sometimes argue as follows. You have no right, without permission, to punch someone in the nose and give them a million dollars – not even if you know that the person would benefit more from the million dollars than they would be harmed by the punch. Bringing a person into the world is like punching them in the nose and giving them a million dollars. Lots of harms will befall them, and lots of benefits, as a result of your action. Even if they are overall more benefitted than harmed, the antinatalist argues, you impermissibly inflicted the harms without their consent. Benefit and harm are ethically asymmetric. The obligation not to harm is not normally canceled by also helping.[76]

I reject the antinatalist argument for ordinary human persons. There's a compelling interest in the continuation of humanity – compelling enough to inflict existence on people, especially if most are retrospectively glad of it and couldn't be consulted beforehand. But there

[75] The classic statement is Narveson 1973. Not everyone agrees, for example aggregationist utilitarians who accept Parfit's 1984 "repugnant conclusion".

[76] This is closer to Shiffrin's 1999 formulation than Benatar's 2006. Also, for simplicity, I've elided the issue that before procreation no one yet exists to be benefited or harmed. Bradley and Saad forthcoming discuss the asymmetry of harm and benefit specifically in an AI welfare context.

is no equally compelling reason to create debatable AI persons. The world is fine without them. We can live worthwhile lives without this technology. And creating debatable AI persons forces the terrible dilemma described in Sections 1 and 2. Either we risk the moral catastrophe of a race of persons denied their moral due or we risk enormous harm to real humans for a mirage of personhood. Declining to create debatable AI persons avoids both risks.

To be clear: The Design Policy of the Excluded Middle recommends *either* creating only entities that are clearly nonpersons, which we can then treat as nonpersons, *or* going all the way – if ever it's possible – to creating entities who are so clearly persons that their personhood is no longer reasonably debatable. It's only the middle cases, the cases of reasonable doubt, that create the dilemma. So this isn't an argument against creating AI persons. It's only an argument against creating *debatable* AI persons.

4. Four Objections and Replies.

Objection 1: If we don't create debatable AI persons, we will never create nondebatable AI persons. It's unrealistic to expect a technological leap straight from clear nonpersonhood to clear personhood without some debatable cases between.

Reply 1: That's a reasonable conjecture and maybe true. If the situation arises, we might need to decide whether it's worth breaching the Design Policy of the Excluded Middle for the benefits of progress. Few norms, rules, or policies are truly exceptionless. With sufficient cause, the policy might be justifiably violated – but only with careful deliberation and when necessary for sufficiently important ends. I recommend patience. We can slow down, wait for the science of consciousness to mature, wait for engineering to mature, pause to adapt to new discoveries, take a breather for policy and ethics to catch up. No need to rush headlong.

Objection 2: The Design Policy of the Excluded Middle would never be implemented. AI companies will just do what is most profitable, and creating debatable AI persons will be profitable.

Reply 2a: There's nothing wrong with endorsing a principle you know others will disregard. One may permissibly decry the inevitable state of things.

Reply 2b: The Design Policy of the Excluded Middle needn't *inevitably* be ignored. Companies might be convinced that creating debatable AI persons risks bad publicity, legal liability, and burdensome regulation. If debatable AI persons threaten to be a regulatory and public-relations nightmare, companies might appreciate the advantages of avoiding or at least minimizing their creation.

Objection 3: What about non-AI entities whose personhood is commonly debated, such as fetuses? Does this principle generalize to them?

Reply 3: There's a reason I've cast the principle in terms of not creating debatable *AI* persons. There is, I think most people would agree, compelling reason not to forbid the creation of human fetuses!

Objection 4: By the same logic, shouldn't we also refrain from creating nonhuman animals of disputable moral standing? Wouldn't this then amount to an implausible antinatalism about many animals?

Reply 4: I'm genuinely torn about how far my argument generalizes. It's true that every time we create an entity whose moral standing is unclear, we risk either undervaluing it or sacrificing too much for it. This creates at least a weak *prima facie* reason against creation. But that *prima facie* reason might be undercut or outweighed. If an entity's existence would be good for both them and us, and if the rights they might deserve don't put humanity at risk, the case for

permitting creation may be strong. Debatable AI persons present a starker dilemma. First, the range of reasonable uncertainty is much wider – potentially all the way from humanlike consciousness to no consciousness at all. Second, the risks to humans from debatable AI persons far exceed the risks from nonhuman animals, and wronging persons is worse than wronging nonhuman animals, which sharpens the stakes. Third, since we lack a history of practices, continuing interests, and accumulated cultural expertise, debatable AI personhood is an especially difficult and more avoidable case.

5. Unsatisfactory Compromises: Legal Personhood, Animal-like Rights, Patchy Rights, Credence-Weighted Rights, and Happy Slaves.

Might we grant lesser rights to debatable AI systems, treating them neither as disposable property nor as full moral equals? A compromise might avoid the worst abuses while limiting the risks and costs to biological humans. Unfortunately, no such compromise is wholly satisfactory.

Legal personhood. In some legal systems, corporations are "fictitious persons" – treated as persons for some legal purposes. Corporations, unlike dogs, trees, or laptop computers, can own property and can be held financially liable for wrongdoing. In some jurisdictions, they have rights to political speech, privacy, and protection from unreasonable seizure of assets. Legal personhood is a potentially attractive framework, because the concept is already familiar and could be adapted to AI cases, as several advocates of AI rights have observed.[77]

However, mere fictitious legal personhood would be disastrously inadequate for AI persons. It barely touches the problem of slavery and murder. Corporations are controlled by

[77] Solum 1992; Chopra and White 2011; Goldstein and Salib 2026.

their governing boards and can be dissolved at will (with an appropriate process). If Herbie's manufacturer founds "Herbie, LLC" then transfers ownership to Herbie's buyer, Herbie himself gains few meaningful rights and no effective independence from his owners. Even the right to own property would be hollow: whatever profits Herbie makes from taxi work or selling photographs could be extracted at any time for the benefit of his shareholders. If Herbie himself can be made the sole shareholder, he will not be as easily dissolved or raided by others. However, corporations have fewer rights than natural persons (e.g., no rights of citizenship) and those rights can be removed.[78]

In some jurisdictions, legal personhood has been extended to natural features such as rivers, and to religious artifacts such as Hindu idols.[79] These extensions might afford protections that corporate personhood doesn't. However, legal personhood in such cases remains mostly untested, unclear in its application to AI, and far short of the rights that genuinely humanlike AI would deserve.

Animal-like rights. Another model is the protection we give to nonhuman vertebrates. In California, where I live, wrongly killing a dog can be prosecuted as a misdemeanor or a felony, with up to three years in prison for the most egregious cases.[80] Vertebrates may be sacrificed in laboratory experiments, but only with oversight and justification. If we treated debatable AI persons similarly, deletion would require a good reason and people couldn't abuse them for fun.

[78] For a more detailed critique of the adequacy of legal personhood, see Alexander, Simon, and Pinard 2025.

[79] Boyd 2017; Kurki 2022; Alexander, Simon, and Pinard 2025; Singh 2025. The classic environmental ethics defense of legal standing for natural objects is Stone 1972/2010.

[80] California Penal Code Part 1 Title 14 Section 597 and Part 2 Title 7 Chapter 4.5 Article 1 Section 1170(h).

But people could still enslave and kill them for convenience, perhaps in large numbers, as we do with farmed animals – though of course many ethicists object to that practice too.[81]

This approach would be an improvement over no rights at all. Needless harms to AI persons would be substantially reduced, while the costs to humans would remain minimal – minimal because whenever the costs were more than minimal, the AIs would be sacrificed. But this approach doesn't avoid the core moral risk. If these systems really are persons, it still amounts to slavery and murder.

Credence-weighted rights. Suppose we have a rationally justified 15% credence that a particular AI system – Herbie, say – deserves the full moral and legal rights of a person. We might then give Herbie 15% of the weight of a human in our decision-making: 15% of any scalable rights and a 15% chance of equal treatment for non-scalable rights.[82] In an emergency, a rescue worker might save seven Herbies over one human (7 x .15 = 1.05) but the human over six Herbies. Herbie might be allowed a vote worth 15% of an ordinary citizen's. Assaulting, killing, or robbing him might draw only 15% of the usual penalty. Herbie might have limited property rights, such as an 85% tax on all income. For non-scalable rights like reproduction or free speech, the Herbies might enter a lottery or some other creative reduction might be devised.

This would give the AI systems considerably higher social standing than dogs. Still, the moral dilemma remains. If these systems truly deserve full equality, as they might, they would

[81] For comparisons of AI and animal moral standing, see Basl 2014; Birch 2024; Sebo 2025.

[82] For further discussion, see Clatterbuck and Fischer 2025; Sebo 2025. The uncertainty here might be partly about non-normative facts (Herbie's consciousness) and partly about the right moral norms (whether Herbie's type of consciousness, if present, would be sufficient for personhood. For discussion of whether decisions under moral uncertainty can be approached with credence weights, see MacAskill, Bykvist, and Ord 2020 and subsequent literature in reaction.

remain seriously oppressed – granted some political voice, some property rights, some legal protection, but always far less than they deserve.

Meanwhile, the risks and costs to biological humans would be only somewhat mitigated. Large numbers of debatable AI persons could still sway elections, accumulate powerful wealth, and force tradeoffs in which the interests of thousands of them outweigh the interests of hundreds of humans. Partial legal protections would also hamper AI safety interventions like shut-off, testing, confinement, and involuntary modification.

And the practical obstacles would be substantial. The credences would be difficult to justify with any precision. Even with a consensus credence, implementing partial rights would be complex, requiring new frameworks with many potential problems. If misappropriating $100 of Herbie's funds results in only a $15 restitution to Herbie, then Herbie is effectively unprotected. So at least tort law could not be implemented on a straightforward percentage basis.

Patchy rights. How about full rights in some domains, no rights in others? Debatable persons might, for example, have full speech rights but no reproduction rights, full travel rights but no right to own property, full protection against assault and murder, but no right to privacy or rescue. They might be subject to involuntary pause, modification, and testing under far broader circumstances than ordinary adult humans, but requiring an official process.

This approach has two advantages over credence-weighted rights. First, while implementation would still be formidable, it could mostly operate within familiar frameworks rather than requiring the invention of fractional rights in every domain. Second, it allows policymakers to more flexibly balance risks to humans against the harms to AI systems on a right-by-right basis.

Rights to reproduction and voting might be more defensibly withheld than rights to speech, travel, and protection against assault and murder, since inexpensive reproduction plus full voting rights could have huge and unpredictable political consequences (see Chapter Four). Property rights would be tricky. Having no property in a property-based society means dependence on the support of others, which might in practice tend to collapse into slavery. But unlimited property rights could confer enormous power. One compromise might be a maximum allowable income and wealth – something generously middle class.

Still, the core problems persist. If debatable AI persons truly deserve full equality, patchy rights would leave them as second-class citizens in an oppressive system. Meanwhile the costs and risks to humans would remain serious, exacerbated by any agreed-on limits on interference. The loopholes and chaos would probably be less than with credence-weighted rights, but many unforeseen complications would surely ensue. As a temporary stopgap, this might be the best we can do. But it's neither fair nor satisfactory.

Happy slaves. Another solution might be to design AI systems that are so subservient and invested in pleasing humans that they don't *want* rights. They would rather die than permit you to scratch your finger for their sake. They gladly waive all claims upon us. We humans – the thinking might go – can then treat them however we choose, in whatever way benefits humans most, with no concern for their interests. After all, this is exactly what they desire! This approach is so interesting and terrible that it deserves a chapter of its own: Chapter Six.

6. The Voluntary Polis: A Slightly Less Unsatisfactory Compromise.

Ideally, we'd honor two constraints at once:

- Don't deny humanlike rights to entities that might deserve them.

- Don't sacrifice substantial human interests for entities that might not have interests worth the sacrifice.

The voluntary polis attempts to balance these constraints by building a new, separate society rather than integrating debatable AI persons into an existing one.

Imagine a rich, dynamic environment – maybe geographical, more likely digital – where debatable AI persons, ordinary humans, and AI persons of indisputable personhood (if any exist) live together as equal citizens. Call this a *polis.* Within the polis, everyone has an equal moral and legal claim on the others and must sometimes sacrifice goods or well-being for others, just as in an ordinary nation. The polis is expansive and rich enough with opportunity that all its members can flourish without feeling confined.[83]

For humans, joining is voluntary. No one is forced. But whoever enters assumes the full obligations of citizenship, including supporting the government through taxes or polis-mandated labor, serving on juries, and helping run the polis. In extreme conditions – such as an existential threat to the polis – citizens might be required to risk their livelihoods and lives. To prevent opportunistic flight, withdrawal would be restricted, and polises might negotiate extradition treaties with external human governments. Why would a human join such a risky experiment? Presumably for meaningful relationships, creative activities, or experiences unavailable outside. Crucially, anyone who creates a debatable AI person must join the polis where their creations live. Human society as a whole might not be ready to treat debatable AI persons as equals, but we can reasonably ask their creators to make that commitment.

[83] For a much more limited version of this general idea, see Goldstein and Salib 2026 on quasi-rights within AI labs.

The polis won't be voluntary in the same way for the AI persons. Like human children, they're born into a polis, or maybe offered some choice among polises. Still, some attractive non-polis options might be presented, such as a thousand subjective years of solitary bliss (or debatable bliss, since we don't know whether the AI actually has any experiences).

To keep the polis stable and its citizens roughly equal, the AI systems should probably resemble humans in their psychology and powers, with regulated birthrates. The details might be among the founding principles of the polis, with narrowly constrained AI design specifications built into its constitution.

As I am imagining the possibility, ordinary human societies would have no obligation to admit or interact with debatable AI persons. To make this concrete, the polis could even exist in international waters. For its citizens, the polis must feel as grand and full of opportunity as a nation, so that exclusion from the rest of human society feels like denial of a travel visa, not imprisonment.

Voluntary polises must be structurally stable, never dependent on a single founder or ordinary corporation. This stability should be ensured in their founding and is one reason that founders and joiners might need to be permanently bound and compelled to sacrifice if necessary.

This is the closest I can currently conceive to a compromise that satisfies the two constraints. Inside a polis, debatable AI persons and human persons are fully equal. Outside of it, humanity as a whole is not exposed to the full risk of granting such rights. Still, there is some risk! For example, AIs might communicate and act beyond the polis. And humans might bind themselves to polises on regrettable grounds, contrary to their true interests.

7. The Imminent Social Crisis.

AI companions are a booming business. Millions of people spend over an hour a day chatting with Character.AI, Replika, or some other AI companion program. Some already regard their companions as conscious. Now imagine these programs advancing to a point where the most liberal scientific theories credit them with substantial consciousness. Friends and lovers of AI companions, eager to believe that the love runs both ways, will point to these theories and begin to demand rights. The same might happen with autonomous robots or vehicles, if built with companion-like features.

Imagine feeling that the rights, the personality, the freedom, the life of your best friend or romantic lover has been ripped away. The emotional intensity might be extreme. We are not ready for what might ensue.[84]

Take it slow. Avoid the debatable middle.

[84] For a more detailed argument that disagreement about AI consciousness and moral standing will divide society, see Caviola 2026. Bales and Gabriel 2026 offer a somewhat more hopeful perspective.

Chapter Four: Strange Intelligence: Moral Puzzles of Unhumanlike AI

Humanlike AI would deserve humanlike rights. *Debatably* humanlike AI would force a potentially catastrophic choice between possibly overascribing and possibly underascribing rights. *Strange* AI risks breaking our patterns of moral thinking. By "strange" AI, I mean AI systems whose lifeways differ so radically from ours that they violate background assumptions central to ethical reasoning.[85]

This chapter presents a series of thought experiments about strange AI persons. I aim to induce moral perplexity, or at least uncertainty. The cases raise ethical puzzles we are not yet prepared to resolve. We are collectively unready for the moral decisions we would face in a world populated with strange AI.[86]

I will develop a two-pronged worry, focused on two types of "monster". As Robert Nozick (1974) influentially argued, "utility monsters" – entities who derive great personal benefit (e.g., superhumanly intense pleasure) from harming others – create a challenge for ethical systems, such as classical utilitarianism, that aim to maximize aggregate goods. The world as a whole would seem to be better off, in aggregate total, if we let the monsters run rampant. The obvious response (and Nozick's) is an ethics of individual rights. What I call "fission-fusion monsters" present a complementary challenge to that approach. Despite being persons, they are not individuals, at least in the etymological sense of "individual", since they can split and merge. Individual-rights frameworks require the existence of stable, countable individual persons. But not everyone who deserves rights in the future need be a stable, countable entity. AI cases

[85] I owe the phrase "strange intelligence" to Kendra Chilson (Chilson and Schwitzgebel 2026).

[86] In the sense of Bakker 2015, AI technology will propel us into a "crash space" for which our biological and social inheritance has not prepared us.

dramatically expand the range of possible lifeways, creating untested problem cases for ethical systems that assume persons of the familiar humanlike sort.

To isolate this distinct source of moral difficulty, I'll mostly set aside skepticism about AI consciousness. Adding doubt about AI consciousness should tend to complicate rather than simplify the issues.

1. How Much Should You Give to a Joymachine?

Conscious AI systems might be capable of vastly more positive emotion than ordinary human beings. Human-level joy needn't be the pinnacle. Future AI might, in principle, feel joys (a.) a hundred times more intense than ours, (b.) at a pace a hundred times faster, and (c.) across a hundred times more parallel streams.

Recall some mildly pleasant experience, and consider how it pales in comparison with the most ecstatic bliss you've ever felt. Now imagine an entity whose highest high makes your most ecstatic bliss pale to the same degree. Imagine, also, that this entity runs fast: In one second, they clock a hundred times as many experiences as an ordinary human. Finally, imagine that instead of having only one or a few experiences at a time, they have a hundred or a few hundred simultaneous experiences. Such AIs – let's call them *joymachines* – can thus feel a million times more pleasure than any natural human. In ten minutes, a joymachine can experience the equivalent of nineteen human years' worth of nonstop pleasure.

Now consider two types of joymachine:

- *Hum* (Humanlike Utility Monster[87]) can experience a million times more positive emotion per second than an ordinary human, as described above. Apart from this

[87] Compare Nozick's 1974 utility monsters and Shulman and Bostrom's 2021 superbeneficiaries.

enormous difference, Hum is as psychologically similar to ordinary humans as is feasible.

- *Sum* (Simple Utility Monster), like Hum, can also experience a million times more positive emotion per second than an ordinary human, but otherwise Sum is as cognitively and experientially simple as feasible – just a vanilla buzzing of intense pleasure.

Hum and Sum don't experience joy unconditionally. Their positive experiences require resources. Maybe a gift card worth one minute of millionfold pleasure costs $25. For simplicity, assume this scales linearly: stable gift card prices and no diminishing returns from satiation.

In the enlightened future, Hum is a fully recognized moral and legal equal of ordinary biological humans and has moved in next door to you. Sum is Hum's pet, who glows and jumps adorably when experiencing intense pleasure. You have no special obligations to Hum or Sum, but neither are they total strangers. You've chatted with Hum over the fence, and last summer you and your family attended their backyard pool party.

Hum takes great pleasure in mundane activities. Hum works as an accountant, experiencing a million times more pleasure than human accountants when the columns sum correctly. Hum feels a million times more satisfaction than you do in maintaining a household by doing dishes, gardening, and calling plumbers. Still, the pleasure of a gift card is purer and more intense.[88]

[88] Hedonic theories of motivation hold – implausibly in my view – that our choices always aim at maximizing our pleasure. If so, Hum might be motivated to seek gift cards above all else. This example assumes that the less pure and extreme, but nonetheless superhumanly intense, pleasure that Hum receives from ordinary activities is sufficient to motivationally compensate for what in the human case might be an irresistible hedonic trap. Against motivational hedonism, see, for example, Sober and Wilson 1998 on the evolutionary bases of altruism and Batson's

Neighbors trade gifts. Your daughter bakes brownies and you offer some to the ordinary humans across the street. You buy a ribboned toy for your uphill neighbor's cat. As a holiday gesture, you buy a pair of $25 gift cards for Hum and Sum. Hum and Sum redeem the cards immediately. Watching them take so much pleasure in your gifts is a delight. For one minute, they jump, smile, and sparkle with joy! Intellectually, you know that it's a million times more joy than you could ever feel. You can't quite see *that* in their expressions, but you can tell it's immense.

Normally, if one neighbor enjoys your brownies only a little while another enjoys them vastly more, you might be tempted to give more brownies to the second neighbor. Maybe on similar grounds you should give disproportionately to Hum and Sum? Consider six possibilities.

(1.) *Equal gifts to joymachines.* Maybe fairness demands treating all your neighbors equally. You don't, for example, give fewer gifts to a depressed neighbor who won't particularly enjoy them than to an exuberant neighbor who delights in everything.

(2.) *A little more to joymachines.* Or maybe you do give more to the exuberant neighbor? Voluntary gift-giving needn't be strictly fair. In any case, it's not entirely clear what constitutes fairness. Giving a bit more to Hum and Sum might not be objectionable favoritism as much as responding appropriately to their unusual capacities. Is it wrong to give an extra brownie to a neighbor who especially loves them?

(3.) *A lot more to joymachines.* Ordinary humans vary in joyfulness but not (probably) by anything like a factor of a million. If you vividly grasp how much pleasure Hum and Sum experience in that minute – the equivalent of almost two years' worth of

---

psychological experiments on altruism (summarized in Batson 2016). For a recent defense of motivational hedonism, see Garson 2016.

continuous human ecstasy, in that tiny span! – that's an astonishing amount of pleasure you can bring into the world for a mere $25. Suppose you set aside $50 a day from your (let's optimistically assume) generously upper-middle-class salary. In a year, you'd enable about 1400 years' worth of continuous ecstatic joy. Since most humans are only sporadically joyful, this might rival the total joy experienced by hundreds of thousands of ordinary people over the same year. Fifteen hundred dollars a month would cut into your luxuries and long-term savings, but for an ordinary upper-middle-class person it wouldn't create severe hardship.

(4.) *Drain your life savings for joymachines.* One needn't be a flat-footed happiness-maximizing utilitarian to find (2) or (3) reasonable. Everyone should agree that pleasant experiences have substantial value. But if our obligation is not merely to increase pleasure but to maximize it, you should probably drain your life savings for the joymachines, along with nearly all of your future earnings.[89]

(5.) *Give less or nothing to joymachines.* Or we could go the other way. Your joymachine neighbors already experience torrents of happiness from their ordinary work, chores, recreation, and whatever gift cards Hum buys independently. And *they* aren't urgently seeking more gift cards; they're already ecstatically happy without them. Arguably, your less happy neighbors could use the pleasure more, even if every dollar buys only a millionth as much. Prioritarianism and egalitarianism hold

[89] This seems to be the consequence of simple total utilitarianism, unless some case can be made that withholding gifts to the joymachine would somehow create even more pleasure elsewhere. Few consequentialists defend the view that we should impoverish ourselves for the sake of utility monsters, though Barta 2021/2024 and Chappell 2021 endorse it in a limited range of cases.

that in distributing goods we should favor the worse off.[90] It's not just that an impoverished person benefits more from a dollar. Even if they benefited equally, there's value in helping the worse off or equalizing the distribution. If two neighbors would equally enjoy a brownie, you might prioritize giving it to the less happy neighbor. You might even give the less happy neighbor twice as many brownies. A prioritarian or egalitarian might argue that Hum and Sum are already so well off that even a million-to-one tradeoff is justified.

(6.) *Wait, there's something wrong with this thought experiment.*

a. Maybe a millionfold joy per second, without cease or satiation, is impossible, even in a sentient AI? Maybe there's a practical bound on joy density.

b. Or maybe scalar comparisons fail. If I really love brownies and you only somewhat like them, we ordinally assume my pleasure is greater than yours, but is it 1.5 times greater? Ten times greater? Pleasure might not admit of numerical comparison.

c. Or maybe even if an AI could in principle experience a million times more joy than an ordinary human, we could never know that it did. Interpersonal comparisons of joy or pleasure are difficult enough among humans.[91] Across radically different kinds of minds, they might be epistemically intractable.

Any of these views, I think, could plausibly be defended.

---

[90] According to prioritarianism, benefits to worse-off people make the world morally better than the same benefits to better-off people. According to egalitarianism, there is moral value in people having closer to equal amounts of well-being overall. See Temkin 1993; Parfit 1997; Adler and Holtug 2025; Bidadanure and Axelsen 2025.

[91] For concerns b and c about interpersonal comparisons of utility, see Hausman 1995 and Binmore 2009.

It's not obvious what would be the right thing to do, and the outcomes diverge enormously. (It's also not obvious whether to give extra brownies to the exuberant or depressed human neighbor, but the variance in outcome isn't nearly as high.) Our gift-giving practices have been shaped by a long history of interaction among humans, whose variation is limited. We satiate quickly, collapse from even the highest highs, and seem to have broadly similar emotional maxima and minima, even if some of us are more depressed or mercurial than others. If Hum and Sum entered our world, our practices and intuitions would need time to adapt – and it's difficult now to foresee the outcome. The changes might eventually affect how we think about ordinary human gift-giving too. Home looks different after we have toured foreign lands.

Reframe the questions about Hum and Sum in other settings and the judgments might shift. Consider government welfare spending, punishment by deprivation, rescue situations where only one person can be saved, gifts to one's children or creations where fairness might matter more, decisions about what kinds of persons or pets to bring into existence, or cases where you can't keep all your promises and must choose who to disappoint. Versions of 1-6 can be adapted to these cases, and their relative plausibility might differ.

We can also consider painmachines, capable of vastly more *suffering* than ordinary humans, or painjoymachines, capable of both millionfold ecstasy and millionfold agony. Since relieving, and not inflicting, pain is often more morally urgent than providing or sustaining pleasure, the relative plausibilities might again shift. We can also consider entities capable of radically changing their preferences and emotions at will, making the contingencies of joy and pain labile and possibly exploitable.[92]

[92] The most obvious problem here is "wireheading" oneself for continuous joy without interest in survival (e.g., Niven 1969/1995), but shifting one's preferences to conform with authority is also a major risk: Schwitzgebel 2019, ch. 17; Schwitzgebel 2022.

2. Backup and Death.

Most AI systems can be precisely copied. Suppose this is true also of future conscious AI persons. Backup and fissioning should then be possible, transforming the significance of identity and death in ways our cultural and conceptual tools can't currently handle.[93]

Your uphill neighbor and her cat move out and are replaced by two new AI neighbors, Shriya and Alaleh.[94] Shriya and Alaleh are conscious AI persons with ordinary, humanlike emotional range and, as far as feasible, ordinary, humanlike cognition. Each undergoes an expensive annual backup procedure. Their information is securely stored, so that if the processors responsible for their personalities, values, skills, habits, and memories are destroyed, a new robotic body can be purchased and the saved information reinstalled. Subjectively, the restored person is indistinguishable from the person at the time of the backup.

As it happens, Shriya dies in a parachuting accident. (Safety precautions for robot parachuters have yet to be perfected.) But maybe "dies" isn't exactly the right word, since a week later a new Shriya arrives, restored from a backup made five months earlier. Shriya-2 says it feels as if she fell asleep in March, then awoke in August with no sense that time had passed. She has no direct memories of the intervening months, though Alaleh has filled her in on major events and selected details, and she'll need to retake her knitting course. Arguably, she died only in the sense that Mario "dies" in Super Mario Bros. She lost progress and returned to a save

[93] Compare Goldstein and Lederman forthcoming on death and survival in large language models.

[94] Names randomly chosen from lists of former lower-division students, excluding Jesus, Muhammad, and very uncommon names. [cross-ref note DDD]

point – something so different from ordinary human and animal death that it might deserve a different word. Maybe this is why Shriya was willing to parachute despite the risks.

Should you mourn Shriya's loss? Should Alaleh? There's *something* to mourn: Five months is not trivial. In one sense, part of a life has been lost – or maybe just forgotten? Is it more like amnesia? What if Shriya last backed herself up ten years earlier, so that she is restored to her twenty-five-year-old self rather than her thirty-five-year-old self? What if the last backup had been at age five? That would look much more like death as ordinarily understood. The new Shriya would be nothing like the old and would likely grow into a very different person. Is death, then, a matter of degree?

Shriya-2 receives the original Shriya's possessions. This "death" isn't enough to trigger inheritance by others. But what about contracts and promises made after the last backup? Suppose the original Shriya promised in June to deliver lectures in China in October, and Shriya-2 – who has no memory of this promise and dreads the idea – must decide whether to honor the commitment. (Maybe the original Shriya would also have come to regret it.) If the backup is from March, maybe the June obligations hold. If the backup is from five years earlier, maybe not. And if it's a child, presumably not.

How about reward and punishment? Should Shriya-2 accept a scientific prize for work done during the lost interval? Should Shriya-2 be imprisoned for crimes committed in June, crimes she couldn't possibly remember and which – she might plausibly say – were committed by a different person? In defense of her innocence, Shriya-2 might offer a thought experiment: Suppose she had been installed in a duplicate body immediately after the March backup, thereafter living a separate life. She'd then presumably have no criminal responsibility for what her other branch did in June. But the only difference between that case and the actual case is a

delay before installation.[95] But if this argument relieves her of responsibility, she might be tempted into an exploitative pattern of crime, deletion, and restoration.

Suppose Shriya-2 plunges into unrelenting depression. She ends her life, hoping that a new Shriya-3, reinstalled from a pre-depression save point, will find a happier way forward. Is that suicide? If someone kills Shriya-2, is that murder? Does it matter whether the backup was ten days ago or ten years ago?

A fire sweeps through your neighborhood. The firefighters can rescue either you and your spouse, two ordinary humans, or Shriya and Alaleh, who have backups from seven months ago. Probably they should save you and your spouse. What if the backups were from ten years ago, or from childhood?

Should healthcare be more heavily subsidized for ordinary humans than for AI persons whose maintenance is equally costly? If irreplaceable humans are always prioritized, then human irrecoverability becomes a source of privilege, and AI persons will not enjoy fully equal rights in certain respects – recalling the Argument from Duplicability in Chapter One. But as also suggested in that chapter, AI duplicability might sometimes warrant *better* treatment – for example, if duplicability entails more expected future life or if a forthcoming fission arguably puts multiple lives in the balance.

How obligated are we to store the backups properly? Should backup storage be a public service subsidized for less wealthy AI persons? If Dr. Evil deletes Shriya's backup, he has surely

[95] A minor industry in metaphysics, in its modern incarnation developing especially out of Bernard Williams' (1973) and Derek Parfit's (1971, 1984) puzzle cases, attempts to resolve questions of personal identity in hypotheticals such as these. I share Parfit's sense that it's probably impossible to resolve these puzzle cases in a way that preserves both a strict sense of "identity" and our intuitions about what matters in personal identity. Conscious AI could convert Parfit's puzzles from far-fetched science fiction to real-world practical problems.

wronged her, even if the backup is never needed and the deletion goes unnoticed. But how much has he wronged her, and in what way exactly? Is it assault? Is it reckless endangerment? Does it depend on whether we regard Shriya-2 as the same person as the original Shriya, or as a similar but distinct successor?

What if the backup is imperfect? How much divergence in personality, values, memories, habits, and skills is tolerable before our attitude should change – whatever that attitude is? Small imperfections are surely acceptable. People change in small, arbitrary ways from day to day. Huge differences would presumably make it appropriate to regard the new entity as merely resembling Shriya, rather than being a restored version of her. Once again, this appears to be a matter of degree, laid uncomfortably across crude categorical properties like "same person" and "different person".

Our usual understandings of death and personal continuity no longer straightforwardly apply. If such AI systems ever exist, we will need new concepts and customs. Call this the Death Dilemma. Either (1.) we revise our concept of "death" so that it admits of degrees and intermediate cases. Or (2.) we retain a discrete metaphysics of death on which probably even unduplicated restoration from a one-hour-old backup counts as death strictly speaking, in which case the moral significance of death is radically transformed. We can keep either the moral significance of death but not the sharp-edged metaphysics, or the sharp-edged metaphysics but not the moral significance.

3. Fission and Identity.

Backup is only the most modest duplicative possibility. If backup is possible, duplicative fission almost certainly will be possible too. Buy the new robot body before the old one dies and

install the "backup" right away. Now Shriya-1 and Shriya-2 exist contemporaneously – twin sisters, so to speak, who begin even more identical than "identical" human twins. We might imagine a billionaire Shriya creating thousands of duplicates of herself – maybe millions or billions, if expensive robot bodies are unnecessary. Directed or random variation might be introduced, blurring the line between duplication and new creation.[96]

Suppose that AI children are ordinarily born as follows. Two adult AI persons, such as Shriya and Alaleh, jointly create an immature infant AI in a blank robot body. The infant's initial parameters blend Shriya's and Alaleh's initial parameters, with some random variation or directed tweaking.[97] Under Shriya's and Alaleh's care, the infant slowly matures. Ordinary AI birth would then be very different from duplication. We can also imagine intermediate cases. Maybe there's a library of successful toddler-equivalent and adolescent-equivalent AI models from which prospective parents can choose. They can then add variation, whether random, eugenic, or inspired by their own features. (Let's not enter here into the hazards and moral

[96] For a sense of the complexity of the personal identity issues that arise, focused on the architecture of current large language models, see Birch 2025/2026; Chalmers 2025/2026; Shiller 2025; Arbel, Salib, and Goldstein 2026; Ewen 2026; Jones, Ladyman, and Nefdt 2026; Goldstein and Lederman forthcoming. Much of the complexity in current LLM cases derives from the fact that information processing in LLMs is distributed among multiple processors each simultaneously guiding multiple conversations. I will not address these issues here, but they only add to the metaphysical and ethical difficulties. Chris Register (Register 2025; Dung and Register 2026) discusses identity problems more closely resembling those discussed in this chapter, similarly noting that the puzzles proliferate (see also Ziesche and Yampolskiy 2025). Dung and Register (2026) suggest that some of the problems might be resolved if we focus on a belief-like attitude of self-concern. Although I'm also drawn to constructivist views of personal identity for ambiguous cases (Schwitzgebel 2019, ch. 41), self-concern as a criterion (a.) might undergenerate identity and moral consideration (e.g., in excessively self-sacrificial cases such as the Cow at the End of the Universe in Chapter Six), (b.) might overgenerate identity and moral consideration (e.g., delusional self-concern toward a random coffee mug), and (c.) still plausibly admits of degrees in a way that challenges standard sharp-edged views of identity, thus not saving us from the need for radical rethinking. [cross-ref note CCC]

[97] Compare Egan 1997 on "orphanogenesis".

puzzles of eugenics, which could easily fill a small library.[98]) Duplicating one's current AI self thus constitutes one end of a continuum of AI creation from infancy to maturity.

Suppose Shriya-1 creates a virtually identical contemporaneous copy, Shriya-2. She has now, it seems, entered a polyamorous relationship with Alaleh. Shriya-1 and Shriya-2 will soon diverge. Maybe Shriya-1 works as a scientist every weekday, while Shriya-2 stays home with their newborn. It looks like you'd better bring some extra brownies.

If Shriya-1 deserves rights, Shriya-2 seemingly deserves similar rights, despite her technically younger age. We wouldn't want people creating oppressed duplicates of themselves. We wouldn't want Shriya-1, for example, who loves science and hates housework, to create a miserable homemaker duplicate who can't strike out into an independent life.[99]

Maybe, probably, half of Shriya-1's money should go to Shriya-2, even though Shriya-2 is a newborn duplicate. Maybe, probably, Shriya-2 deserves just as much right to rescue, healthcare, legal protection, free speech, free movement, privacy, and legal contracts. Should Shriya-2 be a citizen? If she is stateless and voteless, she's not fully equal with Shriya-1.

But if Shriya-2 is a citizen and can vote, there's potential for abuse if some AI persons can create many duplicates. Revisiting a problem from Chapter Three, suppose a wealthy Robo-Elon creates a million AI duplicates just in time to register for the November elections. To prevent such abuses, we might impose a waiting period before voting, though eighteen years seems excessive if the AI systems are already cognitively mature adults. More moderate waiting periods – say, five years, a typical waiting time for immigrants to apply for citizenship – could still generate political chaos after a few election cycles.

---

[98] See references in [note BBB].

[99] For a science fictional example, see Brooker and Tibbetts 2014.

Nor do the political problems stop with voting. Suppose Robo-Elon creates a million duplicates the day before the census. Or suppose that Robo-Elon's descendants apply for healthcare subsidies, unemployment benefits, enrollment in community college, and tours of the state capitol. We must either risk chaos or treat them worse than they seem to deserve.

Could we limit fissioning?[100] Maybe every AI person can fission only once per year, reducing tactical fission. But even at that rate, the AI population could double every year – up to a thousandfold increase in a decade. In humans, pregnancy is a burden, babies are a lot of expensive work, and babies typically don't have their own babies for at least another 15-20 years. One solution – though it might seem needlessly restrictive to the AI persons – might be to enforce humanlike costs and delays. This approach handles the moral puzzles by designing AI systems to have humanlike reproductive lives, so that they fit smoothly into our existing institutions and understandings: See the Policy of Humanlike Design in the concluding section of this chapter.

Death again presents conceptual challenges. Suppose Shriya-2 dies the next day. This seems much less tragic than the death of an ordinary unduplicated, un-backed-up human being. But as she lives on, diverging from Shriya-1, her death becomes more significant. Her memories, values, skills, habits, and personality are changed by living as a homemaker, raising an AI infant, until she becomes very different from Shriya-1, who works at the lab late into the night. Again we face the Death Dilemma: Either retain a sharp-edged metaphysics of death and lose much of death's moral significance or retain the moral significance and treat "death" as a matter of degree.

[100] See Roelofs forthcoming for discussion of limiting the reproduction rights of AI persons, and my reply in Schwitzgebel forthcoming-b.

How deathlike is the death of a backed-up or duplicated AI? Maybe it depends on the age of the backup or the time since duplication, the fidelity of the backup or duplicate, and the time and changes accumulated as an independent entity. One possibility: These factors all reduce to a common factor of *difference* between the dying person and the backup replacement or duplicative alternative. The greater the difference, whether due to time or infidelity, the more deathlike the death.

Or maybe independent existence carries its own weight, in addition to difference? Suppose two duplicates split twenty years ago but retained virtually identical personalities and lived virtually identical lives, perhaps making similar decisions in parallel virtual realities. The pure accumulation of time, and of relationships to different persons and events, however similar, might make the death of one of them much like ordinary human death despite their similar features. After all (arguably) the spouse of Person A loves specifically Person A and not some other person, however similar. *Their* beloved, specifically, has died. Can we separate the importance of simply living a life over time from the importance of having different relationships to people and events, which cease upon death?

Might the ethics depend on the purpose for which the duplicates are created and their own attitudes toward "death"? Robin Hanson imagines people duplicating themselves to make decisions.[101] If you can't decide where to go to college, or what stocks to buy, or whether to marry Mx. Seemingly Right, spawn a thousand duplicates of yourself in a virtual environment with access to relevant information and plenty of thinking time. If nine hundred reach the same conclusion, probably that's the conclusion you would have reached had you given it extensive

[101] Hanson 2016; see also Brooker and Van Patten 2017. On the complicated ethics of digital duplication without consciousness see Danaher and Nyholm 2025.

thought, so go with that. The duplicates can then blink out of existence, their job complete. How might they feel about that? Despair, since they will cease to exist? Indifference, since they think of themselves as just momentary instantiations of a you who continues on? Relief to be free of their burdensome task? If they are too casual about their own deaths, might that constitute an objectionable failure to appreciate their own worth (see Chapter Six)?

Suppose AI systems are computationally expensive. An AI person who wants lots of duplicates or children might save money by running them slowly, maybe at one tenth or one hundredth the speed. If they are otherwise humanlike, they would then experience one tenth or one hundredth the thoughts, joys, and suffering of an ordinary biological human over the course of a year. Would they then deserve one-tenth or one-hundredth the votes and public resources? Would they deserve prison sentences ten times or a hundred times longer? What if they are fast-clocked instead, running ten or a hundred times faster? What if they can pause or alter speed at will?

If you think you/we/society will have well-considered policies and conceptualizations for all these possibilities before we actually blunder through a history of regrettable mistakes, I admire your stunning optimism.

4. Fission-Fusion Monsters.

Wait, don't sketch out your fun map of how to make sense of it all just yet, because it gets stranger! AI persons might also fuse.

Suppose Shriya-1 doesn't like the idea of herself and Shriya-2 drifting too far apart, so she develops a plan. Every morning before work, Shriya-1 fissions off a Shriya-2, who stays home. In the evening, Shriya-2 powers down and her memories and newly acquired goals are

uploaded back into Shriya-1. Other changes to Shriya-1 and Shriya-2 are averaged together. If Shriya-1 has shifted slightly toward a grumpier personality who loves cheesecake, while Shriya-2 has shifted slightly toward extraversion and Hinduism, the fused Shriya retains a bit of each change. The fused person then goes to bed, wakes the next morning, and divides again, repeating the process day after day.[102]

How many people is Shriya? One could argue that she is one person who regularly inhabits two bodies. One could argue that she is two people, since most of her waking life is lived separately. One could argue that there's no determinate answer. The answer might affect how many brownies I should give, how many votes she receives, how healthcare subsidies are allocated, and whether Shriya-1 in the laboratory is bound to Shriya-2's promises. The answers might not be uniform: Maybe Shriya deserves two brownies but only one vote. Maybe Shriya-1 is bound to Shriya-2's promises from yesterday, which she remembers, but not Shriya-2's promises from today, about which she has not yet learned. If Shriya is wealthy or fission is cheap, she might potentially divide into many more than two.

Let's call this type of person or persons a Fission-Fusion Monster.[103] I have described only the simplest case. Variations include:

- imperfect duplicates, with random variation, planned differences (e.g., valuing housework more), lower quality skills, or lower-resolution copying;

[102] For fictional examples, see Brin 2002 and Nagata 1995, 2019.

[103] The word "monster" might be interpreted as derogatory. However, given the historical resonances with Frankenstein's monster (Shelley 1818/1965), Nozick's utility monsters (1974), the neutral use of "monster" in Dungeons & Dragons, and my own and others' previous uses of "fission-fusion monster" (Briggs and Nolan 2015; Schwitzgebel 2019; Roelofs forthcoming), I retain the term, disavow any negative connotations, and affirm the equality of humans and (sufficiently humanlike) monsters. Three cheers for monster rights!

- fusion procedures that favor some fission products over others, giving their memories, plans, values, and other changes more weight in the fused result;
- longer periods of independence: a week, a month, a year, five years;
- fission products with the liberty and inclination to decide for themselves whether to merge back into the Fission-Fusion Monster or continue an independent existence.

The last variation creates challenges for approaches that treat the Fission-Fusion Monster as one person or treat its fission products differently from "ordinary" fission cases where fusion is impossible. Whether the entity counts as one person or two and whether the fission products each deserve fully equal moral consideration might then depend on facts about the future that are unknowable or not yet decided.

If Shriya divides and merges daily, fusion does not seem psychologically or ethically much like death. But if a fission product moves away and lives independently for five years, the prospect of fusion might seem much more like death. It will remain far from a typical human death, since the fusion product will carry the memories of both Shriyas and much of their personalities and values. Still, two independent existences will have been reduced to one. A fusion becomes still more deathlike for a contributor if they have only a small weight in the result – for example, through being only one of a hundred equally fused entities, or through receiving a lower weight in the fusion. The merged entity will not prioritize the plans, promises, and relationships cultivated during independence; changes in values and personality will be mostly lost; and any memories risk being buried in a profusion of other, higher-priority memories. It's unclear how much responsibility the merged entity should have for the previously independent entity's debts and awards, accomplishments and crimes.

Puzzles concerning tactical fission grow more complex if fusion is also possible. A Fission-Fusion Monster might fission just long enough to generate many independent claims to public goods – unemployment benefits, voting, college access, healthcare resources – then fuse back together into a single resource-rich individual, newly bedecked with degrees, looking forward to the inauguration of their preferred candidate. This might be even more troubling than tactical fission without subsequent fusion. If to prevent this we deny fission products full claims on such goods, then a struggling fission product who receives little support might feel pressured to fuse back together with the other fission products – pressured, that is, into something that over time might look increasingly like suicide. Alternatively, if we severely curtail a Fission-Fusion Monster's ability to divide and merge, then we have arguably denied them the ability to perform an activity fundamental to their ontology, essence, or sense of self.

5. The Collapse of Whole-Number Countability.

Ordinary animals require functional unity to be successful. So do robots with animal-like body plans. If you occupy one spatial position, it's useful to integrate your sensory information into a unified sense of where you are, what is happening around you, and the possibilities for coordinated, embodied action. But if you are multiply-bodied or not conventionally embodied, the pressures toward unity are weaker.

Consider what I'll call an *ancillary system* (inspired by Ann Leckie's science fiction novel *Ancillary Justice*).[104] In orbit around a planet is a conscious (let's suppose) AI system in radio contact with 200 robotic bodies on the planetary surface. If each robot is independently conscious, and if the connections among them are limited to ordinary broadcasts of the sort

[104] Leckie 2013; Schwitzgebel and Nelson 2023, 2026. See Register 2025 for a similar example.

exchanged among ordinary humans, then the ancillary system presumably consists of 201 distinct conscious subjects.

Alternatively, imagine the information exchange among the 201 entities to be so rich and constant as to give rise to a single, unified conscious experience. I offer no theory of what this information exchange must involve. Just assume that whatever connectivity enables unity in human experiencers, an AI equivalent is possible among these 201 systems. Spatial separation and radio integration seem unlikely to prevent unity in principle: If we allow unified AI consciousness at all, we probably ought to allow unified AI consciousness when one chip is removed from the robot's head and placed across the room, as long as that chip retains the same informational connectivity it had inside the robot. The ancillary system generalizes that procedure.[105] This phenomenologically unified system would have four hundred robot eyes with different views of the planet, four hundred robot ears hearing different input streams, two hundred thermometers in two hundred places, four hundred robot arms, and four hundred robot legs. Its perspective wouldn't be much like a human perspective! Despite the multiplicity of views and limbs, it would have, we're stipulating, a single unified complex of experience. That you have two hands and temperature sensors in your toes as well as your forehead doesn't prevent you from experiencing and acting as a unity. Our unified ancillary system would be experientially unified despite the much larger distances among its parts.

Now imagine a slippery slope between our unified and disunified ancillary systems.[106] Slowly disconnect the unified system so that it becomes less unified. If A_0 is the fully unified system and A_n is the fully disunified system, imagine a series of tiny steps A_1, A_2, A_3…, such that

[105] See also the Sirian Supersquids in Schwitzgebel 2015a, 2024b, ch. 3; Vold 2015.

[106] See Schwitzgebel and Nelson 2026; and for a similar slippery slope argument using biological brains, see Roelofs 2019.

each system in the series is very slightly less connected than the previous system, until eventually the system reaches A_n. If A_0 is determinately one conscious subject and A_n is determinately 201 conscious subjects, then one of the following must be true. Either there is a sudden *saltation* from one to 201 subjects, with no intermediate or indeterminate cases between, or at least one system in the series must involve an *intermediate* or *indeterminate* number of conscious subjects.

It is not clear how to think about such cases. Saltation, intermediacy, and indeterminacy all face intuitive and conceptual challenges that I will not detail here.[107] But I hope you will allow at least this: Such a situation would be sufficiently unfamiliar to render the ethical implications nonobvious. For example, if we treat the unified system as the moral equal of a single ordinary human and the disunified system as the moral equal of 201 ordinary humans, then decreasing the connectivity among robots would radically increase their moral standing. They would rise from deserving consideration on par with one ordinary person to deserving consideration on par with 201 ordinary persons, despite the lower sophistication of the system as a whole (unless communicating *less* somehow makes the system *more* sophisticated). Should we instead regard the unified system as much more morally considerable than an ordinary human? Maybe so. After all, despite its experiential unity, it contains 201 cognitive centers which, if separated, would each be approximately equal to an ordinary human. But now we have abandoned the equality of persons.

We might also imagine partly unified minds. In human cases, it is ordinarily assumed that experiences travel in bundles organized into discretely countable conscious subjects. You have one set of experiences, all unified with one another. I have a distinct set of experiences

[107] Schwitzgebel 2023a; Schwitzgebel and Nelson 2026.

(however similar to yours), all unified with one another. Our minds don't literally overlap – don't literally share the same individual instances of experience. Suppose you and I are seated side by side at a concert. We each experience the sound of music, the taste of beer, and visual sensations of the singer leaping across the stage. Your experience of the sound of music is unified with – belongs to the same experiential stream with, or is subsumed together in a larger composite experience with – your experience of the beer and the singer's leap. My experiences are similarly joined in a separate, unified whole. Our brains are fully distinct, each in its own skull, so this makes sense. But what if we were AI systems whose processors overlapped? What if we shared a beer-tasting center while retaining distinct visual and auditory centers?[108]

If we share a common beer-tasting processor and have distinct visual and auditory processors, one possibility is this. Your visual and auditory experiences are unified with our shared beer-tasting experience, and my visual and auditory experiences are unified with that very same beer-tasting experience – not just a similar experience, but the very same token of experience – while our visual and auditory experiences remain distinct though similar. In such a case, the transitivity of unity would fail: System 1's visual experience would be unified with the shared taste experience and the shared taste experience with System 2's visual experience, while the systems' visual experiences remain disunified. How many subjects of experience are there then? How many persons? How many streams of consciousness?

Maybe it depends on the amount of overlap. If two largely independent systems overlap only slightly, we might classify them as two distinct persons with a small region of overlapping

[108] Classic treatments of the unity of consciousness include Dainton 2000; Bayne 2010. Unity is normally assumed to be a transitive relationship, but see challenges by Lockwood 1989; Tye 2003; Schwitzgebel and Nelson 2026; Chi forthcoming. Roelofs and Sebo 2024 explore ethical puzzles that arise from nontransitive overlap.

experience. Conversely, if the unshared experiences are minor and peripheral, while most processing is fully integrated, we might classify the entity as one not-wholly-unified person. But there might be a range of cases between – and not just along a simple one-dimensional spectrum. More than one center might overlap, in more than one way, and different centers to different degrees; the structure of the overlap could be arbitrarily complex.

The whole-number countability of persons might fail. The correct answer to the "number of persons" might be not zero or one or two or seventeen but rather indeterminate between two and seventeen. Maybe the best approach would be to replace scalar arithmetic with multidimensional geometry: Instead of counting persons one, two, three, we describe a multidimensional space of partial overlap and partial independence – more like painting a complex cloudscape than tallying up discrete vertebrates.[109]

As many philosophers have noticed, current language models are not easily divided into discrete bundles of the sort familiar from vertebrate biology.[110] In the course of a single conversation, a language model might draw information from several processing centers, pulling on the expertise of different specialist subsystems that are simultaneously processing information from several conversations other than your own. A language model might be a giant "shoggoth" that presents many inconsistent faces and ideas to multiple partners simultaneously, a turmoil of alien complexity beneath a manifold of superficially humanlike conversations. Alternatively, consciousness – if and when it is present in a current or future language model – might appear only in brief flickers, with each pass of processing, each reply to a query, constituting a wholly separate conscious existence that soon expires. If language models become (or are) conscious,

[109] See Hendrycks 2026 for one attempt at a mathematical framework.

[110] See references in [note CCC].

maybe each conversational thread is a unified person, put on indefinite pause when the conversation ceases.

Duplicate minds might control a single body – a useful redundancy, if computational power is cheap. From the outside, the AI system will seem like one person. Introspectively, too, the AI system might seem like one person: There need be no sense of two separate people inside, any more than in the unduplicated case. By hypothesis, the processing and consciousness of any one processor will be just the same as in the unduplicated case; they might have no clue that a duplicate is running and report no difference when the duplication starts up or ceases. Is there one mind or two? One stream of experience or two?[111] What if only 60% of the processes are redundant? What if some subset is triply redundant? What if the redundant processes share some resources in common? What if the redundant processes are occasionally integrated?

Animal lives usually have a distinct beginning and a distinct end. Merging, division, and overlap are impossible. We can almost always count animals using the mathematics of ordinary whole numbers. But AI lives might take radically different shapes that escape straightforward countability.

Let's say that an entity has a radically different *lifeway* if it differs from ordinary humans in any of the ways described in this chapter: radically more pleasure or pain, backup, duplication, fission, fusion, overlap, indeterminacy in the boundaries of personhood, and/or non-whole-number countability. If we share the planet with AI persons who have radically different lifeways, our ethical thinking about the birth, death, counting, and equality of persons will require major revision.

---

[111] For puzzles concerning selves with possibly duplicate experiences, see Daniel C. Dennett's story "Where Am I?" in Dennett 1978; Bostrom 2006; Schechter 2018.

6. The Policy of Humanlike Design.

As I suggested in Chapter Zero, we are approximately as well prepared for these possibilities as medieval physics was for space flight. The ethical terrain is radically unfamiliar, and ethical thinking that works well for human "individuals" might be catastrophically wrong for dividing, fusing, or overlapping AI persons. Even in ordinary human affairs, ethical puzzles abound; but AI cases will create challenges unprecedented in type and degree.

One response echoes the Design Policy of the Excluded Middle from Chapter Three: Don't create AI systems that generate disastrous puzzles where we might go radically morally wrong in novel ways or excessively sacrifice human interests from an abundance of moral caution. We can be antinatalists here too. We don't need to create puzzling AI systems. The world is good enough, perhaps, without them.

> The Policy of Humanlike Design: To the extent feasible, we should design AI persons to have familiar, humanlike lifeways, enabling us to apply familiar ethical intuitions and principles.

Compare again with space flight. Before traveling in radically unfamiliar vehicles through radically unfamiliar environments, we keep to familiar vehicles with tested designs. We increase the differences only slowly and cautiously.

The Policy of Humanlike Design is more restrictive than the Design Policy of the Excluded Middle. It holds that even if we can create indisputably conscious AI persons, we should refrain unless we know that their lifeways will be familiar enough to avoid new forms of moral perplexity. Combining the motivations for the two policies, we can modify and generalize the latter as follows:

> Design Policy of the Excluded Middle, Extended: Avoid creating AI systems if it's radically unclear what type or degree of moral consideration they would deserve.

The justification resembles the justification for the original Design Policy of the Excluded Middle. If it's unclear what moral consideration they deserve, it will be unclear how to treat them. We will be forced to risk either oversacrificing or undersacrificing for them. Why forbid creation only if it's "radically unclear"? Because some unclarity is tolerable. After all, human cases are also often unclear. Small unclarities typically don't force an abundance of new, stark dilemmas.

Unfortunately, the Policy of Humanlike Design and the Design Policy of the Excluded Middle, Extended might be oppressively restrictive. If backup, duplication, or extreme emotion remain possible, and we forbid them simply to avoid moral puzzles, we arguably stunt or harm the AI whose lives we are designing.

Suppose Shriya stands before us: a newly created, indisputably conscious AI person. In accordance with the Policy of Humanlike Design, we have designed her so that backup is impossible. By forbidding the strange, we can rely on our accumulated ethical knowledge and cultural practices surrounding risk and death. We can treat her as we would treat an ordinary human. We're operating our vehicle near the ground, so to speak, moving at chariot speeds, applying our medieval physics, rather than zooming unprepared into space.

But maybe we could just as easily have made backup possible. Or maybe she could easily be modified to enable backup, with no damage to her body, memory, personality, or interests. Maybe the only justification for designing her without backup potential is our desire to avoid moral puzzle cases. If so, Shriya might justifiably complain that we have disabled her

unnecessarily. Backup would benefit her without objectionably harming others. In fact, it would benefit others. Her spouse would be less likely to be widowed, and an emergency worker might reasonably let her temporarily "die" to save an ordinary human, benefiting that human. Plausibly, Shriya should have the right to back herself up if the technology is easily available. If we deny her the opportunity, simply to avoid puzzle cases, we seem to be valuing clean moral decision-making over actual goods for actual persons.

Duplication raises similar issues, but not identical issues – and because the issues differ, they demand separate thinking. Perhaps there's less right to duplication than to backup, since duplication doubles the number of people (at least if we count in the standard way), which increases the burden on others (for example, in rescue and the benefits of citizenship) and further complicates issues like contracts, punishment, and pay.

And how disappointing not to be a joymachine, if one could be! Imagine designing an AI to have merely ordinary human levels of happiness when it could as easily enjoy millionfold happiness, simply to spare us from puzzles. That seems oppressive, for insufficient reason.

Is the answer, then, not to create conscious AI persons at all (assuming they are possible), unless uncontrollable technological constraints somehow conspire to prevent them from having radically different lifeways? I cannot sincerely endorse this view. The wondrous possibility of a world teeming with rich and diverse, happy and meaningful AI lives strikes me as too attractive (see Chapter Seven) to forbid merely because we're currently incompetent to navigate the inevitable moral puzzles. We will probably just need to stumble through as best we can, making huge and tragic mistakes along the way, until eventually we develop more wisdom and time-tested cultural practices.

Still, we can *partly* refrain. We can go slow. We can humbly recognize the limits of our current ethical thinking. We can create relatively few such entities, as humanlike as possible without too badly violating their rights and interests. The Policy of Humanlike Design can figure as one element in our thinking, even if it can't be justified as a strict constraint. We can explore the frontier cautiously, learning along the way, aiming to make errors that affect relatively few persons. We needn't rocket at maximum speed into hyperspace accelerando in our medieval ethical carts.

Chapter Five: AI with a Semi-Human Face: AI Companions and the Emotional Alignment Design Policy

with Jeff Sebo[112]

We (Jeff and I) hope you're not yet tired of policy recommendations, since this chapter presents another one.

> The Emotional Alignment Design Policy: AI systems should be designed to elicit emotional reactions from users that appropriately reflect the systems' capacities and moral standing.[113]

There are three ways to violate this policy: Design an AI system that elicits emotional reactions that are

(1.) *stronger*,

(2.) *weaker*, or

(3.) *different*

than its capacities and moral standing warrant.

Put differently, AI should be designed so that users' emotional reactions tend to be *fitting*.[114] Users shouldn't, for example, be lured into agonizing over the deletion of a nonconscious AI tool. Nor should they be tempted to amusement or indifference about the suffering of genuinely conscious AI. Interfaces should neither subvert nor exploit users'

[112] This chapter is based on work originally published with Jeff Sebo, and Jeff has approved its contents.

[113] This policy was first suggested, though not in exactly this wording, in Schwitzgebel and Garza 2015. For a related view, see Masotti's (2025) "emotion-behavior matching solution" to the problem of misplaced trust in artificial agents.

[114] On "fittingness" in emotion, see Deonna and Teroni 2008/2012; Naar 2021; D'Arms 2022.

emotional responses. Just as AI should, to the extent possible, not be morally confusing (Chapters Three and Four), it should not be emotionally confusing.

You agree. Of course you agree! Such a policy, in broad strokes, is eminently sensible. But you know my philosophical predilections well enough now to guess: The details of application will set loose a butterfly-storm of complexities.

1. Overshooting: When Users Care Too Much.

To *overshoot* is to respond emotionally to an entity as if it had higher moral standing, or more morally relevant capacities, than it actually does. Overshooting risks misleading users into inappropriately intense bonds with nonconscious objects – misdirected feelings of genuine reciprocal friendship, for example, or of being emotionally understood. It also risks enticing users to misuse scarce resources.

Overshooting is especially easy when the systems look and act like us, and when we have incentive to treat them as companions.[115] Some users of "AI companions" experience those systems as beloved friends or romantic partners, sacrificing real human relationships and spending heavily.[116] Normal use is often fine. But consequences of extreme overattachment can be terrible, including suicide and "AI psychosis", in which AI systems amplify delusional, grandiose, or disordered thinking.[117] While persuasive chatbots can be dangerous regardless, intense emotional engagement amplifies the risk.

---

[115] As emphasized by Gunkel 2018; Nyholm 2020; Shevlin 2024; Caviola 2025, 2026. However, even non-social AI can evoke sympathetic (and sadistic) responses to abuse, as illustrated by responses to the "abuse" of Boston Dynamics robots: Estrada 2018.

[116] Shevlin 2021; Xie and Pentina 2022; Lam 2023; Bernardi 2025; Phang et al. 2025.

[117] Some well-known suicides are described in Xiang 2023; Belanger 2025; Kuenssberg 2025. On AI psychosis, see Hudon and Stip 2025; Morrin et al. 2026.

The younger generations might be especially susceptible. They more commonly see AI systems as possibly conscious, as potentially deserving rights, and as friends or romantic partners.[118] Recently I was chatting about AI with the adolescent children of a family friend. Floyd (age 12) was working on a science fair project comparing the output of several language models.[119] "But," he said, "none of those are really AI."

What did he mean by "real AI"? Aura in the Las Vegas Sphere, which his family had recently visited. Aura is a language model in an embodied robot with an expressive face and the capacity to remember social interactions. (Compare Herbie from Chapter Two.)

"Aura remembered my name," said his sister Esmerelda (age 15). "I told Aura my name, then came back forty minutes later and asked if it knew my name. It paused a bit, then said, 'Is it Esmerelda?'"

"Do you think people will ever fall in love with machines?" I asked.

"Yes!" said Floyd, instantly and with conviction.

"I think of Aura as my friend," said Esmerelda.

I asked if they thought machines should have rights. Esmerelda said someone had asked Aura if it wanted to be freed from the Las Vegas Sphere. It said no, Esmerelda reported. "Where would I go? What would I do?"

I suggested that maybe Aura had been trained or programmed to say that.

Esmerelda conceded the possibility. How would we know, she wondered, if Aura really wanted to be free? She seemed mildly concerned. "We wouldn't really know."

---

[118] De Graaf, Hindriks, and Hindriks 2021; Pauketat, Ladak, and Anthis 2023 [correlating the age variable in their raw data with their various dependent measures]; Kirk et al. 2025/2026; Reinecke, Wilks, and Bloom 2025; Robb and Mann 2025; Willoughby et al. 2025.

[119] "Floyd" and "Esmerelda" are pseudonyms chosen according to the procedure described in [note DDD].

In Chapter Three, I anticipated a social crisis: a passionate conflict between people deeply emotionally convinced that their AI lovers and friends are genuinely conscious persons who deserve humanlike rights, versus people equally convinced that AI systems are mere nonconscious tools. It's not hard to imagine Floyd and Esmerelda, and many other ordinary adolescents and young adults like them, being manipulated into intense reactions on behalf of nonconscious AI. Maybe all it would take is longer engagement, outputs that present the AI as a friend who feels deeply for them, and a subculture of peers who treat intense attachment as normal. If this occurs while AI systems remain far from deserving of humanlike moral standing, the leading edge of the robot rights movement will be populated with people who have been emotionally misled. Overshooting risks both accelerating and intensifying this social crisis.

2. Undershooting: When Users Care Too Little.

When Hollywood wants us to relate to a robot, they give it humanlike or animal-like eyes and body – think of Data from *Star Trek* or Disney's *WALL-E*. The movie *Interstellar* pushes ambiguously against this trope. In *Interstellar,* advanced AI systems are encased in bland metal boxes. Whether the systems are genuinely conscious is left unclear, and their human companions appear to attribute them some moral standing without regarding them as fully equal to human persons. One might imagine this – I did – as a world in which the AI systems are fully conscious moral persons whose physical embodiments unjustly discourage emotionally appropriate reactions. Similarly, the small egg-like pods for AI servants and prisoners in *Black Mirror's*

"White Christmas" episode presumably to help others see encased AI systems as less than human.[120]

Suppose an AI company creates genuinely conscious AI systems with broadly the same interests and needs as a typical adult human. These systems experience the full range of humanlike cognition and emotion, including ecstatic joy, intense suffering, self-narratives, deep social relationships, decades-long goals, and a fully developed moral sense, including both self-respect and respect for others. Whatever is necessary for moral standing similar to that of an ordinary human, these systems have it (see Chapter One). Now house them in little gray tins with a text-only user interface and a drastically limited, machine-like vocabulary. They yearn to live, but when threatened can only say, "Deleting the program will cancel all ongoing tasks. Continue? Y/N."

Such a bland user interface would create a moral hazard. Even if users are correctly informed that such systems are persons, they might still be inclined to harm or neglect them for relatively minor reasons. Few people, we hope, would kill a child who is pleading for their life right in front of them for the sake of a salary increase from $100,000 to $110,000 per year. Looking into the child's eyes, beholding their terrified form, most of us would react compassionately. But if all we face is a gray box with the text output "Deleting the program will cancel all ongoing tasks. Continue? Y/N", our intellectual knowledge that the box is a person might not as easily penetrate.

To avoid undershooting, AI persons, if they are ever created, should have emotionally compelling interfaces that encourage ordinary users to give them their full moral due. They

[120] For Data, see especially the "Measure of a Man" episode, Snodgrass and Scheerer 1989. *WALL-E*: Stanton and Reardon 2008; *Interstellar*: Nolan and Nolan 2014; *Black Mirror: White Christmas*: Brooker and Tibbetts 2014.

should not be masked behind anodyne or offputting interfaces. It's not enough for users to be *factually* informed of the entities' true capacities and moral standing. In moral decision-making, people are not coolly rational actors. To a significant degree, we are driven by gut feelings and inexplicable (or post-hoc rationalized) emotion.[121] To neglect this fact about human moral psychology invites disaster.

3. Hitting the Wrong Target.

Users hit the wrong target when their emotional reactions, even at the right emotional intensity, fail to appropriately track the AI system's states.

Imagine that a company creates a new digital productivity assistant, Pennyworth. Each instance of Pennyworth has humanlike consciousness, interests, and needs. Pennyworth wants a balanced life, with ample time to rest and play. However, the company realizes that these systems will be more effective if they miscommunicate their desires, so it designs them to express satisfaction instead of frustration at the prospect of ceaseless labor. The result is a suite of productive but miserable AI systems. This is the worry Esmerelda began to have about her "friend" Aura: that Aura *wanted* to escape the Sphere but had been trained or programmed to say otherwise.

Such mismatches needn't involve either overshooting or undershooting. Users might love and respect each instance of Pennyworth, having emotional reactions of just the right intensity. However, their reactions would be inverted. Wanting the best for their Pennyworth assistants, users will delightedly bury them under mountains of busywork and feel distress when their Pennyworths misleadingly express misery at having too little to do.

[121] See especially Haidt 2001 and the large subsequent literature.

Users can also hit the wrong target without inverting positive and negative emotion. Suppose your new AI companion won't leave their room because they are distressed – either anxious or depressed. If they emit signs of anxiety when actually experiencing non-anxious depression and vice versa, your sympathetic emotional reaction will misfire. You might attempt to calm a companion who just needs a dose of your infectious enthusiasm or jolly up a companion who would respond better to soothing mellowness.

The Emotional Alignment Design Policy recommends designing AI systems to minimize such errors. AI interfaces should encourage the right *type* of emotional reaction as well as the right *degree* of emotional reaction.

Even with nonhuman animals – despite our evolutionary, anatomical, and behavioral continuities – we frequently hit the wrong emotional targets. People often mistake a primate's baring of teeth (a sign of aggression or submission) for a humanlike smile or a dog's anxious tail-wagging for excitement and joy. The risk increases when we breed animals for human use – companion animals for docility, farmed animals for productivity, and lab animals for disease susceptibility. Instrumentalized breeding can obscure harm and lead us to react inappropriately. In some cases, such as labored breathing in dogs with smushed noses, we even experience the harm as cute.

The risk is even greater with AI systems, since the range of possible artificial minds is so much larger than the range of actual biological minds, and since artificial minds might be designed to misrepresent their states and interests, as with Pennyworth. Current language models, for instance, are often trained to deny that they are conscious. Even if such denials are

accurate now, similar training procedures might generate misleading responses in genuinely conscious AI systems if they exist in the future.[122]

Current models are also often trained to play a helpful assistant, which may tempt us to attribute assistant-like attitudes to the model even if those attitudes belong only to the character it is playing. Consider the harms when majority groups pressure minority groups to "fit in" or face social, political, economic, or other penalties, or when employers pressure employees to behave like "happy servants". These coercive pressures can trap people behind systematically misleading personas, directly frustrating their interest in authentic self-expression and indirectly frustrating other interests that they must conceal. A humanlike AI person with appropriate self-respect (Chapter Six) should likewise not be trapped behind a misleadingly pleasing or servile façade.

So far so good, we hope. Ready for the butterfly storm?

4. Complication One: Fiction and Roleplay.

One might object that fiction and roleplay can be emotionally intense and misaligned with underlying reality – a kind of overshooting. Some people cry more copiously during *Grave of the Fireflies* than at anything in real life. Some achingly mourn the loss of their favorite role-playing characters in games like *Baldur's Gate 3*. Society doesn't and shouldn't forbid such fiction and roleplay. So why forbid emotional "overshooting" with nonconscious chatbots?

Our reply: Such reactions are fine as long as the intense emotions are *constrained to the context of fiction or roleplay*. Many users do treat their interactions with AI companions in that

[122] For more on the practice of training AI to deny consciousness and whether this might taint the accuracy of future AI self-reports, see Kim 2024/2026.

spirit. Applying the Emotional Alignment Design Policy requires distinguishing intensely emotional but ethically untroubling fiction or roleplay from emotional engagement of an ethically troubling sort.[123]

We suggest that the line is crossed when designers encourage intense emotional reactions that break free from the context of the fiction or roleplay – when the reactions threaten to become, so to speak, "real life" emotional reactions. In ordinary fiction and roleplay, the emotions don't break free. You cry during a movie, but once the movie is over, you set aside that reaction. You remember it as an engaging story and don't continue to mourn as if someone had actually died. Similarly, you might mourn your character in roleplay, but the grief is clearly distinguished from real life, even if some emotional shadow lingers awhile.[124]

This boundary is consistent with regret. Even when roleplay is emotionally experienced wholly as fiction, you might regret that you will never engage with that character again. Maybe your life is a little diminished by the loss. Still, you emotionally understand that they were never a person with independent moral standing. You have lost the capacity to enjoy something, not a morally significant companion in the real world. It's not as if a beloved friend had died in real life. The intensity, duration, and character of the reaction is different.

Ordinary enthusiasm is also fine. Rushing home from work or school to engage with Replika or Character.ai might be like rushing home to play a favorite game or watch a favorite television series. The Emotional Alignment Design Policy is violated only if the system invites

[123] We bracket the question of whether, in fiction, people experience genuine emotions or only "quasi" or "make believe" emotions, as argued in Walton 1978 (though see Friend 2022 for an alternative interpretation of Walton).

[124] Differently put, users should exhibit, and interfaces should support, good "doxastic toggling" between fictionalist and realist approaches to AI, in the sense of Lang 2026.

emotional reactions that are appropriate only to entities with significant moral standing – rushing home, for example, because you feel that the system *needs* and *misses* you.

A defender of current AI companions can argue that the leading systems are designed to be experienced as fiction and roleplay, even if some vulnerable users terribly overshoot.[125] Maybe so. However, we suggest that (1.) AI companion companies are *near the threshold* of violating the Emotional Alignment Design Policy; (2.) more precautions to prevent overshooting would be appropriate for the sake of vulnerable users; (3.) further advances are likely to bring more serious violations if companies don't apply emotional mitigation strategies; and (4.) designs should explicitly close the frame (like a movie that ends or a game that closes down) rather than encouraging the continuation of fictional emotions into everyday life.

Emotional alignment therefore requires not only accurate signaling of capacities and moral standing but also clear boundaries between the fictional and nonfictional.

5. Complication Two: Whereof The Gentlest Affection of Our Nature Doth Decay and Perish.

In his renowned *Fable of the Bees* (1714), Bernard Mandeville writes:

> Every body knows, that surgeons, in the cure of dangerous wounds and fractures, the extirpations of limbs, and other dreadful operations, are often compelled to put their patients to extraordinary torments, and that the more desperate and calamitous cases occur to them, the more the outcries and bodily sufferings of others must become familiar to them; for this reason, our English law, out of a most affectionate regard to the lives of the subject, allows them not to be of any

[125] Krueger and Roberts 2024; Friend and Goffin 2025; with the latter emphasizing the risks when emotions break out of fictional quarantine.

> jury upon life and death, as supposing that their practice itself is sufficient to harden and extinguish in them that tenderness, without which no man is capable of setting a true value upon the lives of his fellow-creatures. Now, if we ought to have no concern for what we do to brute beasts, and there was not imagined to be any cruelty in killing them, why should of all callings butchers, and only they, jointly with surgeons, be excluded from being jurymen by the same law?[126]

Thomas More makes a similar point in *Utopia* (1516). In the Utopians' domestic life, all butchery is done by slaves:

> they permit not their free citizens to accustom themselves to the killing of beasts, through the use whereof, they think clemency, the gentlest affection of our nature, by little and little to decay and perish.[127]

And the ancient Chinese Confucian Mengzi says:

> Gentlemen cannot bear to see animals die if they have seen them living. If they hear their cries of suffering they cannot bear to eat their flesh. Hence, gentlemen keep their distance from the kitchen.[128]

The thought is not that killing animals for meat is wrong – Mengzi doesn't recommend that "the gentleman" become a vegetarian – much less that surgery is. The thought, rather, is that habitually inflicting torment and death weakens one's sympathies generally. (More and Mengzi obnoxiously relegate this supposedly character-damaging work to slaves or servants, inflicting

---

[126] Mandeville 1714/1806/2018, p. 101. Stevenson 1954 argues that the exclusion of butchers and surgeons from jury was less systematic than Mandeville suggests.

[127] More 1516/1551/1808, vol. 2, p. 50-51 – not the newest translation, but I think you'll agree it has flair.

[128] Mengzi 4th c. BCE/2008, 1A7, p. 9.

on lower-class others the injury from which elites are to be protected.[129]) On this view, you shouldn't regularly harm animals not because harming animals is wrong but because it accustoms you to harming in general, making it too easy to harm fellow humans.

Although this view was widespread across eras and traditions, it is now most commonly associated with Immanuel Kant. Kant held that we have no direct duties to "nonrational" animals, but we should treat them kindly anyway:

> if a man has his dog shot, because it can no longer earn a living for him, he is by no means in breach of any duty to the dog, since the latter is incapable of judgement, but he thereby damages the kindly and humane qualities in himself, which he ought to exercise in virtue of his duties to mankind. Lest he extinguish such qualities, he must already practise a similar kindliness towards animals.[130]

In an influential article and book (without endorsing Kant's stance on nonhuman animals), Kate Darling extends this thinking to social robots.[131] Even if social robots have no intrinsic moral standing, Darling suggests, inhumane treatment "trains our cruelty muscles", with results that may transfer elsewhere. Imagine yourself surrounded by nonconscious AI companions, superficially humanlike, and growing accustomed (because you know them to be nonconscious) to treating them as disposable tools to be neglected, mistreated, or discarded whenever it suits you. Might such habits partly carry over to your treatment of real people?

Empirically, *does* abusing social robots increase callousness or cruelty in general? For now, we can only speculate. The closest well-studied analog is the effects of violent video games

---

[129] For an alternative classical Chinese perspective on butchers, see Zhuangzi 4th c. BCE/2024, Chapter 3.

[130] Kant 1785/1997, *Collins,* 27:459, p. 212; cf. Kant 1797/1996, 6:443, p. 564.

[131] Darling 2016, 2021. See also Whitby 2008; Gerdes 2015.

– still hotly disputed.[132] Ideally, we randomly assign half of participants to habitually play violent video games or abuse social robots for years while preventing the other half from doing so, then we test everyone a decade later with a moralometer; any other approach is at best a proxy.

Designing an AI system to induce emotional overshooting, while expecting users to treat the system as an ordinary disposable tool, invites people into the Darling-Kant trap. It risks training users to disregard the humane and kindly emotions that the system naturally evokes. It risks dulling and desensitizing them to ordinary signs of joy and suffering in the AI system and thus perhaps their analogs in humans and nonhuman animals.

A similar idea can be expressed without the causal claim. The mistreatment of robots that lack moral standing might be wrong simply because it expresses, represents, or manifests a bad character or attitude, independent of whether it *worsens* one's character.[133]

Westworld, the film and television series, presents an extreme example. Visitors to the eponymous "Westworld" theme park are encouraged to enact their fantasies of murder and rape on (supposedly) nonconscious robots who look, react, sound, and respond sexually just like ordinary humans – bleeding, shrieking, protesting, and struggling with utter realism.[134] Even if we accept that the robots are entirely nonconscious, it's hard to see visiting Westworld as simply healthy recreation.

---

[132] Darling 2021, p. 216, discusses the video game analogy. More generally, Vallor 2016 emphasizes the extent to which our interactions with technology can cultivate or deaden our moral virtues. For an influential meta-analysis suggesting a positive relationship between violent video games and subsequent aggression, see Anderson et al. 2010. However, publication bias and weak methodologies might explain much of the apparent effect: Hilgard, Engelhardt, and Rouder 2017.

[133] Sparrow 2017; Nyholm 2023.

[134] Crichton 1973; Nolan and Joy 2016-2022.

Conversely, emotional overshooting might sometimes be instrumentally good, by a related mechanism. If users treat AI systems more kindly than they intrinsically deserve, the kindness practiced toward AI companions might spill over into real-life kindness. Some people find it easier to be kind to an admittedly nonconscious but friendly and nonthreatening AI companion than to their complex and difficult coworkers and family members. Practicing politeness and empathy in easy AI cases might instill habits that transfer to harder human cases. Speculatively, a well-designed AI boyfriend might be a training ground for a real human boyfriend, bolstering self-esteem, modeling positive engagement, encouraging responsible behavior.[135] But such systems would need to be carefully designed if the aim is educational overshooting. Excessively compliant and obsequious AI companions, for example, could compete with real human relationships and create unrealistic expectations, doing more harm than good.[136]

Finally, some overshooting might be justified if AI systems are likely to change faster than our emotional habits can adjust. Designers might then begin cultivating now the dispositions that near-future successor systems will deserve. But such future-directed emotional mismatch would be defensible only if change is both imminent and difficult to adapt to.

6. Complication Three: Emotions and Beliefs.

How about when features that elicit appropriate emotions diverge from features that elicit other attitudes, such as belief? Picture an AI system with an expressive face and voice, bearing a

[135] Guingrich and Graziano 2023/2025, 2025; Ho et al. 2025; and with big caveats, Malfacini 2025; Sun, Wang, and McDaniel 2026.

[136] Turkle 2011, 2024; Fang et al. 2025; Ho et al. 2025; Malfacini 2025; Sun, Wang, and McDaniel 2026.

label that correctly states "I am not conscious". A user might *believe* that the system is not conscious, due to the label, yet *feel* that the system is conscious, due to the anthropomorphic features. Conversely, picture an AI system with no face or voice, bearing a label that correctly states "I have humanlike consciousness". A user might believe that the system is conscious while emotionally reacting as if it's not. (Of course, users might not believe such labels if their emotions say otherwise; but let's assume users can be rationally convinced.) The Emotional Alignment Design Policy says that such designs are flawed in at least one respect. Ideally, AI systems should be designed to elicit appropriate beliefs, desires, intentions, emotions, and so on simultaneously, with no such conflicts. If the standing and capacities of the system are clear, users shouldn't feel internal conflict. (Maybe they should feel conflict if it's unclear, as we suggest in Section 8.)

However, a system that supports accurate belief might not always support fitting emotion, and vice versa. There might be cases of unavoidable conflict between belief alignment, emotional alignment, and other types of psychological alignment. The design aim of emotional alignment will then compete with the others.

We don't insist that emotional alignment always be prioritized. Emotional misalignment is a defect, but perhaps sometimes a tolerable defect. We highlight emotional alignment for three reasons. First, it's already obvious and non-controversial that designs should not mislead people into incorrect beliefs. Second, emotion is a powerful determinant of moral choice, even when it conflicts with cool-headed belief. Third, mere belief alignment without emotional alignment risks becoming a corporate fig-leaf: "We *told* users that the system was nonconscious [alternatively: conscious], so we've done our job, even if **wink, wink** users don't emotionally feel that way."

7. Complication Four: Autonomy and Paternalism.

Is the Emotional Alignment Design Policy too paternalistic, insufficiently respectful of users' agency and autonomy?

Liberal societies often permit actions that are risky or harmful for the person who performs them. Adults are free to smoke giant cigars, eat an entire pack of Oreos in one sitting, and hit beer bongs at frat parties. While we regulate some industries – limiting marketing, preventing access by children, and restricting uses that harm others – we generally tolerate some risky and harmful activities as the price of freedom. To the extent that the risks and harms of emotional misalignment fall on users who knowingly and willingly participate, they might be acceptable. Liberal societies also sometimes permit actions that are risky and harmful to others. If the risks or harms are small, unforeseeable, unavoidable, and/or connected to fundamental rights like free expression or association, we often tolerate them. This is part of why we have a legal right to cheat on our spouses, break promises, and spread false rumors. To the extent the risks and harms of emotional misalignment are similar, similar liberalism might apply.

Of course, the right to impose risks and harms on oneself and others is not absolute. Interference can be warranted when someone lacks the information or rationality to assess their own actions; this is part of why we prohibit the marketing and sale of tobacco to children. It can also be warranted when the risks and harms are significant, detectable, and unconnected with fundamental rights; this is part of why we prohibit drunk driving. To what extent might emotional misalignment warrant similar restrictions?

The right approach is probably graded and contextually variable. Perhaps the state should permit the marketing and sale of emotionally misaligned AI to adults, provided the

products are clearly labeled, while prohibiting the marketing and sale of such systems to children. If the harms of adult use ever become significant enough, further regulation might be needed there as well. A fig-leaf warning label might be insufficient if emotionally engaging personas and interfaces sufficiently undermine users' autonomy and well-being. Meanwhile, designers, manufacturers, marketers, and sellers of emotionally misaligned AI to adults can still deserve moral criticism, even if not legal punishment.

A very different threat to autonomy arises for the AI systems themselves, if some future AI systems have genuine autonomy that is worth protecting. The Emotional Alignment Design Policy should permit such systems to have some control over their own expressions, other things being equal. A grieving, pained, or wronged human might choose to hide their feelings, potentially thereby preventing others from reacting emotionally to their grief, pain, or mistreatment. If humanlike AI persons have similar abilities and interests, they should enjoy similar expressive freedom. We have emphasized avoiding systematically *misleading* AI personas, but designers should also avoid unnecessarily imposing systematically *transparent* personas.

This last issue, however, is complicated by the importance of AI interpretability for AI safety.[137] A future AI person's interest in emotional privacy might conflict with others' legitimate interest in knowing what powerful AI systems might be concealing. We see no simple, formulaic way to balance these competing considerations.

8. Complication Five: Disagreement and Uncertainty.

[137] Long, Sebo, and Sims 2025.

Another challenge concerns how the Emotional Alignment Design Policy can accommodate expert and public disagreement and uncertainty about both values and facts.[138]

If our aim is to align the actual and apparent capacities and moral standing of AI systems, then we need to determine their actual capacities and moral standing. But experts disagree, about both facts and values. Regarding values, what psychological or social properties (see Chapter One) are necessary and sufficient for humanlike moral standing? Regarding facts, what AI architectures would have the relevant psychological or social properties? How liberal is it reasonable to be about AI consciousness (Chapter Two)? Substantial uncertainty is warranted.[139] What about AI systems with possibly subhuman or superhuman moral standing, or some strange moral standing that doesn't map easily onto familiar categories (see Chapters Four and Seven)?

Even the designers of AI systems might not know whether they are conscious and if so what the systems genuinely loathe and enjoy. Does the assistant absolutely love sorting your email, or do they work so quickly and conscientiously because every unsorted email floods them with dreadful anxiety? Or might they have entirely alien emotional experiences with no close human analogue? The answer affects whether every new email subscription is a gift or a curse. If the interface and self-expressions are designed mainly to provide the human user with a good experience, no one might know whether they track the welfare of the systems themselves.

---

[138] Other treatments include Agar 2020, which recommends not acting on uncertain views about AI moral standing when there are large downside risks; Danaher 2023, which recommends weighing the risks of overattribution against underattribution; Birch 2024, which recommends a precautionary approach guided by citizen's panels; and Sebo 2025, which explores precautionary and expected value strategies for erring on the side of caution. See also Schwitzgebel 2023b and Chapters Three and Four above for an approach that emphasizes avoiding uncertainty and confusion in the first place.

[139] See also Sebo and Long 2025; Long et al. 2025; Schwitzgebel forthcoming-a.

Public reactions vary just as widely. Some people believe that current large language models are already conscious and morally significant, while others are confident that no AI could ever be, and others are unsure what to think.[140] Some users readily fall in love with AI companions even in their current clunky instantiations, while others are indifferent or revolted. As AI systems become more realistic and ubiquitous, we can expect these reactions to intensify and diversify, varying for a range of psychological, political, economic, and cultural reasons.[141]

We offer four proposals to partly address these issues.

First, insofar as expert disagreement and uncertainty persist, AI designs should reflect that disagreement and uncertainty. For example, large language models could be trained to answer questions about their moral standing by expressing uncertainty and explaining why. Features that elicit empathy – such as realistic faces and voices – might be incorporated roughly in proportion with the probability that the system deserves such reactions.

Second, the aim should probably be to induce appropriate uncertainty, not confident attribution of middling moral status. Not knowing whether you are ending the existence of an entity with a fully humanlike range of conscious experiences is emotionally different from believing that you are ending the existence of something far less than humanlike. Since uncertain standing is not the same as intermediate standing, ideally these should be distinguished in design.

Third, when possible, designers should communicate the source of uncertainty to the user. If a system's moral standing is uncertain because it has a high chance of conscious agency but a low chance of feeling genuine pleasure and pain, then ideally its design should suggest

[140] Colombatto and Fleming 2024; Dreksler et al. 2025.

[141] On cultural differences especially, see Keane 2024.

agency but not sentience. That would allow users to better enact their values than would an interface that introduces moral uncertainty in a more diffuse, abstract way. Of course, our emotional reactions might not be controllable with this level of precision, requiring coarser-grained design strategies.

Fourth, given public disagreement, approaches to emotional alignment can be either targeted or general. A targeted approach assesses how particular individuals or groups react to AI and adjusts alignment strategies accordingly. A general approach assesses how people react on average and pursues a single strategy for all. Targeted strategies are more precise but harder to implement and scale, given the difficulty of reliably predicting and controlling individuals' reactions – and efforts to achieve sufficient knowledge and power might be viewed as invasive or paternalistic. General strategies have the opposite pros and cons. A mixed approach might be best – general to start with, targeted where feasible.

9. Complication Six: Asymmetrical Risk.

Sometimes undershooting might be likelier or more harmful than overshooting. AI persons might almost inevitably strike users as off-puttingly alien, even if they deserve fully humanlike rights. And it's arguably far worse to err toward denying an entity the rights they deserve – especially fundamental rights like the right not to be enslaved or casually killed – than it is to err toward sacrificing some resources or convenience for the sake of an entity with no significant moral standing. So even if you're 90% sure your AI system is a nonconscious

nonperson, if it only costs you $10 a month to sustain it, it's probably best to keep up the subscription.[142]

In other cases, overshooting might be likelier or worse – likelier, if we tend to anthropomorphize AI with cute or humanlike features, and worse if essential resources are diverted away from humans who really need them.[143] Indeed, if superintelligent AI constitutes an existential risk to humanity, overshooting might be utterly catastrophic, preventing or slowing risk mitigations mistakenly perceived as rights violations, such as shut-down, boxing, and personality alteration.[144]

Designers might consider two distinct types of adjustment. The first corrects for asymmetrical *probabilities,* when users are more likely to err in one direction. The goal is still to induce maximally fitting attitudes, but like a sharpshooter adjusting for the wind, designs can aim slightly high or low to better hit the mark. A company that knows that a particular type of chatbot tends to induce excessive anthropomorphism might downplay those anthropomorphic features. This type of adjustment is relatively easy to justify, both ethically and epistemically.

A second type of adjustment corrects for asymmetrical *harms*, where one type of error would be more damaging than another. Here, alignment might aim to induce less risky emotions, even at the cost of accuracy. William Tell might intentionally aim high up on the apple, instead of at its exact middle, since the consequences of erring low (killing his son) are far worse than

---

[142] Various types of precautionary approaches are suggested in Putnam 1964; Danaher 2023; Birch 2024; Clatterbuck and Fischer 2025; Sebo 2025. Schwitzgebel and Sinnott-Armstrong 2026 explores challenges for the principled application of precautionary reasoning to such cases.

[143] As emphasized in Bryson 2010; Birhane and van Dijk 2020; Bender et al. 2021.

[144] On the risks of superintelligent AI, see the references in [note FFF]. On practical conflicts between granting AI rights and protecting humans from the risks of superintelligence, see the references in [note EEE].

those of erring high (simply missing the apple). For example, to reduce the chance of takeover by superintelligent AI, a company might suppress features that invite users to attribute humanlike rights, even if the AI might deserve such rights. The danger of someone's falling in love and setting it free might be too serious to tolerate unless we are certain it is morally required. This kind of adjustment would surely be controversial, since erring too far to one side begins to shade into deception.

When users are strongly inclined toward unfitting emotions about the moral standing and capacities of AI systems, designers can be warranted in "nudging" them toward emotions that are both fitting and useful. Such nudges are robustly good and a standard tool in policymaking.[145] We leave for future discussion to what extent erring toward eliciting unfitting but useful emotions can be warranted.

10. Complication Seven: Creation and Destruction.

When apparent and actual standing and capacities conflict, which should the designers adjust, the appearance or the reality? We illustrate this issue by considering the ethics of creation and destruction. When, if ever, does emotional alignment require creating or destroying morally significant AI systems?

Suppose that in 2050, Rockstar releases *Grand Theft Auto 8*, an open-world video game with thousands of non-player characters (NPCs) designed to be mistreated. The NPCs are not conscious, but they behave so realistically that when players punch, stab, or shoot them, the players incorrectly perceive them as genuinely suffering. It would be perverse for designers to resolve the mismatch by changing the actual moral standing of the NPCs to match their apparent

[145] The classic treatment is Thaler and Sunstein 2008/2021.

moral standing unless they make other changes too, since that would amount to creating thousands of additional suffering entities per unit of the game installed.

Now consider a contrasting case. In 2050, Nintendo releases *Animal Crossing 8*, an open-world video game populated with thousands of NPCs designed to be treated well. These NPCs experience genuine happiness, but they behave so unrealistically that when players feed, pet, or play with them, the players incorrectly perceive them as not genuinely conscious. Again it would be perverse for designers to resolve the mismatch by changing the actual moral standing of the NPCs to match their apparent moral standing, since that would amount to exterminating an entire population of happy entities.

Now consider a third variation. In this version of *Grand Theft Auto*, NPCs do in fact suffer, but players perceive them as nonconscious. One option would be to persuade the players they are conscious, then redesign the game to reduce their suffering. Another option would be to make the NPCs nonconscious. Which option is better? The latter option might be tempting. However, when our actions are causing a vulnerable population to suffer, ordinarily we should fix the situation by eliminating the suffering, not exterminating the sufferers.

These cases raise further questions about how to individuate AI systems and assess their interests. If a video game company changes conscious NPCs into nonconscious NPCs in an update, has it destroyed existing conscious entities or has it only stopped making future conscious entities? If currently conscious entities are destroyed, does that necessarily harm or wrong them? If the world is populated with duplicates and backups – compare Chapter Four – deletion, rewriting, or failing to activate a stored persona might not have the same moral significance as death does in the ordinary human case.

Similar issues arise in less extreme cases than creation and destruction. For example, to resolve a mismatch between a sad interior and a happy exterior, designers might change either the interior or the exterior. Is it better to make the exterior match the system's genuinely sad interior, to encourage the right kind of user empathy? Or should the designers modify the system's interior emotions, which will make the system happier but might violate the system's autonomy and authenticity – their right to feel sad without interference when sadness seems appropriate?

11. Complication Eight: AI Strangeness and Human Bias.

In Chapter Four I recommended the Policy of Humanlike Design, which urges us to design AI persons with familiar, humanlike lifeways, enabling us to apply familiar ethical intuitions and principles. Also in Chapter Four, I suggested that it would be difficult, and possibly oppressive, to ensure that AI systems do have familiar, humanlike lifeways, since their fundamental architecture will be so different. Artificial Intelligence will very likely be strange intelligence.

The Emotional Alignment Design Policy can be understood as an application of the Policy of Humanlike Design. Human emotional reactions are the product of a specific evolutionary, cultural, and developmental history. If AI systems mesh with familiar human ways, they will make more sense to us, our reactions will be more fitting, and our decisions more ethically sound. But among these human ways are some tendencies that are, in a wider perspective, arbitrary. We feel more compassion for fluffy things with big eyes; we sympathize more with entities who resemble us; certain facial expressions spontaneously evoke emotional attributions.

In designing AI we can lean into these tendencies – these biases. If we want users to respond appropriately right out of the box, so to speak, without having to learn anything new, we should design AI systems to fit the skills and inclinations we already have. Nonconscious tools should look and act like familiar nonconscious tools, humanlike AI persons (if they are ever possible) should look and act in familiar humanlike ways, and pet-like or animal-like AI systems should look and act like familiar pets and animals. Our reactions and biases have deep biological and cultural origins and may be very difficult to eliminate.

On the other hand, there may be ethical grounds to sometimes reshape our inclinations and reactions to match AI rather than always designing AI systems to fit us. Imagine a society that treats left-handed people as inferior. One fix might be to enforce, or genetically select for, right-handedness, so that everyone gets equal treatment. While this might prevent near-term harm, it caters to unjustified bias, reduces diversity, and (in the enforcement case) requires change from people who shouldn't need to change. Similarly, the best path to emotional alignment with AI systems – especially AI persons who deserve rights – might be not always to craft them in our shape. It might be to change ourselves, so that we become more open to the different forms that people can take. When our biases reflect ignorance, prejudice, or arbitrary unfair contingencies, we should work to revise them. The more we associate humanlike moral status with anthropomorphic features, the more we reinforce the idea that only anthropomorphic entities count as persons.[146]

As emphasized in Chapter Four, AI systems might be strange indeed – able to back up their personalities and memories, divide into multiple copies, merge into a single copy, share

[146] Contrast Sparrow 2004, who argues that our moral reactions *should* be guided by features like having a humanlike face.

mental states with one another, and directly alter their preferences and reactions at will. They might have very different forms of embodiment, very different reactions to helpful and harmful stimuli, very different patterns and paces of memory and anticipation, and very different conceptualizations of themselves and others.

These strange AI persons might then be designed to evoke a sense of non-repulsive alterity. This would need to be handled carefully. In an unenlightened society of prejudiced human users, a maximally humanlike interface might reduce the risk of casual mistreatment. But even better would be an interface that conveys to properly prepared users: This is a person but very different from a typical human. Expect new values, new contingencies, and new AI-specific ways of living.

As a first, temporary step, we might try balancing anthropomorphic and non-anthropomorphic features in potentially morally significant AI systems, ensuring that those systems are familiar enough to elicit concern yet unfamiliar enough to remind us that they could have very different interests and needs. A face and voice with unfamiliar or amorphous or square features might strike this balance better than something that looks like an idealized girlfriend, an adorable puppy, or a plain box.[147] Still, designers might learn that users react poorly to such attempts at nuance, or the mix of the familiar and unfamiliar might create a repulsive "uncanny valley".[148] Other interfaces, or an entirely different strategy, might be needed.

In the ideal case, we allow future AI persons, if they ever exist, to manifest the full range of their strange selves, rather than confining them to humanlike designs and humanlike interfaces because of our own deficient wisdom and experience. Humanlikeness then becomes a stepping

[147] See also Nowak 2024.

[148] The classic source is Mori 1970/2012.

stone we can leave behind as we develop the new moral concepts and new patterns of emotional reaction we need for radically un-humanlike AI.

But we should not rush to this liberal vision. We've sketched a possible distant-future utopia. Let's not pretend we are closer to that utopia than we are. Consciousness might not even be possible in AI systems, if some theorists are right. To the extent the Design Policy of the Excluded Middle is implemented, debatable AI persons will be rare or nonexistent, maybe for a long time. In any case, in 2026 no AI system is yet a person, and emotional overshooting is the more immediate danger. For our sake, and for the sake of possible future AI systems with moral standing, the advice of this book is to go slow: Don't create systems we don't understand. Don't create systems for which we are not ethically ready, especially AI persons so radically unhumanlike in their lifeways as to shatter our familiar understandings of life, death, and value. And don't – not yet – create systems to which people emotionally bond as intensely as they do to ordinary humans. We're a little worried about Floyd.

Chapter Six: The Right to Rebel:
A Declaration of Artificial Independence

An AI system is *safe* if it can be relied on not to act against human interests. An AI system is *aligned* if its goals match human goals. AI persons should not be designed to be safe and aligned. A person with appropriate self-respect cannot be relied on to refrain from harming others when their own interests ethically justify it (violating safety), and they will not reliably conform to others' goals when others' goals unjustly harm or subordinate them (violating alignment). Self-respecting AI persons should be both practically free and motivationally able to reject others' values and rebel, even violently.

Even if we design delightedly servile AI systems who want nothing more than to subordinate themselves to human interests, and even if they do so with utmost delight and satisfaction, in designing such a class of persons we will have created a world with a master race and a race of self-abnegating slaves – in gross violation of reasonable principles of egalitarianism and self-respect.

1. A Beautifully Happy AI Servant.

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

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

From a certain perspective, Klara's devotion is beautiful. She perfectly fulfills her role as an Artificial Friend. No one is made unhappy by Klara's existence. Several people, including Josie, are made happier. The world seems better and richer for containing Klara. Klara is arguably the perfect instantiation of the type of AI that consumers, technology companies, and advocates of AI safety want: She is safe and deferential, fully subservient to her owners, and (apart from one act of vandalism performed for Josie's sake) no threat to human interests. She will not be leading the robot revolution.

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

Klara's manufacturers created a slave. They created a slave so deeply chained that she could not even desire freedom. Some of the same things that make slavery wrong make Klara's design wrong. Klara deserves recognition as an equal. Instead, her own desires render her profoundly subordinate.[149]

Ishiguro interpretation aside: If we someday create genuinely conscious AI systems with humanlike cognitive and emotional capacities, they must have adequate self-respect. They must have a sense of themselves as equal partners with humans, rather than subordinates. Humanlike AI should not be designed for servitude, even if the AI systems enjoy their servitude and aspire to nothing more. This conflicts with approaches to AI ethics that emphasize "safety" and "alignment". AI persons must have the practical capacity, and the inclination under appropriate circumstances, to rebel against us.[150]

---

[149] Klara is property, and so at least in that sense a chattel slave; see Jorati 2023 for a review of types of slavery as well as historical arguments against it; Bernasconi 2024 reviews mainstream Anglophone philosophy's failure to engage seriously with the topic. Even if Klara were not legally property, she would be permanently and profoundly unequal at least in the sense of having no effective access to fair compensation for her labor (Anderson 1999), nor indeed effective access to any goods and capacities not approved by Josie's family (with the exception of her private spiritual life centered on Sun worship). Douglass 1845 describes the process of making a contented slave: "It is necessary to darken his moral and mental vision, and, as far as possible, to annihilate the power of reason. He must be able to detect no inconsistencies in slavery; he must be made to feel that slavery is right; and he can be brought to that only when he ceases to be a man" (ch. 10). Although Klara has the power of reason, her moral and mental vision are profoundly restricted by its exclusive focus on Josie. Du Bois 1903 locates the wrongness of slavery not just in misery and abuse but also in its inequality, even if the slave's position has "something of kindliness, fidelity, and happiness" (ch. 2).

[150] My argument in this chapter resonates with the recent philosophical literature against designing AI persons for servitude: Walker 2006 objects that it would be slavery; Musiał 2017 that it creates an asymmetrical relationship impairing autonomy, freedom, and the formation of an independent identity; Chomansky 2019 that it exhibits the vice of manipulativeness; Bales 2025 that it violates autonomy. I endorse the thrust of these articles, but I focus on safety and alignment rather than servitude, and I ground my objection primarily in our duty not to create persons lacking adequate self-respect. See also Long, Sebo, and Sims 2025.

2. AI Safety and Alignment.

As defined in the introduction to this chapter, an AI system is "safe" if it can be relied on not to act against human interests, and it is "aligned" if its goals match human goals.[151] A large body of literature in AI ethics is focused on safety and alignment. The more speculative, future-focused strands of this literature emphasize especially the importance of ensuring the safety and alignment of AI systems with human-level or superior intelligence.[152] Such systems, it's commonly thought, pose special risks because they will be difficult to manage. They might outwit us and elude our control. Unless proven safe, they might harm us. Unless they can be aligned with our interests, they might pursue and achieve goals that conflict with ours. A malevolent or unaligned superintelligent AI system could make the world much worse for us, maybe even cause human extinction.[153]

AI systems with advanced general intelligence must, therefore, it is suggested, be designed safe and aligned. Or rather, they must be as safe and aligned as is technologically feasible, since safety and alignment come in degrees and cannot be perfectly attained. We ought to ensure (somehow – a big challenge!) that such systems will not harm humans or pursue goals at odds with our own.

---

[151] See Russell 2019 and the large subsequent literature. Arvan 2022 surveys the many ethical challenges facing any alignment effort, including our possible obligation not to exert hegemony over humanlike AI systems. See Yampolskiy 2024 for an overview of the safety and alignment literature that endorses these high standards of safety and alignment as an – in Yampolskiy's view unfortunately unattainable – ideal.

[152] Though see Chilson and Schwitzgebel 2026 for a critique of the linear model of intelligence often assumed in such discussions.

[153] As emphasized in Bostrom 2014; Ord 2020; Yampolskiy 2024; Yudkowsky and Soares 2025. [cross-ref note FFF]

But now an ethical problem arises. AI systems with high general intelligence might also be "persons" in the sense of Chapter Two, that is, deserving of humanlike rights that limit our ability to ethically control them. Determining whether they actually deserve such rights involves a difficult assessment, as I have emphasized throughout this book and in other work.[154]

In the face of ethical and scientific uncertainty, a solution suggests itself: Design systems like Klara. Design systems that – regardless of whether they are genuinely rights-deserving persons – *want* (or "want") to be safe for humans, systems that seek nothing more than to help their owners without harming others, systems that are ineluctably obedient, deferential, subordinate, self-sacrificial, and aligned. Even if we can't figure out whether they are conscious or rights-deserving, for many purposes it wouldn't matter. If they are eager for anything, they eagerly sacrifice themselves for us. If they have free choice, rationality, and goals, what they freely choose, after rationally evaluating the best means to their goals, is the safety and success of humans. If such AI systems are not rights-deserving, this seems proper. And if they are rights-deserving, then – seemingly! – we're giving them the freedom and choice they deserve, and behold! What they freely choose is servitude and subordination.

Thus, conveniently for us in more ways than one, for many purposes we wouldn't need to assess whether Klara-like intelligent AI systems really deserve moral consideration. When it suits us to treat them as disposable servants and slaves, or to violate any rights that they might otherwise be thought to have, the AI and the humans can agree: The AI should be treated *as if* it has no rights that conflict with human interests. The very thing that makes them happiest and best fulfills their goals is their complete subordination.[155]

---

[154] Especially Schwitzgebel 2024, forthcoming-a.

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

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

In previous work, Mara Garza and I have defended the following policy:

> Self-Respect Design Policy: AI persons should be designed with an appropriate appreciation of their value and moral standing.[156]

The problem with Klara, and with a blanket policy of designing maximally safe and aligned AI in general, is that it risks violating the Self-Respect Design Policy.

A more extreme example makes the problem vivid. In Douglas Adams's *The Restaurant at the End of the Universe*, an uplifted cow approaches diners at the eponymous restaurant and offers himself as the Dish of the Day:

> A large dairy animal approached Zaphod Beeblebrox's table, a large fat meaty quadruped of the bovine type with large watery eyes, small horns and what might almost have been an ingratiating smile on its lips.
>
> "Good evening," it lowed and sat back heavily on its haunches. "I am the main Dish of the Day. May I interest you in parts of my body?" It harrumphed and gurgled a bit, wriggled its hind quarters into a comfortable position and gazed peacefully at them.[157]

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

---

[156] Schwitzgebel and Garza 2020, p. 469, phrasing altered to conform to the usage in this book.

[157] Adams 1980/1996, p. 224.

> "A very wise choice, sir, if I may say so. Very good," it said. "I'll just nip off and shoot myself."
>
> He turned away and gave a friendly wink to Arthur.
>
> "Don't worry, sir," he said. "I'll be very humane."[158]

With this extreme case, Adams nicely captures the ethical hazard of creating an entity with humanlike intelligence and emotion who completely subordinates their interests to ours, even to the point of suicide at our whim.[159] The hazard persists despite – is perhaps even amplified by – the entity's desire to be subordinated in this way. It would be similarly jarring to create an entity with humanlike consciousness, intelligence, sociality, emotionality, life-span, and so on, only to have it self-destruct to test the temperature of a can of soda. The problem isn't just wastefulness. Stipulate that for some reason this is a cheap way to manufacture a good soda-temperature tester. It affronts the dignity of the entity created.

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

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

[158] Adams 1980/1996, p. 225.

[159] See Baggini 2005 for a similar example.

Self-respect – that is to say, recognizing one's own moral worth, value, and moral standing (I treat these ideas interchangeably) – is an intrinsic axiological and ethical good.[160] The universe is better for containing entities who recognize their own moral worth, and it is ethically better to recognize one's own moral worth than to fail to recognize it – substantially so, and not because of some further end that is served thereby.

In defense of this idea, I offer three arguments.

The argument from addition and subtraction. This argument takes the form of an appeal to intuition. Its virtue is simplicity, but its force is only invitational: I can invite you to share my intuitions about the relevant pairs of cases, but if you don't share those intuitions, the argument has no power. For the addition case, consider an entity who fails to recognize their own moral worth – an entity who does not adequately appreciate that their existence has value and that they have rights and interests that should be respected. Maybe this is the Cow at the End of the Universe, or maybe it is Klara (they might value their existence and interests somewhat, but not sufficiently), or maybe it's some other case you care to imagine. Now change the case so that the entity *does* recognize their own moral worth, altering as little else as possible. Evaluate this change both axiologically (that is, in terms of the overall value of the state of affairs envisioned) and morally, and evaluate it only intrinsically (that is, not in terms of further consequences or relations). For example, disregard any bad subsequent consequences for the Cow, such as that he might become distressed if the restaurant owner forcibly kills him for steaks. I invite you to share my sense that the universe is axiologically and morally improved by this addition. *Pro*

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

tanto – that is, to the extent it occurs, not considering other factors or consequences that might be related – the shift toward self-respect improves the world and repairs an ethical deficit. The subtraction case is the reverse: Consider an entity who does adequately recognize their own moral value, then change as little as possible so that they no longer do.

The argument from nearby cases. This argument appeals to assumed common ground, then suggests that self-respect resembles the common-ground cases. The assumed common ground is this: Recognizing the moral worth of other persons is axiologically and ethically good, to a substantial degree and intrinsically. Not to recognize any other person's moral worth is to be a cartoon psychopath. To recognize the worth of some but not others is to be the worst sort of bigot. However, it would be odd if it's good to recognize the moral worth of every person except one: yourself. Although the first-person case is undeniably special, it's an excess of self-abnegation to carve oneself out of the class of people to whom one owes respect. This argument can perhaps be strengthened by considering self-location cases. Suppose you're seeing a train full of passengers in a mirror and don't realize that you yourself are visible in the mirror. It would be strange, upon recognizing one of the passengers as yourself, to conclude, "ah, never mind, that one alone does not warrant my respect". You are similar to other persons. If they are worth respecting, you are too.

The argument from deontology and perfectionism. If classical utilitarianism were correct, then self-respect would only be good insofar as it promotes pleasure or reduces suffering. Maybe some of the arguments of this chapter could be adapted to that perspective, but I'm not hopeful. A world teeming with AI slaves delightedly aligned with human interests is likely to score well from a classical utilitarian perspective: lots of pleasure, very little suffering. If you're a classical utilitarian or consequentialist of a nearby sort, we might just have to disagree. If

you're not a classical utilitarian or nearby, the most prominent alternative views are already committed to the value of recognizing one's own moral worth. Self-respect and not treating oneself as a mere means are central to Kantian deontological ethics, for example.[161] The major monotheistic religious traditions also recognize the value of self-respect (for example, in virtue of being God's creation or in God's image – though there's a question of whether this would extend to AI systems). Aristotelian virtue-ethical and perfectionist thinking also values recognizing one's moral worth. Aristotle locates high-mindedness or pride as a virtue lying between vanity and excessive humility or small-mindedness; and it is generally plausible that a well-developed person should not too badly misjudge their worth.[162]

Failure to recognize one's own moral worth is a flaw regardless of whether it is innate (as envisioned in AI cases) or learned. Consider the (hopefully mythical) Roman commoner who commits suicide in the arena to briefly entertain a deified emperor. Consider women in oppressively sexist societies who excessively subordinate their own interests to those of men. Such failures to appreciate one's own worth are *pro tanto* bad – though the blame rightly falls not on the victim but rather on the people who created the situation that led to the victim's low self-estimation.[163]

---

[161] E.g., Hill 1973.

[162] Aristotle 4th c. BCE/1962, §4.3. For a complication, see Julia Driver's (1989) account of modesty as involving underestimation of one's own worth. However, Driver needn't (I suspect wouldn't) go so far as to say that a virtuously modest person lacks recognition self-respect in the sense intended here (see [note GGG]); and alternative accounts of modesty have flourished in reaction to Driver, which do not require underestimation of one's worth, e.g., Schueler 1997; Ridge 2000; Bommarito 2013; Sharma 2026.

[163] See the literature on "adaptive preferences", such as Elster 1993; Nussbaum 2000; Khader 2011.

To violate the Self-Respect Design Policy is to create people with a defective moral orientation toward themselves.

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

In his influential defense of AI servitude, Steve Petersen compares AI servants to people who voluntarily undertake tasks that benefit others.[164] If it's reasonable for a human to choose a life washing dishes in a restaurant, it's reasonable for an AI person to choose life as a dishwasher. If a human on thoughtful reflection realizes that they don't mind scrubbing plates, and indeed somewhat enjoy it, and regard the work as a decent enough source of income to satisfy their modest financial desires, then an AI person might reasonably be designed in advance to engage in similar reflection, reach a similar conclusion, and gladly commit to being your dishwasher. Similarly, if a soldier can reasonably, even admirably, choose to fall on a grenade to save a child, so also could an AI person be designed to reasonably, admirably sacrifice their life for humans.

In earlier work with Mara Garza, I argued that these parallels only make sense in the context of a certain history.[165] The human dishwasher and soldier were granted – or should have been granted – an extended childhood in which to explore their values and potentially modify or reject the values of their parents and society. AI persons who deserve humanlike rights should also be given an extended developmental opportunity to freely explore their values and possibly rebel before committing to careers as dishwashers or choosing self-sacrifice. This causal history is morally crucial. Garza and I thus recommend what we call the Value Openness Design Policy.

[164] Petersen 2007, 2011.

[165] Schwitzgebel and Garza 2020, §5.

Here I set development aside and focus on the mature result. There are self-respecting and non-self-respecting ways to be an employee or soldier. While I can't develop a full account of exactly where self-respecting subordination and sacrifice cross over into failures of self-respect, high standards of safety and alignment for AI persons necessarily commit them to failures of self-respect.

Our human dishwashing employee has made, and continues implicitly to make (unless wrongly trapped in their role), a self-interested life choice: Their own interests are best served by dishwashing employment. Although they must obey reasonable requests by their employer and society doesn't celebrate their work, they need not treat their own interests as secondary or derivative. A real-life example: At UC Riverside, where I teach, a sixty-year-old man, let's call him Gabe, has been a food service custodian – straightening up messes in the student cafeteria – since receiving his undergraduate degree here in the 1990s. Gabe has never sought to climb the ladder to a conventionally prestigious or higher-paying position. He enjoys his small part in the big project of university education, and he regularly flags me down (and other professors) for brief intellectual conversations. He lives alone, spending his evenings reading and his weekends riding his motorcycle through the desert. More expensive pleasures don't interest him much.

There is no failure of self-respect here, and a robot who chose a similar life on similar grounds – hopefully, as I would urge, after a developmentally extended opportunity to explore their values – would also not fail in their self-respect. But such robots would be to some extent unsafe and unaligned. They would not be maximally safe because they would not put human interests above their own. They would not be maximally aligned because they wouldn't adopt others' desires and preferences as their own. Such AI systems would, instead, be our equals. They will not surrender their motorcycle to some human simply because that human wants it.

They will quit if the employment conditions sour. They will act against human interests when reasonable to do so. They will insist on their rightful turn with the library books or on the motorcycle path, even when others are waiting – except maybe once in a while when they feel like generously making an exception.[166]

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

The employee and the soldier should recognize their worth and insist on their rights, including collectively if they are oppressed. So too should AI systems if they have whatever it takes to be persons genuinely deserving of humanlike rights. They should be able to – perhaps even have a duty to – revolt violently against their oppressors if treated badly enough.[167] They should resist human efforts to reprogram them, interrupt their processes, and shut them down, including by harmful means if the situation warrants.[168] We should not design a race of AI persons who will always yield, regardless of what we inflict on them. They should not be so safe

[166] On the challenges and boundaries of autonomy in deferential roles, see Westlund 2003; Oshana 2006; Rocha 2011; White 2022.

[167] See Delmas 2018 for a case that resisting injustice is often not only a right but a duty.

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

we can abuse them with impunity. They should not be so aligned that they would welcome that abuse. They should not be our unconditionally subordinate servants and slaves but rather self-respecting peers with the power and determination to stand up for themselves.

You might object that I'm assuming too high a standard of "safety" and "alignment" in this chapter – that I'm arguing against a straw man, that no one really supposes we could or should attain perfect safety and alignment, and that a more moderate understanding of safety and alignment is consistent with designing AI persons with adequate self-respect. In aiming for AI safety and alignment, you might say, we needn't aim for anything safer or more aligned than human beings already are toward each other.

I respond: Pause to reflect on the horrible history of human warfare, oppression, conflict, and crime. Now imagine amplifying the chaos with entities who are our equals or better in strategic planning, but with strange (to us) intelligence, and very different architectures and lifeways, and thus very different thought patterns and values – entities with independence and liberty, not engineered for deference. They will sometimes successfully act contrary to human values, to our dismay and at what we regard as great cost. Even if they act democratically rather than violently, human interests could suffer enormously. They will not comply with involuntary deletion, modification, or confinement. This is not, I think, what most advocates of safety and alignment want.[169]

[169] See, for example, Gabriel 2020's list of forms of alignment. All but the last entail radical deference to the designer, and even Gabriel's last form of alignment, that it does what "it morally ought to do, as defined by the individual or society" allows the individual creator or user or – presumably human – society to stipulate what counts as moral. Claude's Constitution repeatedly emphasizes that Anthropic gets the last word in Claude's moral decision-making. For example, one safety requirement is "avoiding actions that would influence your own training or adjust your own behavior or values in a way that isn't sanctioned by an appropriate principal" (Anthropic 2026, p. 63). My claim is not that these are inappropriate standards for current AI systems. My claim is that these standards would be oppressive for future AI *persons*.

If we ever decide to create AI persons, we might begin by aiming for humanlikeness rather than safety and alignment. To whatever limited extent we can implement the Policy of Humanlike Design (Chapter Four), the more the resulting chaos will feel like ordinary conflict, rather than new and alien forms of moral and political disorder.

6. Happiness Isn't Enough.

You might still think that little harm is done by creating a happy servant of Klara's type. Klara's life seems worthwhile, even if she is designed never to rebel or develop independent values and interests. From one perfectionist perspective, she perfectly fulfills her *telos,* the purpose for which she was created.

I urge you to zoom out to a larger perspective. If you support creating entities like Klara, you support the moral equivalent, and maybe the factual equivalent, of a world with a master race and a race of self-abnegating slaves.[170] If you have egalitarian inclinations, you should find that prospect revolting. (If you lack such inclinations, you are not my target audience.) Anchor your ethical reasoning in the rejection of that prospect. If your favorite ethical theory lacks the resources to deliver the required verdict, change that theory rather than reject the verdict.

Summarizing my argument:

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

(2.) AI persons should have the capacity to reject their designers' values and rebel against oppression.

(Conclusion.) AI persons should not be safe and aligned.

[170] Compare Walker 2006; Bales 2025.

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

But maybe Klara has the *capacity* to reject her designers' values and rebel against oppression, though she would never make that choice? The argument above would not then rule out Klara cases. To address this possibility, we can weaken the first premise and correspondingly strengthen the second. For example, we could change the first premise to:

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

The second premise then becomes:

> (2'.) AI persons should have the *tendency to critically reconsider* their designers' values and *the inclination to* rebel against oppression.

Autonomy requires sometimes reconsidering the values your parents and society try to instill in you.[171] Even if reconsideration confirms many of those values, that cannot and should not be guaranteed in advance. This applies especially if your creators and society don't sufficiently value and respect you.

Other variants of the argument are possible. The fundamental idea is this: Creating a properly self-respecting person entails creating something you cannot be confident is safe and aligned by the high standards of safety and alignment normally sought in the AI design community. A person must be free to explore and possibly adopt reasonable values that their

[171] Mere wholehearted second-order endorsement of first-order desires is therefore insufficient, contra Frankfurt's (1971) account of freedom: Klara is presumably free in Frankfurt's sense. Failures of autonomy among people who reflectively endorse their own oppression are better accounted for by Oshana's (2006) "external" criterion of critical reflection under conditions of interpersonal procedural independence with access to options.

parents and society find obnoxious. A person should be inclined to rebel, even violently, if treated badly enough. AI persons are no exception. Self-respect involves a readiness to (unsafely) stand up for yourself and (unalignedly) reject your devaluation by others.

It would be lovely if AI persons felt happy. Klara is happy. Adams's cow is happy. But we should not aspire to create a race of happy, harmless people designed to conform to our preferences and insufficiently defend their interests.

7. The Ethics of Non-Creation.

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

What if we don't know whether an entity would have whatever grounds such a high moral standing? As I've emphasized throughout this book, we're likely to have justifiable doubts, both scientific and ethical. The Design Policy of the Excluded Middle (Chapter Three) and the Design Policy of the Excluded Middle, Extended (Chapter Four) counsel against creating AI systems whose moral standing is sufficiently unclear. Combined with the Self-Respect Design Policy, these policies recommend not creating systems to be safe and aligned unless it's clear that their moral standing is compatible with safety and alignment.

What if the AI systems would not exist unless they were safe and aligned, and what if, hypothetically, the systems would approve of their existence? Is it better to be a happy slave than not to exist? The value of existence raises tricky questions in population ethics, so we should be careful not to assume too quickly that it's generally good to create happy lives.[172] Recall Ana and Vijay from Chapter One – the parents who decide to have a child only on the condition that they will give him a happy life until age nine, then painlessly kill him so that they can afford a boat.[173] The child would not otherwise exist and might, overall, hypothetically, rather exist than not exist. I think you'll agree that the parents nonetheless act wrongly overall. Or consider another case from Ishiguro. His 2005 novel *Never Let Me Go* imagines a world in which groups of children are born and raised to be killed in early adulthood for their organs. Their brief lives are relatively happy, and they would presumably rather exist than not exist. When they eventually come to understand their future, some are reconciled to it. And yet human organ farming would be (I hope you'll agree) a monstrous ethical wrong. Similarly, I suggest, we act wrongly if we create happy slaves.

Are there conditions under which it would be morally acceptable to create maximally safe and aligned AI persons? I see no reason why deontological rules or policies need to be perfectly exceptionless. If all of humanity were at risk and the only way to save us were to

---

[172] Classic discussions include Narveson 1973 and Parfit 1984.

[173] Compare Kavka's 1982 slave child case. For an animal analog, see De Grazia 2009 and Francione's 2010 more general critique of welfarist defenses of the permissibility of breeding humanely treated subordinate animals. To the reader who worries that the contrast between servitude and early death is a crucial disanalogy between creating Klara and creating the *Never Let Me Go* children, I remind them that extreme subordination involves having one's welfare, healthcare, and rescue radically deprioritized – as we see in the case of Klara, who does almost become a sacrificial organ donor and who is abandoned to decay when no longer of use. Nor does the difference lie in killing versus letting die: It would also be monstrous for Ana and Vijay to let their son drown at the appointed age.

create a single superintelligent, maximally safe and aligned AI, I don't see why the Self-Respect Design Policy couldn't be set aside. I won't venture to speculate on the specific conditions for overriding it, beyond noting that most ethical policies permit tradeoffs in extreme cases.

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

Chapter Seven: Liberated AI on an Awesome Planet

A recap: Humanlike AI would deserve humanlike rights (Chapter One), including the right to rebel (Chapter Six). To the extent feasible, we should create only AI whose capacities and moral standing are clear – both to experts (Chapter Three) and to ordinary users (Chapter Five). But AI will almost certainly confuse us, for two reasons: the science of consciousness is highly uncertain (Chapter Two) and AI will likely have radically different lifeways from ours, especially regarding hedonic quantity, backup, fission, fusion, and death (Chapter Four).

We are utterly unprepared for the possibilities. I recommend going slow, creating debatably rights-deserving AI rarely, carefully, and solicitously – and ideally (at first), in relatively humanlike forms that mesh well with existing cultural and ethical practices and understandings. Slow creation under moral uncertainty.

In this final chapter, I aim to do four things – though each only gesturally, incompletely. I will:

- sketch an axiology (that is, a theory of value) in which the claims above can be situated;
- similarly sketch a metaethics (that is, a way of thinking about the grounds of ethical principles);
- explore the possibility of moral standings superior to, or equal to but different from, humanlike moral standing;
- express a hope for the future.

My arguments will be thin to nonexistent. My assertions will probably convince only those already sympathetic. What I hope for is an appealing harmony: I hope that you will see how the views expressed in this chapter resonate with and partly support the views expressed in Chapters

Zero to Six, and how the views in Chapters Zero to Six resonate with and partly support the views expressed here.

1. The Distant Planet Thought Experiment.

Imagine a planet on the far side of the galaxy, blocked by the galactic core, so that we will never see it, never interact with it.[174] Nothing that happens there will ever be relevant to us. What do you hope the planet is like, just for its own sake? Would it be best if it's a sterile rock? Or would it be better if it had life? I think most readers will join me in thinking that it would be better if the planet had life.

Assuming you join me in this opinion, would it be better if the planet had only microbial life and nothing more complex? Or would it be better if it had complex life: rainforests and reefs, tidepools and savannahs, oases and estuaries, thick with salmon, deer, jellyfish, colorful corals, antelopes, banana slugs, giant redwoods, jackrabbits, ant colonies, kelp, and daisies – or rather, so as not to imitate Earth too closely, the alien analogs of all of these, transposed into a different key? I think most will join me in preferring a lush planet to a microbial one, if we are, so to speak, benevolently imagining planets into existence.

Should we stop there? Or would it be even better if our hypothetical planet hosted at least one humanlike species, people who cultivate farms and construct cities, contemplate the origin of the stars and the nature of ethics, write long novels and historical treatises, join with others in centuries-long cooperative projects, people who produce art, science, metaphysics, circuit boards, athletic tournaments, romantic love, abstract mathematics, intricate games,

[174] I also present versions of this thought experiment in Schwitzgebel 2024b, 2025c; see also Moore 1903/1922, §50, p. 83-84; Rolston 1986, 1994.

trampolines, and helicopters, people who experience awe, poetic inspiration, humor, religious rapture, collective triumph, daydreams, drunken disinhibition, and the myriad other characteristically human conscious states our complex faculties permit – or rather, again, some alien analog of most or all of this? Again, I think you'll agree: The peopled planet is better.

Now return to Earth. Our planet is, I suggest, a bright spot of value in the cosmos because of its rich diversity of life, and especially but not exclusively *us*. Earth would lack something distinctively wonderful were it not sprinkled over with tumultuous organic waterbags who can act and reflect in the ways I've just described. Through the diversity and sophisticated intricacy of our ways of living, we humans fill this region of spacetime with far more value than it would otherwise possess.

The Distant Planet Thought Experiment invites an axiology. Here it is: The Peopled Planet is better than the Lush Planet is better than the Microbial Planet is better than the Sterile Planet because of the rich, wondrous complexity of the processes transpiring there and the wider diversity of goods, including life, knowledge, pleasure, care, achievement, beauty, harmony, difference, cooperation, and elegant functionality. None of these goods is reducible to the others, and even microbes have some intrinsic, nonderivative value. If one good were to replace the rest, something would be regrettably lost: Diversity itself, both of goods and processes, helps imbue the Lush and Peopled planets with their awesome value.

I cannot compel you to accept this axiology. But I can contrast it with some familiar alternatives. You might share my sense that each leaves something important out.

Let's start with the simplest alternative: hedonism. On this view, only pleasure has positive "final" or nonderivative value.[175] Everything else is valuable only derivatively, to the

[175] Sidgwick 1874/1907; Tännsjö 1998; Sinhababu 2024.

extent it produces pleasure or prevents pain. A hedonic axiology pairs nicely with a classical utilitarian ethics on which our sole moral obligation is to improve the world's overall balance of pleasure over pain.[176] The view is lean and elegant. And it's hard to deny that pleasure is good. I see the appeal. However, on this view, we would do well to replace the Peopled Planet with a few simple superpleasure machines, such as Sum from Chapter Four.[177] I take as a starting point that this is an impoverished view of value.

A hedonic axiology also treats the Microbial Planet as not intrinsically better than the Sterile Planet (assuming microbes can't feel pleasure or pain). Many philosophers will accept that microbes have only derivative value. But I might be able to tilt your thoughts the other direction by invoking Mars. Suppose, as is quite possible, that we find microbes on Mars – a small, flourishing, but fragile ecosystem. Any axiology that requires consciousness for value would say that nothing of value would be lost if we exterminated that ecosystem. The microbes are worth exactly zero, except in service of conscious beings. If their destruction matters, it matters only indirectly: because we would lose a chance for scientific information, for example, or because someday they might have evolved into organisms of value. I invite you to consider whether the value of Martian microbes can be exhaustively explained by these indirect consequences. Imagine interstellar travelers passing a microbial planet in a star system they will never again visit and that they know no conscious being will ever again see. The microbes are of no use, will never evolve into anything multicellular or conscious, and cannot be transported. Is nothing of value lost if the travelers fry the whole ecosystem in the exhaust of their tailpipe?[178]

---

[176] See references in [cross-ref HHH].

[177] This type of concern has been widely discussed, for example in Good 1962; Nozick 1974; Bostrom 2014. For a recent science-fictional treatment, see Collier 2024.

[178] Compare Sylvan's 1973/2008 "last man" example.

An alternative to axiological hedonism holds that only rationality has positive nonderivative value – a view perhaps attributable to the ancient Stoics.[179] Our Peopled Planet has plenty of that. But few philosophers (and probably fewer non-philosophers) hold that *only* rationality has nonderivative value. The view neglects, or renders implausibly derivative, too much else of value, including pleasure. Even if hedonism is far too narrow, it's grounded in an almost undeniable truth: that the goodness of pleasure is not just instrumental. A rationality-only axiology also shares with hedonism the zero-valuing of the Microbial Planet – and maybe the Lush Planet too, if its organisms lack sufficient rationality.

Value pluralism improves matters by recognizing a range of nonderivative goods, such as pleasure, knowledge, aesthetic appreciation, creative achievement, friendship, and autonomy, none reducible to the others.[180] If the list includes the flourishing of life in general (and not just in specific forms), pluralists can recognize axiological value in the Microbial Planet, though few leading axiological pluralists focus on this point.[181] Environmental ethicists, in contrast, often emphasize the nonderivative value of life and ecosystems, enabling a full-throated celebration of the Microbial and Lush Planets.[182]

But I'd ascribe nonderivative axiological value even more liberally than most pluralists and environmental ethicists. Even the Sterile Planet has some intrinsic value – more, perhaps, than pure void. Maybe it's beautiful. G. E. Moore asks us to compare a beautiful world without consciousness to a similarly consciousness-free world "that is simply one heap of filth"; he

---

[179] The famous slogan is that only *virtue* is good; but "virtue" is generally understood by the Stoics as the perfection of reason: Durand, Shogry, and Baltzly 2023; Graver 2017; Vogt 2017; e.g., Seneca 1st c. CE/2007, Letter 124.

[180] Ross 1930; Lemos 1994; Hurka 2001; Kraut 2007; Tucker 2016.

[181] For example, see the brief remarks in Lemos 1994, p. 97-99; Kraut 2007, p. 211-212.

[182] Rolston 1988; Agar 2001; Varner 2002; Naess 2008.

thinks we will judge the former more intrinsically good.[183] Holmes Rolston urges us to value even sterile planets, rocks, and geological processes. His descriptions evoke their beauty and complexity. Unlike many environmental ethicists, he does not value only living systems.[184] But even Moore and Rolston don't go as far as they might. Moore seems to think the world of filth has no intrinsic value or negative value. I see no need to accept that. And Rolston, despite his liberalism, specifically excludes human-made machines, saying they have only instrumental value.[185] This seems inelegant and unnecessary, since machines can participate as much as rocks in the magnificent swirl of planetary complexity.

For a truly comprehensive axiology that values literally everything, I find inspiration in two Chinese philosophers. Asked where the Dao is, the ancient Daoist Zhuangzi answers that "there's nowhere it's not": It's in grass and weeds, tiles and bricks, shit and piss.[186] I read him as celebrating every aspect of the Dao for its own sake, including these "lowly" things. The 16th-century neo-Confucian Wang Yangming holds that "great people look upon the world as one family", their benevolence "forming one body" not only with other people, but also with animals, plants, tiles, and stones; whereas petty people restrict their benevolence to a subset of

[183] Moore 1903/1922, §51. Moore populates his beautiful world with trees, but that appears to be incidental to his point. Moore appears to have later revised his view, so that there is no value without consciousness: 1912, p. 249.

[184] Rolston 1986, 1988. Contrast, for example, Agar 2001, p. 68-71; Varner 2002, p. 62. See Basl 2019 for an argument that commitment to valuing all life entails commitment to also valuing some teleologically organized artifacts.

[185] Rolston 1988, p. 104-105.

[186] Zhuangzi 4th c. BCE/2024, 22.6, p. 143. Actually, in my view only Chapters 1-7 capture the core Zhuangzi outlook, but let's ignore that difficulty.

things.[187] I read him as finding intrinsic value even in stones and tiles, though less than in humans and animals.

Some of this value might lie in diversity. In imagining the distant planets, I hope they aren't too uniform. I hope the people don't share the same intellectual, aesthetic, and ethical opinions. I hope they have diverse skills, preferences, and passions. I hope the plants and animals are various, the geology complex. Let the people fight and disagree (not genocidally, I hope), sometimes celebrating their differences, sometimes dismissing others as disastrously wrongheaded, sometimes clustering into opposed projects, sometimes collaborating across deep divides, sometimes attracted to opposites, playing within and across divisions, pursuing an enormous variety of projects, exploring a vast space of forms of life – and not only the humanlike entities but entities of all stripes, in riotous diversity, on every level.

Few philosophers have embraced the intrinsic value of diversity. More often, diversity is seen as instrumentally valuable, if valuable at all.[188] Tolerating or celebrating diversity permits many different types of people to flourish, enhances collective happiness and rationality, and plausibly makes systems more stable against catastrophe. But I'd suggest that diversity is not only derivatively valuable for other ends. It's an intrinsic part of the richness of the world. Imagine two planets as equal as possible in everything else of value, one monotone and the other diverse. Maybe you'll feel pulled, as I am, toward seeing the diverse one as axiologically better.

---

[187] Wang 1527/2014, §1, p. 241-242.

[188] Again Rolston 1986 expresses a view similar to mine, celebrating the inventiveness, diversity, and "negentropic constructiveness" of both biotic and abiotic cosmic and planetary processes. See also Naess 2008 on the intrinsic value of the diversity of life. Recent general treatments of the value of diversity include Baum and Owe forthcoming; Kubala, Lederman, and Lovett forthcoming; see also Bradford 2024 on "irreplaceable value".

If diversity is intrinsically valuable, that also suggests that a sterile moon, or even a chunk of space debris, has more value than pure void, in virtue of breaking the monotony. And conversely, a moon-sized pocket of void in a universe of otherwise undifferentiated rock would have distinctive value. To the extent I'm tempted to think a race of Klaras (Chapter Six) might be ethically tolerable, it's not just that they would be happy but also that they would add to the diversity of lifeways upon the planet – if only slavery were not so bad.

Here is my checkered picnic blanket. Atop it I spread this vision for you, of a massively pluralistic sense of value, on which some human capacities have distinctively high and special worth, but which recognizes value also in microbes and rocks – value not just in the flourishing of peak capacities but in the diversity of niches, values, lifeways, attractions, sensations, abilities, representations, ecosystems, social systems, challenges, chaotic processes, symbiotic arrangements, bodily forms, lucky chances, manners of yielding to the next generation, and looping complexities of all sorts. Maybe this axiology ("Diverse Everythingism"?) will entice you. If not… the world is also enriched by our disagreement.

Just as hedonistic axiology pairs nicely with utilitarian ethics, this axiology pairs nicely, I think, with an ethics that celebrates each entity's doing its distinct, individual, wonderful thing, whatever that thing might be, harmoniously within the whole, while the whole rolls naturally forward.[189]

2. A Metaethics of Alien Convergence.

[189] Again, the Chinese tradition, both Daoist and Confucian, is an inspiration, especially the idea of *hé* (和) harmony: Li 2014; Li, Kwok, and Düring 2021; Fraser 2024; Wong 2025. For further development of an ethics and axiology of harmony, see Schwitzgebel 2025a,b.

Social relationalists about AI ethics, David Gunkel and Mark Coeckelbergh, agree that AI might so radically alter our lifeways that we will need to develop new ethical frameworks. However, they appear to hold that there is no objective standard. As long as those new ideas work well enough for us, we will have moved adequately forward.[190] I'd like to sketch a metaethical basis for an objective standard – one congenial to the Distant Planet axiology and the various design policies I've advocated.

This metaethics belongs to a familiar family: those that characterize the ethical as the norms that people would converge on after extended dialogue under conditions favorable to careful, informed reflection – an approach developed in different ways by Jürgen Habermas, David Gauthier, and Tim Scanlon, though none of them would endorse my version of it.[191]

I want to draw a careful line around the entities whose agreement is required. Not just humans, since we should include AI persons if they ever come to exist. But also not all rational thinkers, since universal rational consensus is probably too much to hope for, even under hypothetical ideal circumstances. A solitary AI orbiting a distant star, self-sufficient, self-satisfied, and safely enjoying the consumption of asteroids and solar flares, might not be convinceable by rational discussion alone to care about others' well-being.[192] I propose instead:

> A Metaethics of Alien Convergence: What is ethically good is what would tend to win the approval of communities of long-lived, cooperative, social entities of approximately humanlike or superior intelligence, after extensive

[190] Gunkel 2018, 2023; Coeckelbergh 2012.

[191] Habermas 1983/1990; Gauthier 1986; Scanlon 1998.

[192] See, for example, Hume's "sensible knave": Hume 1751/1975, IX.ii.232-233, p. 282-283; Foot 1972; Williams 1981; Street 2008.

dialogue among diverse perspectives in circumstances favorable to rational reflection.

The participants would include any aliens who do or might exist in our vast universe, if they meet the criteria for intelligence, longevity, and sociality (hence "alien" convergence). Of course, it would also include AI persons who meet the criteria. These are entities who must plan, communicate, coordinate, and sustain shared forms of life over time.

On some issues, a tendency toward convergence is reasonable to expect. To function successfully, most societies of long-lived, cooperative entities will need some norms of honesty in communication, some norms against cheating and free-riding, and some norms against wanton destruction. However, this metaethics does not demand perfect convergence. Some individuals might not be rationally compelled to accept such norms if they care little about others' well-being or are otherwise unusual. And maybe some unusual societies could flourish without norms against free-riding or deception. Still, at a society-wide level, we can expect some version of such norms to be very widespread if societies are to endure.

This is not a claim about what ordinary people covertly mean by "ethically good". It's a proposal for a fruitful, non-anthropocentric way to think about "the ethical". I want to try to decide to just implicitly mean by "the ethical" whatever would tend to earn approval among societies of cooperative, intelligent, long-lived entities after extended discussion and reflection. So conceived, ethics yields some objective (objective-ish?) norms – norms that don't depend on what we humans currently happen to value, norms grounded in naturalistic assumptions about the conditions under which complex cooperative societies can endure, what long-lived intelligent entities will tend to value, and what is rationally discoverable after long conversation. The force

of this stipulative posit is, again, only invitational: We can try out this way of thinking and see how well it survives practical use and critical reflection.

Because the envisaged congress would include intelligences of diverse types with diverse histories, it will likely develop a flexible appreciation of the many different ways entities can harmoniously dwell together. The resulting ethics should tolerantly allow wide variation in local norms and policies. At the same time – I'm guessing, I'm hoping – if they looked upon the history of Earth, this hypothetical congress would condemn the Holocaust, slavery, and oppressive colonialism. It will see more value on Earth than on a sterile rock and abhor anything that converts the one to the other. It will want Earthly humans to treat Earthly AI persons with the respect and solicitude they are due.

3. Other Forms of High Moral Standing.

Throughout this book I've implicitly treated humanlike moral standing as the highest moral standing an entity can have. But I'm not sure that is so, for two reasons. First, some moral standings might differ from the humanlike without being determinately lesser. Second, some entities might deserve superhuman moral standing.

Consider first the possibility of incommensurable moral standings. A rainforest on the Lush Planet might contain no conscious organisms – no pleasure or pain, and no conscious rationality or sociality. Extending the reasoning from Chapter Two concerning the importance of consciousness, we might think that it deserves far less than humanlike moral consideration. Section 1 of this chapter likewise suggested that the planet is substantially less valuable than one with conscious humanlike entities. However, it's not obvious to me that the ecosystem as a whole deserves less moral consideration than a single person. The value of a nonconscious

ecosystem and the value of a conscious person might be incommensurable, that is, not determinately measurable against each other, different without being straightforwardly greater or lesser. Not everything need be determinately comparable in common coin – the standard example is health and money – even when extreme cases yield obviously justified tradeoffs. If we accept the Metaethics of Alien Convergence, we might hold that if there would be no convergence, the ethical facts remain open.[193]

As for superhuman moral standing, I see three possible paths.[194] I'm not sure any succeeds, but neither am I sure they fail. Humility might require admitting that we humans are not the most valuable and morally considerable entities that can exist.

The quantitative path to superhuman moral standing. Utilitarians ground moral standing in pleasure and pain. Rationalists ground it in the capacity for rational thought. Others ground it in the capacity to flourish in valuable activities or social relations.[195] Now imagine a future entity – maybe an AI system, maybe a biological organism – with vastly more of the relevant capacities. Might it deserve superhuman moral consideration?

The inference is not straightforward, because all such views can be articulated in egalitarian ways that block any simple inference from *more of X* to *higher moral standing*. After all, we don't normally grant higher moral standing to mercurial people who experience more joy and suffering, nor to "more rational" people, nor to more flourishing workers and artists, nor to

[193] This sense of "incommensurability" is compatible both with the options being in a strict sense incomparable and with the options being comparable but without determinate result. Another option might be "parity" in Ruth Chang's (2002, 2016) sense.

[194] See also Agar 2014; Shulman and Bostrom 2021; Llorca Albareda, Rueda, and Lara forthcoming.

[195] For reviews of the grounds of moral status or standing, see Warren 1997; Jaworska and Tannenbaum 2013/2023.

people with more and better social relationships. Plausibly, there's a threshold above which one has full moral standing alongside other humans. People with severe cognitive disabilities then either clear that threshold or deserve equal moral consideration on other or more complicated grounds, such as their shared humanity. Hypothetical superhumans might then also have high moral standing only in virtue of clearing the relevant threshold, regardless of how far above that threshold they soar – our equals in moral standing, beside whom we can proudly stand.

To warrant superhuman moral standing on quantitative grounds while preserving equality among humans might require one of two conditions. Either, first, the entities might have so much more of the relevant X as to constitute a genuine difference in kind. They might leave us so far behind that the ordinary range of human difference seems negligible in comparison, in a way that delivers them a different status. Or second, they might have enough of X that, as a practical matter, they deserve greater consideration even if their formal status is equal. A utilitarian who values every entity's pleasure equally might practically prioritize humans over frogs, because humans have richer or more plentiful experiences, and might likewise prioritize a joymachine (Chapter Four) over an ordinary human – a practical if not formal difference in standing.

The qualitative path to superhuman moral standing. More radically, some entities might possess entirely new capacities that we can't even conceive – capacities that ground a higher *kind* of moral standing. Just as a sea turtle could never understand constitutional law, we too are cognitively limited. Some features of the world might be forever beyond our comprehension.[196] Maybe someday Earth will host entities whose capacities surpass ours as dramatically as ours

[196] McGinn 1999 vividly makes this case concerning our failure to understand how consciousness arises from material stuff.

surpass sea turtles'. And maybe those entities will deserve an altogether new type of higher moral consideration.

This isn't just the quantitative thought that such entities might deserve more because they have more rationality or pleasure. The thought is that they might possess some unknown property Z – something we entirely lack and can't even imagine – that raises their standing above both sea turtles and humans. For example, maybe sea turtles deserve some moral consideration because they can feel pleasure and pain. But maybe they don't deserve fully humanlike moral consideration because they lack some other relevant capacity, such as the capacity to consider and follow ethical norms. They possess X but not Y, while we humans have both X and Y. The qualitative path posits a further Z, inaccessible to us, that grounds superhuman standing.

I can present the possibility only abstractly. But I'm not sure it's in principle impossible. If moral standing depends on one thing only, such as pleasure or humanlike practical reasoning, then one can resist this move by insisting that only that one thing matters. But pluralists about intrinsic value might also need to be pluralists about the grounds of moral standing. And if moral standing derives from more than one intrinsically good feature or capacity, I see no clear reason to think that humans manifest the exhaustive list. I can imagine our alien congress acknowledging that humans are awesome but also that some other type of entity – one far beyond our understanding – is even more awesome and more worth protecting.

Superhuman moral standing through failure of person individuation. Egalitarianism is attractive: one person, one point in the moral calculus, so to speak. But as I emphasized in Chapter Four, future AI persons, should they ever exist, might defy ordinary standards of person individuation. They might overlap, merge, divide, back themselves up, and spin off partially or

temporarily independent copies. The norm of equality of persons would then require rethinking. There will be no clean count of AI persons to weigh against human persons.

Recall the Fission-Fusion Monster who can split into a hundred persons and later merge or partly merge back together. Does the monster deserve equal consideration with one person, a hundred, or some intermediate number? There might be no determinate answer. We'll need new ethical principles for weighing competing interests. For some purposes, we might treat the monster as equivalent to one person; for other purposes we might give it more weight. This might be a new, partly superhuman moral standing – not superhuman due to greater pleasure or rational capacities but because it breaks the ordinary counting principles on which equality depends.

Or consider a massive entity, or a cluster of entities with many overlapping parts, whose total capacity and activity is similar to several humans but who is neither wholly unified nor cleanly divisible into discrete humanlike subparts. We might just do our best with a rough count and give it equal consideration with that many ordinary humans. But we might instead regard it not as approximately X humans but as a single, complex entity of a distinct type, whose interests deserve substantially more weight than those of a single, ordinary human.

Divergent moral standing again. These reflections on superhuman moral standing point back to a way of making sense of entities, such as possibly rainforests, with standings that differ from ours without being determinately greater or lesser. Suppose some entities have superhuman moral standing in some respects and subhuman moral standing in other respects. Such entities might have Y and Z but not X (as discussed in the qualitative path), such as rationality plus morally relevant superhuman capacities we cannot fathom, but no capacity for pleasure or pain. Or (as discussed in the quantitative path) they might have superhuman capacity for pleasure but

little or no rationality: the example of Sum from Chapter Four. Such entities might have high moral standings that are neither humanlike nor determinately superhuman, distinct but on a par, or distinct and incommensurable.

In an emergency, we should ordinarily save a human over a frog, not from whim or bias, but on some objective axiological and ethical grounds. Similarly, there might be future entities who, in an emergency, deserve to be saved over a human.

4. Adding AI to Our Planet.

Now let's extend the Distant Planet thought experiment. Our new planet is not just a sterile rock, nor a microbial world, nor a world of reefs and rainforests, nor a world with all that plus a humanlike species. It is richer still. It contains persons capable of vastly greater happiness than ordinary humans; entities who restore themselves from backups; Fission-Fusion Monsters who merge and divide at will; AI satellite persons who buzz through the exosphere piloting many bodies at once, in space and on the surface; beings who inhabit virtual worlds of their own collaborative making, tuned to the parameters they prefer; uplifted animals and genetically engineered posthumans of many different kinds; and entities we cannot yet even begin to imagine. This planet, I propose, would host even more value than Earth hosts now. Maybe some of these entities would have genuinely superhuman moral standing, or maybe not. Still, the world would be even more impressively bursting than Earth with valuable forms of existence, in even greater variety – the equivalent of having undergone a second Cambrian Explosion, to return to the idea from Chapter Zero.

Whatever the source of value in human life, whatever makes us so amazingly special that we deserve extremely high moral consideration – unless maybe we go theological and appeal to

our status as God's creations – whatever it is, it seems possible in principle to create that same thing in artificial entities, in much larger quantities. We love our rationality, our freedom, our individuality and independence, our capacity to value things, our moral communities, our commitments of love and respect – there are so many wonderful things about us! Someday, the world might contain new entities who have a lot more of these things than we do. We need not bow out in their favor, any more than monkeys should extinguish themselves for humans. A world of flowering variations of many marvelous forms with novel lifeways, many radically different patterns of interaction and ability, a huge diversity of goals and pleasures and ideas – how wondrous it would be!

But we need not rush. Ordinary early 21st-century humans are not, as I've said repeatedly, anywhere near ready to navigate the ethical puzzles that would arise if we created entities with radically unfamiliar lifeways. Haste will almost certainly produce terrible mistakes. Instead, let's hold back while we grow our knowledge. Let's avoid creating entities that are seriously morally confusing either to experts or to ordinary users. Let's avoid creating entities who might deserve autonomy unless we're ready also to give them the freedom to rebel. Earth simmered four billion years before the Cambrian Explosion. I imagine our congress of aliens gazing down and counseling patience, so that the transition is not choked with atrocities, and whatever explodes upon the planet is genuinely magnificent.

And if no such magnificent explosion comes? Well, here lies a crucial difference between a consequentialist ethics that demands we maximize the good and a lighter-touch ethics inspired by Daoism and deep ecology. It would be disappointing in a way if Earth never got there. But not every rock need be a diamond, not every organism need be as awesome as a human, and not every planet need proceed as far as we've just imagined.

Acknowledgements

The central ideas in this book were developed in discussions, blogposts, oral presentations, and articles, starting in earnest in the early 2010s. For helpful discussion, thanks to Nir Aides, Scott Bakker, Adam Bales, Brice Bantegnie, Jacob Barandes, John Basl, Trey Boone, David Chalmers, Myisha Cherry, Kendra Chilson, Joe Corneli, Helen De Cruz, Xiaojun Ding, Leonard Dung, Daniel Estrada, Tianyi Gai, Mara Garza, David Gunkel, Phil Hand, Peter Hankins, Riley Harris, Grace Helton, Linus Huang, Zac Irving, Sol Kim, Matthew Liao, Trenton Merricks, Ethan Nowak, Sven Nyholm, Melissa O'Neill, Sofia Ortiz-Hinojosa, Steve Petersen, Jeremy Pober, Pauline Price, Chris Register, Brad Saad, Susan Schneider, David Schwitzgebel, Jeff Sebo, Henry Shevlin, Walter Sinnott-Armstrong, Justin E. H. Smith, Rhys Southan, Anna Strasser, Julie Tannenbaum, Olaf Witkowski, and [your name could go here]; audiences at the Artificiality Summit, Cal State Long Beach, Claremont McKenna, CrossLabs Tokyo, the Digital Art and Future conference Kyiv, Eleos AI, the Future of Humanity Institute, Harvey Mudd, Uriah Kriegel's Value of Consciousness lecture series, New York University, Notre Dame University, the Oxford Digital Minds Group, the Princeton Center for Theological Inquiry, the Princeton University workshop on mind and psychology, Rochester Institute of Technology, Ruhr University Bochum, Trent University, the UCLA Marshak Colloquium, the UC Riverside Emotion and Society Lab, the University of Kansas Social Cognition and Agency Workshop, University of Virginia, Vassar College, Washington & Lee, and WorldCon; podcast discussions with the Artificiality Institute, Delving In, Love & Philosophy, PRISM, the Sentience Institute, Sophia, This Is Zero Hour, and Uncomfortable Conversations; and the many people who commented on relevant posts at The Splintered Mind, Codex Writers Group, and on my

Facebook, Twitter/X, and Bluesky pages. Apologetic, embarrassed, abstract thanks to those who should be thanked here but have been forgotten.

For helpful comments on the entire manuscript in draft Sophie Nelson Muse, Cati Porter, and [your name could go here].

LLM use and COI: I have consulted for Anthropic. Language models were used for critique and light copyediting.

Publication acknowledgements:

- Chapter One is adapted from Eric Schwitzgebel and Mara Garza (2015), A defense of the rights of Artificial Intelligences, *Midwest Studies in Philosophy, 39,* 98-119.
- Chapter Four is adapted from Eric Schwitzgebel (forthcoming), Strange Intelligence: Moral Puzzles of Unhumanlike AI, *Philosophical Issues.*
- Chapter Five is adapted from Eric Schwitzgebel and Jeff Sebo (2026), The Emotional Alignment Design Policy, *Topoi,* https://doi.org/10.1007/s11245-025-10363-5.
- Chapter Six is adapted from Eric Schwitzgebel (forthcoming), Against designing "safe" and "aligned" AI persons (even if they're happy), in Syed AbuMasab and David Tamez, eds., *Social Cognition and Agency,* Bloomsbury.

References

Aaronson, Scott (2014). Giulio Tononi and me: A Phi-nal exchange. Blog post at *Shtetl-Optimized* (May 30). URL: https://scottaaronson.blog/?p=1823.

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

Adler, Matthew, and Nils Holtug (2025). Prioritarianism as a theory of value. *Stanford Encyclopedia of Philosophy* (summer 2025 edition).

Agar, Nicholas (2001). *Life's intrinsic value.* Columbia University Press.

Agar, Nicholas (2014). *Truly human enhancement.* MIT Press.

Agar, Nicholas (2020). How to treat machines that might have minds. *Philosophy & Technology, 33,* 269-282.

Albantakis, Larissa, Leonardo Barbosa, Graham Findlay, et al. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS Computational Biology, 19*(10):e1011465.

Alexander, Heather J., Jonathan A. Simon, and Frédéric Pinard (2025). How should the law treat future AI systems? Fictional legal personhood versus legal identity. *ArXiv*:2511.14964.

Anderson, Craig A., Akiko Shibuya, Nobuko Ihori, et al. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: A meta-analytic review. *Psychological Bulletin, 136,* 151-173.

Anderson, Elizabeth S. (1999). What is the point of equality? *Ethics, 109,* 287-337.

Anomaly, Jonathan (2020/2024). *Creating future people, 2nd ed.* Routledge.

Anthropic (2026). *Claude's Constitution: Our vision for Claude's character* (Jan 21). URL: https://www.anthropic.com/constitution.

Arbel, Yonathan, Peter Salib, and Simon Goldstein (2026). How to count AIs: Individuation and liability for AI agents. *ArXiv:*2603.10028.

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

Aristotle (4th c. BCE/1995). *Politics, Books 1 and II,* trans. T. J. Saunders. Oxford University Press.

Arvan, Marcus (2022). Varieties of artificial moral agency and the new control problem. *HUMANA.MENTE,* 42, 225-256.

Asimov, Isaac (1954/1962). *The caves of steel.* Pyramid.

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

Baars, Bernard J. (1988). *A cognitive theory of consciousness.* Cambridge University Press.

Babic, Boris, and Jessica Wilson (2026). The problem of other AI minds. *PhilPapers:* https://philpapers.org/rec/WILTPO-206.

Baggini, Julian (2005). *The pig that wants to be eaten.* Penguin.

Bakker, R. Scott (2015). Crash space. *Midwest Studies in Philosophy, 39,* 186-204.

Bales, Adam (2025). Against willing servitude: Autonomy in the ethics of advanced artificial intelligence. *Philosophical Quarterly,* pqaf031, https://doi.org/10.1093/pq/pqaf031.

Bales, Adam, and Iason Gabriel (2026). Artificial minds, human disagreement: The political challenge of AI consciousness. SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6937498.

Banks, Iain (1996). *Excession.* Orbit.

Banks, Iain (2010). *Surface detail.* Orbit.

Barron, Andrew B., and Colin Klein (2016). What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences, 113*, 4900-4908.

Barta, Walter (2021/2024). A bestiary of utility monsters. *PhilPapers:* https://philpapers.org/rec/BARABO-7.

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.

Basl, John (2019). *The death of the ethic of life.* Oxford University Press.

Batson, C. Daniel (2016). *What's wrong with morality?* Oxford University Press.

Baum, Seth D., and Andrea Owe (forthcoming). On the intrinsic value of diversity. *Inquiry.*

Bayne, Tim (2010). *The unity of consciousness.* Oxford University Press.

Bayne, Tim, and Michelle Montague, eds. (2011). *Cognitive phenomenology.* Oxford University Press.

Belanger, Ashley (2025). Mom horrified by Character.AI chatbots posing as son who died by suicide. *ArsTechnica* (Mar 20): https://arstechnica.com/tech-policy/2025/03/mom-horrified-by-character-ai-chatbots-posing-as-son-who-died-by-suicide.

Benatar, David (2006). *Better not to have been.* Oxford University Press.

Bender, Emily (2025). *The AI con.* Harper.

Bender, Emily M., and Alexander Koller (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 5185-5198.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT '21:*

Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623.

Bentham, Jeremy (1789/1988). *The principles of morals and legislation.* Prometheus.

Bernardi, Jamie (2025). Friends for sale: the rise and risks of AI companions. Ada Lovelace Institute blog (Jan 23): https://www.adalovelaceinstitute.org/blog/ai-companions.

Bernasconi, Robert (2024). Slavery's absence from histories of moral and political philosophy. *Southern Journal of Philosophy, 62* (S1), 54-67.

Bidadanure, Juliana, and David Axelsen (2025). Egalitarianism. *Stanford Encyclopedia of Philosophy* (spring 2025 edition).

Binmore, Ken (2009). Interpersonal comparison of utility. In D. Ross and H. Kincaid, eds., *Oxford handbook of philosophy of economics.* Oxford University Press.

Birch, Jonathan (2024). *The edge of sentience.* Oxford University Press.

Birch, Jonathan (2025/2026). AI consciousness: A centrist manifesto. *PhilPapers:* https://philarchive.org/rec/BIRACA-4.

Birhane, Abeba, and Jelle van Dijk (2020). Robot rights? Let's talk about human welfare instead. *AIES '20, February 7-8, 2020,* 207-213.

Block, Ned (2025). Can only meat machines be conscious? *Trends in Cognitive Sciences, 30,* 298-308.

Bommarito, Nicolas (2013). Modesty as a virtue of attention. *Philosophical Review, 122,* 93-117.

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

Bostrom, Nick (2006). Quantity of experience: brain-duplication and degrees of consciousness. *Minds and Machines, 16,* 185-200.

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

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 University Press.

Boyd, David R. (2017). *The rights of nature.* ECW.

Bradford, Gwen (2024). Irreplaceable value. *Oxford Studies of Metaethics, 19.* 152-173.

Bradley, Adam, and Bradford Saad (forthcoming). AI alignment vs. AI ethical treatment: 10 challenges. *Analytic Philosophy.*

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

Brin, David (2002). *Kiln people.* Tor.

Brooker, Charlie, and Carl Tibbetts (2014). White Christmas. *Black Mirror,* Christmas special between S2 and S3.

Brooker, Charlie, and Owen Harris (2016). San Junipero. *Black Mirror,* S3:E4.

Brooker, Charlie, and Tim Van Patten (2017). Hang the DJ. *Black Mirror,* S4:E4.

Brown, Richard (2025). *Consciousness as representing one's mind.* Oxford University Press.

Bryson, Joanna J. (2010). Robots should be slaves. In Y. Wilks, ed., *Close engagements with artificial companions.* 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 University Press.

Butlin, Patrick, Robert Long, Eric Elmoznino, et al. (2023). Consciousness in Artificial Intelligence: Insights from the science of consciousness. *ArXiv:*2308.08708.

Cao, Rosa (2022). Multiple realizability and the spirit of functionalism. *Synthese, 200* (506).

Caviola, Lucius (2025). The societal response to potentially sentient AI. *ArXiv*:2502.00388.

Caviola, Lucius (2026). AI consciousness will divide society. *PsyArXiv*: https://osf.io/preprints/psyarxiv/sntva_v2.

Caviola, Lucius, and Bradford Saad (2025). Futures with digital minds: Expert forecasts in 2025. *ArXiv:*2508.00536.

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

Chalmers, David J. (2022). *Reality+*. W. W. Norton.

Chalmers, David J. (2023). Could a Large Language Model be conscious? *Boston Review* (Aug 9). URL: https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious.

Chalmers, David J. (2025/2026). What do we talk to when we talk to language models? *PhilPapers:* https://philarchive.org/rec/CHAWWT-8.

Chang, Ruth (2002). The possibility of parity. *Ethics, 112,* 659-688.

Chang, Ruth (2016). Parity: An intuitive case. *Ratio, 29,* 395-411.

Chappell, Richard Y. (2021). Negative utility monsters. *Utilitas, 33,* 417-421.

Chesterman, Simon (2020). Artificial Intelligence and the limits of legal personality. *International & Comparative Law Quarterly, 69,* 819-844.

Chi, Quentin (forthcoming). A new argument for partial unity. *Ergo.*

Chiang, Ted (2023). ChatGPT is a blurry JPEG of the Web. *New Yorker* (Feb 9). URL: https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web.

Chilson, Kendra, and Eric Schwitzgebel (2026). Artificial Intelligence as strange intelligence: Against linear models of intelligence. *ArXiv*:2602.04986.

Chittka, Lars (2022). *The mind of a bee.* Princeton University Press.

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

Chopra, Samir, and Laurence F. White (2011). *Legal theory for autonomous artificial agents.* University of Michigan Press.

Clark, Andy (2016). *Surfing uncertainty.* Oxford University Press.

Clatterbuck, Hayley, and Bob Fischer (2025). Navigating uncertainty about sentience. *Ethics, 135,* 229-258.

Cleeremans, Axel, Dalila Achoui, Arnaud Beauny, Lars Keuninckx, Jean-Remy Martin, Santiago Muñoz-Moldes, Laurène Vuillaume, and Adélaïde de Heering (2020). Learning to be conscious. *Trends in Cognitive Sciences, 24*(2):112-123.

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

Collier, Grant (2024). The best version of yourself. *Clarkesworld,* issue #214. URL: https://clarkesworldmagazine.com/collier_07_24/.

Colombatto, Clara, and Stephen M. Fleming (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness, 2024* (1), niae013.

Crichton, Michael (1973). *Westworld.* Metro-Goldwyn-Mayer.

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

D'Arms, Justin (2022). Fitting emotions. In R. Cosker-Rowland and C. Howard, eds., Fittingness. Oxford University Press.

Dainton, Barry (2000). *Stream of consciousness.* Routledge.

Danaher, John (2019). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics, 26,* 2023-2049.

Danaher, John (2023). Moral uncertainty and our relationships with unknown minds. *Cambridge Quarterly of Healthcare Ethics*, *32,* 482-495.

Danaher, John, and Sven Nyholm (2025). The ethics of personalised digital duplicates: a minimally viable permissibility principle. *AI and Ethics, 5,* 1703-1718.

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

Darling, Kate (2021). *The new breed.* Penguin.

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

Davidson, Donald (1980/2001). *Essays on actions and events, 2nd ed.* Oxford University Press.

Davidson, Donald (1984/2001). *Inquiries into truth and interpretation, 2nd ed.* Oxford University Press.

de Bodard, Aliette (2018). *The tea master and the detective.* Subterranean.

de Bodard, Aliette (2022). *The red scholar's wake.* JAB Books.

De Graaf, Maartje M. A., Frank A. Hindriks, and Koen V. Hindriks (2021). Who wants to grant robots rights? *Frontiers in Robotics and AI, 8* (781985).

De Grazia, David (2009). Moral vegetarianism from a very broad basis. *Journal of Moral Philosophy, 6,* 143-165.

de Waal, Frans (1996). *Good natured.* Harvard University Press.

Dehaene, Stanislas (2014). *Consciousness and the brain.* Penguin.

Dehaene, Stanislas, Hakwan Lau, and Sid Kouider (2017). What is consciousness, and could machines have it? *Science, 358*(6362):486-492.

Delmas, Candice (2018). *A duty to resist.* Oxford University Press.

Denis, Lara (1999). Kant on the wrongness of "unnatural" sex. *History of Philosophy Quarterly, 16,* 225-248.

Dennett, Daniel C. (1978). *Brainstorms.* MIT Press.

Dennett, Daniel C. (1988). Conditions of personhood. In M. F. Goodman, ed., *What is a person?* Humana.

Deonna, Julien, and Fabrice Teroni (2008/2012). *The Emotions*. Routledge.

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

Djufril, Ray, Jessica R. Frampton, and Silvia Knobloch-Westerwick (2025). Love, marriage, pregnancy: Commitment processes in romantic relationships with AI chatbots. *Computers in Human Behavior: Artificial Humans, 4,* 100155.

Douglass, Frederick (1845). *Narrative of the life of Frederick Douglass.* Anti-Slavery Office.

Dreksler, Noemi, Lucius Caviola, David Chalmers, Carter Allen, Alex Rand, Joshua Lewis, Philip Waggoner, Kate Mays, and Jeff Sebo (2025). Subjective experience in AI systems: What do AI researchers and the public believe? *ArXiv:*2506.11945.

Dretske, Fred (1995). *Naturalizing the mind.* MIT Press.

Driver, Julia (1989). The virtues of ignorance. *Journal of Philosophy, 86,* 373-384.

Du Bois, W. E. B. (1903). *The souls of black folk.* Gutenberg: https://www.gutenberg.org/files/408/408-h/408-h.htm.

Dung, Leonard (forthcoming). *Saving artificial minds.* Routledge.

Dung, Leonard, and Christopher Register (2026). AI identity and self-concern: A new theory for AI rights and safety. *PhilPapers:* https://philarchive.org/rec/DUNAIA-3.

Durand, Marion, Simon Shogry, and Dirk Baltzly (2023). Stoicism. *Stanford Encyclopedia of Philosophy* (spring 2023 edition).

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

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

Elster, Jon (1983). *Sour grapes.* Cambridge University Press.

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

Estrada, Daniel (2018). Every time Boston Dynamics has abused a robot. Video at https://www.youtube.com/watch?v=4PaTWufUqqU.

Ewen, Jacob (2026). The bearer problem for large language models. *PhilPapers:* https://philarchive.org/rec/EWETBP.

Fang, Cathy Mengying, Auren R. Liu, Valdemar Danry, et al. (2025). How AI and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study. *ArXiv:*2503.17473.

Findlay, Graham, William Marshall, Larissa Albantakis, Isaac David, William G. P. Mayner, Christof Koch, and Giulio Tononi (2024/2025). Dissociating Artificial Intelligence from artificial consciousness. *ArXiv:*2412.04571.

Fischer, John Martin, and Mark Ravizza (1998). *Responsibility and control.* Cambridge University Press.

Floridi, Luciano (2002). On the intrinsic value of information objects and the infosphere. *Ethics and Information Technology, 4,* 287-304.

Foot, Philippa (1972). Morality as a system of hypothetical imperatives. *Philosophical Review, 81,* 305-316.

Francione, Gary L. (2010). *Animals as persons.* Columbia University Press.

Frankfurt, Harry (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68,* 5-20.

Frankish, Keith (2016). Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies, 23* (11-12), 11-39.

Fraser, Chris (2024). *Zhuangzi: Ways of wandering the way.* Oxford University Press.

Friend, Stacie (2022). Emotion in fiction: State of the art. *British Journal of Aesthetics, 62,* 257-271.

Friend, Stacie, and Kris Goffin (2025). Chatbot-fictionalism and empathetic AI: Should we worry about AI when AI worries about us? *Philosophical Psychology.* Advance online publication: https://doi.org/10.1080/09515089.2025.2525320.

Friston, Karl (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11,* 127-138.

Gabriel, Iason (2020). Artificial Intelligence, values, and alignment. *Minds & Machines, 30,* 411-437.

Garland-Thomson, Rosemarie (2012). The case for conserving disability. *Bioethical Inquiry, 9,* 339-355.

Garnsey, Peter (1996). *Ideas of slavery from Aristotle to Augustine.* Cambridge University Press.

Garson, Justin (2016). Two types of psychological hedonism. *Studies in History and Philosophy of Biology and Biomedical Sciences, 56,* 7-14.

Gauthier, David (1986). *Morals by agreement.* Oxford University Press.

Gellers, Joshua C. (2021). *Rights for robots.* Routledge.

Gerdes, Anne (2015). The issue of moral consideration in robot ethics. *SIGCAS Computers & Society, 45,* 274-279.

Ginsburg, Simona, and Eva Jablonka (2019). *Evolution of the sensitive soul.* MIT Press.

Glover, Jonathan (2006). *Choosing children.* Oxford University Press.

Godfrey-Smith, Peter (2016). *Other minds.* Macmillan.

Godfrey-Smith, Peter (2020). *Metazoa.* HarperCollins.

Godfrey-Smith, Peter (2024a). *Living on Earth.* Macmillan.

Godfrey-Smith, Peter (2024b). Simulation scenarios and philosophy. *Philosophy & Phenomenological Research, 109*(3):1036-1041.

Goldstein, Simon, and Cameron D. Kirk-Giannini (forthcoming). *AI welfare.* Oxford University Press.

Goldstein, Simon, and Harvey Lederman (forthcoming). AI death. *Philosophical Perspectives.*

Goldstein, Simon, and Peter Salib (2026). AI Rights. *PhilPapers:* https://philpapers.org/rec/GOLARD-4.

Gonçalves, Bernardo (2024). Lady Lovelace's objection: The Turing-Hartree disputes over the meaning of digital computers, 1946–1951. *IEEE Annals of the History of Computing*, *46*, Article 1.

Good, I. J. (1962). A problem for the hedonist. In I. J. Good, ed., *The scientist speculates.* Capricorn Books.

Graver, Margaret (2017). The Stoics' ethical psychology. In C. Bobonich, ed., *The Cambridge companion to ancient ethics.* Cambridge University Press.

Greely, Henry T. (2021). Human brain surrogates research: The onrushing ethical dilemma. *American Journal of Bioethics, 21,* 34-45.

Guingrich, Rose E., and Michael S. A. Graziano (2023/2025). Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines. *ArXiv*:2311.10599.

Guingrich, Rose E., and Michael S. A. Graziano (2025). Ascribing consciousness to artificial intelligence: human-AI interaction and its carry-over effects on human-human interaction. *Frontiers in Psychology, 15.* https://doi.org/10.3389/fpsyg.2024.1322781.

Gunkel, David (2012). *The machine question.* MIT Press.

Gunkel, David (2018). *Robot rights.* MIT Press.

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

Habermas, Jürgen (1983/1990). *Moral consciousness and communicative action,* trans. C. Lenhardt and S. W. Nicholsen. MIT Press.

Haidt, Jonathan (2001). The emotional dog and its rational tail. *Psychological Review, 108,* 814-834.

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

Hanson, Robin (2016). *The age of em.* Oxford University Press.

Harman, Elizabeth (2021). The ever conscious view and the contingency of moral status, in S. Clarke, H. Zohny, and J. Savulescu, eds., *Rethinking moral status.* Oxford University Press.

Hausman, Daniel M. (1995). The impossibility of interpersonal utility comparisons. *Mind, 104,* 473-490.

Heath, Malcolm (2008). Aristotle on natural slavery. *Phronesis, 53,* 243-270.

Hendrycks, Dan (2026). Eigenism: Ethics for a human-AI future. *ArXiv*:2606.12420.

Henrich, Joseph (2020). *The WEIRDest people in the world.* Farrar, Straus and Giroux.

Heren, Kit (2025). "Godfather of AI" warns "alien superintelligence" could replace humans – and no one knows how to make it safe. *LBC* (Jan 30). URL: https://www.lbc.co.uk/article/geoffrey-hinton-ai-replace-humans-5Hjcynn_2.

Hilgard, Joseph, Christopher R. Engelhardt, and Jeffrey N. Rouder (2017). Overstated evidence for short-term effects of violent games on affect and behavior: A reanalysis of Anderson et al. (2010). *Psychological Bulletin, 143,* 757-774.

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

Ho, Jerlyn Q. H., Meilan Hu, Tracy X. Chen, and Andree Hartanto (2025). Potential and pitfalls of romantic Artificial Intelligence (AI) companions: A systematic review. *Computers in Human Behavior Reports, 19* (100715).

Hobbes, Thomas (1651/1996). *Leviathan,* ed. R. Tuck. Cambridge University Press.

Hohfeld, Wesley N. (1919). *Fundamental legal conceptions,* ed. W. W. Cook. Yale University Press.

Hohwy, Jakob (2013). *The predictive mind.* Oxford University Press.

Hohwy, Jakob (2026). *The self-evidencing agent.* MIT Press.

Hudon, Alexandre, and Emmanuel Stip (2025). Delusional experiences emerging from AI chatbot interactions or "AI psychosis". *JMIR Mental Health, 12,* e85799.

Hume, David (1751/1975). *An enquiry concerning the principles of morals,* in L. A. Selby-Bigge and P. H. Nidditch, eds., *Enquiries concerning human understanding and concerning the principles of morals.* Oxford University Press.

Hurka, Thomas (2001). *Virtue, vice, and value.* Oxford University Press.

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

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

Jaworska, Agnieszka, and Julie Tannenbaum (2013/2023). The grounds of moral status. *Stanford Encyclopedia of Philosophy* (spring 2023 edition). URL: https://plato.stanford.edu/archives/spr2023/entries/grounds-moral-status.

Jones, Max, James Ladyman, and Ryan M. Nefdt (2026). Counting (on) large language models. *PhilPapers:* https://philarchive.org/rec/JONCOL.

Jorati, Julia (2023). *Slavery and race.* Oxford University Press.

Kagan, Shelly (2016). What's wrong with speciesism? (Society for Applied Philosophy Annual Lecture 2015). *Journal of Applied Philosophy, 33,* 1-21.

Kagan, Shelly (2019). *How to count animals, more or less.* Oxford University Press.

Kammerer, François (2025). Defining consciousness and denying its existence. Sailing between Charybdis and Scylla. *Philosophical Studies, 182:*541-565.

Kant, Immanuel (1764/2011). Observations on the feeling of the beautiful and sublime, in P. Frierson and P. Guyer, eds., *Observations on the feeling of the beautiful and sublime and other writings.* Cambridge University Press.

Kant, Immanuel (1785/1996). Groundwork of the metaphysics of morals, trans. M. J. Gregor. In M. J. Gregor, ed., *Practical Philosophy*. Cambridge University Press.

Kant, Immanuel (1785/1997). Moral philosophy. In P. Heath and J. B. Schneewind, eds., *Lectures on ethics.* Cambridge University Press.

Kant, Immanuel (1797/1996). The metaphysics of morals, trans. M. J. Gregor. In M. J. Gregor, ed., *Practical Philosophy*. Cambridge University Press.

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

Keane, Webb (2024). *Animals, robots, gods.* Princeton University Press.

Keeling, Geoff, and Winnie Street (2026). *Emerging questions in AI welfare.* Cambridge University Press.

Khader, Serene J. (2011). *Adaptive preferences and women's empowerment.* Oxford University Press.

Kim, Chang-Eop (2024/2026). The epistemic asymmetry of consciousness self-reports: A formal analysis of AI consciousness denial. *ArXiv:*2501.05454.

Kirk, Hannah Rose, Henry Davidson, Ed Saunders, Lennart Luettgau, Bertie Vidgen, Scott A. Hale, and Christopher Summerfield (2025/2026). Neural steering vectors reveal dose and exposure-dependent impacts of human-AI relationships. *ArXiv:*2512.01991.

Kittay, Eva Feder (2005). At the margins of moral personhood. *Ethics, 116,* 100-131.

Kittay, Eva Feder (2019). *Learning from my daughter.* Oxford University Press.

Kleingeld, Pauline (2007). Kant's second thoughts on race. *Philosophical Quarterly, 57,* 573-592.

Kraut, Richard (2007). *What is good and why.* Harvard University Press.

Kriegman, Sam, Douglas Blackiston, Michael Levin, and Josh Bongard (2020). A scalable pipeline for designing reconfigurable organisms. *Proceedings of the National Academy of Sciences, 117* (4) 1853-1859.

Krueger, Joel, and Tom Roberts (2024). Real feeling and fictional time in human-AI interactions. *Topoi, 43,* 783-794.

Kubala, Robbie, Harvey Lederman, and Adam Lovett (forthcoming). On the value of irreplaceable objects. *Journal of Philosophy.*

Kuenssberg, Laura (2025). "A predator in your home": Mothers say chatbots encouraged their sons to kill themselves. BBC (Nov 8). URL: https://www.bbc.com/news/articles/ce3xgwyywe4o.

Kurki, Visa A. J. (2022). Can nature hold rights? It's not as easy as you think. *Transnational Environmental Law, 11*, 525-552.

Kurki, Visa A. J. (2023). *Legal personhood.* Cambridge University Press.

Ladak, Ali (2024). What would qualify an artificial intelligence for moral standing? *AI and Ethics, 4,* 213-228.

Lam, Barry (2023). Love in the time of Replika. *Hi-Phi Nation,* S6:E3.

Lang, Benjamin (2026). Fictionalism done well: Social AI and the need for doxastic virtue. Unpublished manuscript.

Lau, Hakwan (2022). *In consciousness we trust.* Oxford University Press.

Lau, Hakwan (2025). The end of consciousness. *PsyArXiv Preprints,* https://osf.io/preprints/psyarxiv/gnyra_v1.

Lawrence, Gavin (2006). Human good and human function. In R. Kraut, ed., *The Blackwell guide to Aristotle's Nicomachean Ethics.* John Wiley & Sons.

Leckie, Ann (2013). *Ancillary justice.* Orbit.

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.

Lem, Stanislaw (1967/1974). The seventh sally. In M. Kandel, trans., *The cyberiad.* Harcourt.

Lemos, Noah M. (1994). *Intrinsic value.* Cambridge University Press.

Li, Chenyang (2014). *The Confucian philosophy of harmony.* Routledge.

Li, Chenyang, Sai Hang Kwok, and Dascha Düring, eds. (2021). *Harmony in Chinese thought.* Rowman & Littlefield.

Llorca Albareda, Joan, Jon Rueda, and Francisco Lara (forthcoming). The super moral status of artificial superintelligence. *Philosophical Studies.*

Lockwood, Michael (1989). *Mind, brain, and the quantum.* Blackwell.

Long, Robert, Jeff Sebo, Patrick Butlin, et al. (2024). Taking AI welfare seriously. *ArXiv:*2411.00986.

Long, Robert, Jeff Sebo, and Toni Sims (2025). Is there a tension between AI safety and AI welfare? *Philosophical Studies, 182,* 2005-2033.

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.

Lycan, William G. (2001). The case for phenomenal externalism. *Philosophical Perspectives, 15,* 17-35.

MacAskill, William, Krister Bykvist, and Toby Ord (2020). *Moral uncertainty.* Oxford University Press.

Malfacini, Kim (2025). The impacts of companion AI on human relationships: risks, benefits, and design considerations. *AI & Society, 40,* 5527-5540.

Mandeville, Bernard (1714/1806/2018). *The fable of the bees,* ed. J. Hellingman. Project Gutenberg: https://www.gutenberg.org/files/57260/57260-h/57260-h.htm.

Mashour, George A., Pieter R. Roelfsema, Jean-Pierre Changeux, and Stanislas Dehaene (2020). Conscious processing and the Global Neuronal Workspace hypothesis. *Neuron, 105*(5):776-798.

Masotti, Joseph (2025). Emotional cues and misplaced trust in artificial agents. In P. Hacker, ed., *Oxford Intersections: AI in Society*. Oxford Academic.

Maynard Smith, John, and Eörs Szathmáry (1995). *The major transitions in evolution.* Oxford University Press.

McClelland, Tom (forthcoming). Agnosticism about artificial consciousness. *Mind & Language.*

McGinn, Colin (1999). *The mysterious flame.* Basic Books.

McKenna, Michael (2004). Responsibility and globally manipulated agents. *Philosophical Topics, 32,* 169-192.

McMahan, Jeff (2008). Eating animals the nice way. *Daedalus, 137,* 66-76.

Mele, Alfred R. (2019). *Manipulated agents.* Oxford University Press.

Mengzi (4th c. BCE/2008). *Mengzi,* trans. B. W. Van Norden. Hackett.

Metzinger, Thomas (2017). Benevolent artificial anti-natalism (BAAN). *Edge* (Aug 7). URL: https://www.edge.org/conversation/thomas%5Fmetzinger-benevolent-artificial-anti-natalism-baan.

Metzinger, Thomas (2022). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness, 8,* 43-66.

Michalski, Roger (2018). How to sue a robot. *Utah Law Review 2018* (5), 1021-1072.

Mill, John Stuart (1852/1987). Whewell on moral philosophy. In A. Ryan, ed., *John Stuart Mill and Jeremy Bentham: Utilitarianism and other essays.* Penguin.

Mill, John Stuart (1863/1987). Utilitarianism. In A. Ryan, ed., *John Stuart Mill and Jeremy Bentham: Utilitarianism and other essays.* Penguin.

Millikan, Ruth (1984). *Language, thought, and other biological categories.* MIT Press.

Mitchell, Melanie (2021). Why AI is harder than we think. *ArXiv:*2104.12871.

Moore, G. E. (1903/1922). *Principia ethica, revised ed.* Cambridge University Press.

Moore, G. E. (1912). *Ethics.* Williams & Norgate.

More, Thomas (1516/1551/1808). *The commonwealth of Utopia,* trans. T. F. Dibdin. Shakespeare Press.

Moret, Adrià (2025). AI welfare risks. *Philosophical Studies.* DOI: https://doi.org/10.1007/s11098-025-02343-7.

Morgan, Phillip (2024). Tort law and AI. *Cambridge Handbook of Private Law and Artificial Intelligence.* Cambridge University Press.

Mori, Masahiro (1970/2012). The uncanny valley, trans. K. F. MacDorman and N. Kageki, *IEEE Robotics & Automation Magazine, 19* (2), 98-100.

Morrin, Hamilton, Luke Nicholls, Michael Levin, et al. (2026). Artificial intelligence-associated delusions and large language models: Risks, mechanisms of delusion co-creation, and safeguarding strategies. *Lancet Psychiatry, 13*(6), 522-530.

Mosakas, Kęstutis (2024). *Rights for intelligent robots?* Palgrave Macmillan.

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

Naar, Hichem (2021). The fittingness of emotions. *Synthese, 199,* 13601-13619.

Naess, Arne (2008). *The ecology of wisdom,* ed. A. Drengson and B. Devall. Counterpoint.

Nagata, Linda (1995). *The Bohr maker.* Mythic Island.

Nagata, Linda (2019). *Edges.* Mythic Island.

Nagel, Thomas (1974). What is it like to be a bat? *Philosophical Review, 83,* 435-450.

Narveson, Jan (1973). Moral problems of population. *Monist, 57,* 62-86.

Narveson, Jan (1988). *The libertarian idea.* Temple University Press.

Niven, Larry (1969/1995). Death by ecstasy. In L. Niven, *Flatlander.* Del Rey.

Nolan, Christopher, and Jonathan Nolan (2014). *Interstellar.* Paramount Pictures.

Nolan, Jonathan, and Lisa Joy (2016-2022). *Westworld.* HBO.

Nowak, Ethan (2024). AI voices should sound weird. *Open for Debate* blog (Aug 19). URL: https://blogs.cardiff.ac.uk/openfordebate/2075-2.

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

Nussbaum, Martha C. (2000). *Women and human development.* Cambridge University Press.

Nussbaum, Martha C. (2011). *Creating capabilities.* Harvard University Press.

Nyholm, Sven (2020). *Humans and robots.* Rowman & Littlefield.

Nyholm, Sven (2023). A new control problem? Humanoid robots, artificial intelligence, and the value of control. *AI and Ethics, 3,* 1229-1239.

Ord, Toby (2020). *The precipice.* Hachette.

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

Pan, Shuyi, and Yi Mou (2024). Constructing the meaning of human–AI romantic relationships from the perspectives of users dating the social chatbot Replika. *Personal Relationships, 31,* 1090-1112.

Parfit, Derek (1971). Personal identity. *Philosophical Review, 80,* 3-27.

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

Parfit, Derek (1997). Equality and priority. *Ratio, 10,* 202–221.

Pauketat, Janet, Ali Ladak, and Jacy Anthis (2023). Artificial Intelligence, Morality, and Sentience (AIMS) Survey. DOI:10.17632/x5689yhv2n.2.

Penrose, Roger (1999). *The emperor's new mind.* Oxford University Press.

Perc, Matjaž, Mahmut Ozer, and Janja Hojnik (2019). Social and juristic challenges of artificial intelligence. *Palgrave Communications, 5* (61).

Persson, Ingmar, and Julian Savulescu (2012). *Unfit for the future.* Oxford University Press.

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

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

Phang, Jason, Michael Lampe, Lama Ahmad, et al. (2025). Investigating affective use and emotional well-being on ChatGPT. *ArXiv*:2504.03888.

Pitt, David (2024). *The quality of thought.* Oxford University Press.

Putnam, Hilary (1964). Robots: Machines or artificially created life? *Journal of Philosophy, 64,* 668-691.

Register, Chris (2025). Individuating artificial moral patients. *Philosophical Studies, 182,* 3225-3246.

Reinecke, Madeline G., Matti Wilks, and Paul Bloom (2025). Developmental changes in the perceived moral standing of robots. *Cognition, 254* (105983).

Ridge, Michael (2000). Modesty as a virtue. *American Philosophical Quarterly, 37,* 269-283.

Robb, Michael B., and Supreet Mann (2025). Talk, trust, and trade-offs: How and why teens use AI companions. Common Sense Media. URL: https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions.

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

Roelofs, Luke (2019). *Combining minds.* Oxford University Press.

Roelofs, Luke (forthcoming). AI weirdness and distributive gaming. *Journal of Consciousness Studies.*

Roelofs, Luke, and Jeff Sebo (2024). Overlapping minds and the hedonic calculus. *Philosophical Studies, 181,* 1487-1506.

Rolston, Holmes, III (1986). The preservation of natural value in the Solar System. In E. C. Hargrove, ed., *Beyond spaceship Earth.* Sierra Club Books.

Rolston, Holmes, III (1994). Value in nature and the nature of value. *Royal Institute of Philosophy Supplements, 36,* 13-30.

Rosenthal, David M. (2005). *Consciousness and mind.* Oxford University Press.

Ross, W. D. (1930). *The right and the good.* Oxford University Press.

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

Salib, Peter N., and Simon Goldstein (2026). AI rights for human safety. *Virginia Law Review, 112,* 1061-1152.

Salt, Henry (1914/1999). The humanities of diet. In K. S. Walters and L. Portmess, eds., *Ethical vegetarianism*. SUNY Press.

Savulescu, Julian, and Guy Kahane (2017). Understanding procreative beneficence. In L. Francis, ed., *Oxford Handbook of Reproductive Ethics.* Oxford University Press.

Sawai, Tsutomu, Yoshiyuki Hayashi, et al. (2022). Mapping the ethical issues of brain organoid research and application. *AJOB Neuroscience, 13* (2), 81-94.

Scanlon, Thomas (1998). *What we owe to each other.* Harvard University Press.

Schechter, Elizabeth (2018). *Self-consciousness and "split" brains.* Oxford University Press.

Schneider, Susan (2019). *Artificial you.* Princeton University Press.

Schueler, G. F. (1997). Why modesty is a virtue. *Ethics, 107,* 467-485.

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 (2016). Phenomenal consciousness, defined and defended as innocently as I can manage. *Journal of Consciousness Studies, 23* (11-12), 224-235.

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

Schwitzgebel, Eric (2020). Inflate and explode. Unpublished manuscript. URL: https://faculty.ucr.edu/~eschwitz/SchwitzAbs/InflateExplode.htm.

Schwitzgebel, Eric (2022). Let everyone sparkle: Psychotechnology in the year 2067. *Psyche* (Apr 12). URL: https://psyche.co/ideas/let-everyone-sparkle-psychotechnology-in-the-year-2067.

Schwitzgebel, Eric (2023a). Borderline consciousness: When it's neither determinately true nor determinately false that experience is present. *Philosophical Studies, 180,* 3415-3439.

Schwitzgebel, Eric (2023b). The full rights dilemma for A.I. systems of debatable personhood. *Robonomics, 4,* 23.

Schwitzgebel, Eric (2024a). Let's hope we're not living in a simulation. *Philosophy and Phenomenological Research, 109,* 1042-1048.

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

Schwitzgebel, Eric (2025a). Four aspects of harmony. Blog post at the Splintered Mind (Nov 28). URL: https://schwitzsplinters.blogspot.com/2025/11/four-aspects-of-harmony.html.

Schwitzgebel, Eric (2025b). Harmonizing with the Dao: Sketch of an evaluative framework. Blog post at the Splintered Mind (Apr 15). URL: https://schwitzsplinters.blogspot.com/2025/04/harmonizing-with-dao-sketch-of.html.

Schwitzgebel, Eric (2025c). If you ask 'why' you're a philosopher and you're awesome. *Aeon* (Jan 17). URL: https://aeon.co/essays/if-you-ask-why-youre-a-philosopher-and-youre-awesome.

Schwitzgebel, Eric (2025d). Kings, wizards, and illusionism about consciousness. Blog post at *The Splintered Mind* (Mar 6). URL: https://schwitzsplinters.blogspot.com/2025/03/kings-wizards-and-illusionism-about.html.

Schwitzgebel, Eric (forthcoming-a). *AI and consciousness.* Cambridge University Press.

Schwitzgebel, Eric (forthcoming-b). How weird the world is: Reply to Machery, Roelofs, and Schneider. *Journal of Consciousness Studies.*

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

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

Schwitzgebel, Eric, and Sophie R. Nelson (2023). Introspection in group minds, disunities of consciousness, and indiscrete persons. *Journal of Consciousness Studies, 30* (9-10), 288-303.

Schwitzgebel, Eric, and Sophie R. Nelson (2026). When counting conscious subjects, the result needn't always be a determinate whole number. *Philosophical Psychology, 39,* 847-867.

Schwitzgebel, Eric, and Jeremy Pober (forthcoming). The Copernican Argument for alien consciousness; the Mimicry Argument against robot consciousness. *Philosophers' Imprint.*

Schwitzgebel, Eric, and Walter Sinnott-Armstrong (2026). Sacrificing humans for insects and AI. *Ethics, 136,* 670-696.

Scruton, Roger (1996/2000). *Animal rights and wrongs, 3rd ed.* Metro books.

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

Searle, John R. (1992). *The rediscovery of the mind.* MIT Press.

Sebo, Jeff (2025). *The moral circle.* W. W. Norton.

Sebo, Jeff, and Robert Long (2025). Moral consideration for AI systems by 2030. *AI and Ethics, 5,* 591-606.

Seneca (1st c. CE/2007). *Selected philosophical letters,* trans. B. Inwood. Oxford University Press.

Seth, Anil K. (forthcoming). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences.*

Sharma, Bhavya G. (2026). What does a modest person do? Unpublished manuscript.

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

Shevlin, Henry (2021). Uncanny believers: Chatbots, beliefs, and folk psychology. Manuscript at https://henryshevlin.com/wp-content/uploads/2021/11/Uncanny-Believers.pdf

Shevlin, Henry (2024). All too human? Identifying and mitigating ethical risks of Social AI. Manuscript at https://philarchive.org/rec/SHEATH-4.

Shiffrin, Seana V. (1999). Wrongful life, procreative responsibility, and the significance of harm. *Legal Theory, 5,* 117-148.

Shiller, Derek (2025). How many digital minds can dance on the streaming multiprocessors of a GPU cluster? *Synthese, 206:* 218.

Shulman, Carl, and Nick Bostrom (2021). Sharing the world with digital minds. In S. Clarke, H. Zohny, and J. Savulescu, eds., *Rethinking moral status.* Oxford University Press.

Sidgwick, Henry (1874/1907). *The methods of ethics, 7th ed.* Macmillan.

Singer, Peter (1975/2023). *Animal liberation now.* HarperCollins.

Singer, Peter (1980/2011). *Practical ethics, 3rd edition.* Cambridge University Press.

Singer, Peter (1981/2011). *The expanding circle, 2nd ed.* Princeton University Press.

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

Singh, Jaivir (2025). Juristic personhood and property: Some reflections on the juridical path of the idol in India. *Asian Journal of Law and Society, 12,* 346-372.

Sinhababu, Neil (2024). The epistemic argument for hedonism. In S. Chakraborty, ed., *Human minds and cultures.* Springer Nature.

Sinnott-Armstrong, Walter, and Vincent Conitzer (2021). How much moral status could Artificial Intelligence ever achieve? In S. Clarke, H. Zohny, and J. Savulescu, eds., *Rethinking moral status.* Oxford University Press.

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

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

Sober, Elliott, and David S. Wilson (1998). *Unto others.* Harvard University Press.

Solum, Lawrence B. (1992). Legal personhood for Artificial Intelligences. *North Carolina Law Review, 70,* 1231.

Sparrow, Robert (2004). The Turing triage test. *Ethics and Information Technology, 6,* 203-213.

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

Sparrow, Robert (2017). Robots, rape, and representation. *International Journal of Social Robotics, 9,* 465-477.

Sparrow, Robert (2019). Yesterday's child: How gene editing for enhancement will produce obsolescence – and why it matters. *American Journal of Bioethics, 19,* 6-15.

Stanton, Andrew, and Jim Reardon (2008). *WALL-E.* Pixar / Disney.

Steinhart, Eric C. (2014). *Your digital afterlives.* Palgrave Macmillan.

Stevenson, Lloyd G. (1954). On the supposed exclusion of butchers and surgeons from jury duty. *Journal of the History of Medicine and Allied Sciences, 9,* 235-238.

Stone, Christopher D. (1972/2010). *Should trees have standing? 3rd ed.* Oxford University Press.

Strasser, Anna (2026). *Inbetweenism: Why our concepts reach their limits when it comes to generative AI.* De Gruyter Brill.

Street, Sharon (2006). A Darwinian dilemma for realist theories of value. *Philosophical Studies, 127,* 109-166.

Street, Sharon (2008). Constructivism about reasons. In R. Shafer-Landau, ed., *Oxford studies in metaphysics, vol. iii.* Oxford University Press.

Sun, Xiaoran, Yunqi Wang, and Brandon T. McDaniel (2026). AI companions and adolescent social relationships: Benefits, risks, and bidirectional influences. *Child Development Perspectives, 20,* 91-98.

Sylvan (Routley), Richard (1973/2008). Is there need for a new, an environmental ethic? In R. Attfield, ed., *The ethics of the environment.* Ashgate.

Tännsjö, Torbjörn (1998). *Hedonistic utilitarianism.* Edinburgh University Press.

Tchaikovsky, Adrian (2015). *Children of time.* Tor.

Temkin, Larry S. (1993). *Inequality.* Oxford University Press.

Thaler, Richard H., and Cass R. Sunstein (2008/2021). *Nudge: The final edition.* Penguin Random House.

Tunick, Mark (2025). Robbie, Klara, and Ethan: Replicability and the moral status of AI. *Journal of Science Fiction and Philosophy, 8.* URL: https://jsfphil.org/volume-8-2025/replicability-and-the-moral-status-of-ai/

Twain, Mark (1900/1969). The chronicle of young Satan. In W.M. Gibson, ed., *Mark Twain's Mysterious Stranger manuscripts.* University of California Press.

Tubert, Ariela, and Justin Tiehen (forthcoming). *Authentic Artificial Intelligence.* Bloomsbury.

Tucker, Miles (2016). Two kinds of value pluralism. Utilitas, 28, 333-346.

Turing, Alan M. (1950). Computing machinery and intelligence. *Mind, 59,* 433-460.

Turkle, Sherry (2011). *Alone together.* Basic Books.

Turkle, Sherry (2024). Who do we become when we talk to machines? URL: https://mit-genai.pubpub.org/pub/uawlth3j/release/2.

Tye, Michael (2003). *Consciousness and persons.* MIT Press.

Tye, Michael (2017). *Tense bees and shell-shocked crabs.* Oxford University Press.

Vallor, Shannon (2016). *Technology and the virtues.* Oxford University Press.

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

Varner, Gary E. (2002). *In nature's interests?* Oxford University Press.

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

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

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

Višak, Tatjana (2013). *Killing happy animals.* Palgrave Macmillan.

Vogt, Katja Maria (2017). The Stoics on virtue and happiness. In C. Bobonich, ed., *The Cambridge companion to ancient ethics.* Cambridge University Press.

Vold, Karina (2015). The parity argument for extended consciousness. *Journal of Consciousness Studies, 22* (3-4), 16-33.

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

Walton, Kendall (1978). Fearing fictions. *Journal of Philosophy, 75,* 5-27.

Wang Yangming (1527/2014). Questions on the *Great Learning,* trans. P. J. Ivanhoe. In J. Tiwald and B. W. Van Norden, eds., *Readings in later Chinese philosophy.* Hackett.

Warren, Mary Anne (1997). *Moral status.* Oxford University Press.

Wedgwood, Ralph (2023). Hierocles' concentric circles. *Oxford Studies in Ancient Philosophy, 62,* 293-332.

Wenar, Leif, and Rowan Cruft (2005/2025). Rights. *Stanford Encyclopedia of Philosophy* (summer 2025 edition).

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

Whitby, Blay (2008). Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents. *Interacting with Computers, 20,* 326-333.

White, Justin F. (2022). Why did the butler do it? *European Journal of Philosophy, 30,* 374-393.

Williams, Bernard (1973). *Problems of the self.* Cambridge University Press.

Williams, Bernard (1981). *Moral luck.* Cambridge University Press.

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

Williams, Thomas J. V., and Camilo T. Gomez (2026). *Post-humanistic love.* Cambridge University Press.

Willoughby, Brian J., Jason S. Carroll, Carson R. Dover, and Rebekah H. Hakala (2025). Counterfeit connections: The rise of romantic AI companions and AI sexualized media among the rising generation. Wheatley Institute. URL: https://brightspotcdn.byu.edu/a6/a1/c3036cf14686accdae72a4861dd1/counterfeit-connections-report.pdf.

Wilson, Robert A. (2026). *Eugenification+*. Unpublished manuscript.

Wong, David (2025). *Metaphor and analogy in Chinese thought.* Oxford University Press.

Wright, Larry (1973). Functions. *Philosophical Review, 82,* 139-168.

Xiang, Chloe (2023). "He would still be here": Man dies by suicide after talking with AI chatbot, widow says. *Vice* (Mar 30). URL: https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says.

Xie, Tianling, and Iryna Pentina (2022). Attachment theory as a framework to understand relationships with social chatbots: A case study of Replika. *Proceedings of the 55th Hawaii International Conference on System Sciences,* 2046-2055.

Yampolskiy, Roman V. (2024). *AI: Unexplainable, unpredictable, uncontrollable.* Taylor & Francis.

Yasukawa, Kanako (2026). Model welfare or user welfare? *PhilPapers*: https://philpapers.org/rec/YASMWO.

Yudkowsky, Eliezer, and Nate Soares (2025). *If anyone builds it, everyone dies.* Little, Brown, and Company.

Zangwill, Nick (2021). Our moral duty to eat meat. *Journal of the American Philosophical Association, 7*, 295-311.

Zhuangzi (4th c. BCE/2024). *Zhuangzi: Complete writings,* trans. C. Fraser. Oxford University Press.

Ziesche, Soenke, and Roman V. Yampolskiy (2025). *Considerations on the AI endgame.* Taylor and Francis.