Strange Intelligence: Moral Puzzles of Unhumanlike AI

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

June 25, 2026

Strange Intelligence: Moral Puzzles of Unhumanlike AI

Abstract: Future AI persons might both (1.) deserve moral consideration and rights fully equal with natural human persons, and (2.) have lifeways so radically different from ours as to break familiar patterns of moral thinking by violating our ordinary background assumptions. This article presents a series of thought experiments about strange AI persons, centering on a two-pronged worry featuring two types of "monster". "Utility monsters", who derive great personal benefit from harming others, create a well-known challenge for ethical systems that aim to maximize aggregate goods. The less-discussed case of "fission-fusion monsters", who can divide and merge at will, presents a complementary challenge to ethical systems focused on individual rights, since individual rights frameworks require the existence of stable, countable individual persons. AI cases dramatically expand the range of possible lifeways, creating untested problem cases for ethical systems that assume persons of the familiar humanlike sort.

Word Count: ~9300 words

Keywords: Artificial Intelligence, ethics, personal identity, robot rights, utility monster

Strange Intelligence: Moral Puzzles of Unhumanlike AI

Sufficiently humanlike AI would deserve humanlike rights. This is, I hope, uncontroversial, even if the *basis* of moral considerability (sentience? autonomy? social relations?) is debatable.[1] If some artificial entities ever become humanlike enough in all the ways that plausibly matter, they will be our moral equals. They will be, that is, moral *persons,* deserving all the rights and consideration of personhood.[2] However, moral personhood is plausibly consistent with being radically unhumanlike in other ways. The existence of AI persons with radically unhumanlike lifeways risks breaking our familiar patterns of moral thinking by violating our ordinary background assumptions.

This article presents a series of thought experiments about strange AI persons, aiming to induce moral perplexity, or at least uncertainty.[3] The cases raise, I contend, 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.[4]

I will develop a two-pronged worry, focused on two types of "monster". As Robert Nozick (1974) has 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

[1] See Schwitzgebel and Garza 2015; Gunkel 2018; Long et al. 2024; Dung forthcoming; Goldstein and Kirk-Giannini forthcoming.

[2] Even if it's impossible for AI systems ever to be humanlike in the relevant respects (for example, because AI consciousness is impossible and consciousness is necessary for moral considerability), the equality claim remains true, just empty of instances.

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

[4] 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.

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

For purposes of this article, I assume that humanlike consciousness is possible in near-to-medium-term future AI systems. This permits us to focus cleanly on the target issue of different lifeways. Adding justified doubt about AI consciousness should tend to complicate rather than simplify the issues, strengthening the case for skepticism and unpreparedness.[5]

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,

[5] In Schwitzgebel forthcoming-b, I explore arguments for and against near-to-medium-term future AI consciousness, arguing that neither the arguments pro nor the arguments con are decisive.

they have 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[6]) can experience a million times more positive emotion per second than an ordinary human, as described above. Apart from this 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.

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

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.[7]

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.

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

(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 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. 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.[8]

(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 that in distributing goods we should favor the worse off.[9] 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.

[8] 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.

[9] 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.

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.[10] Across radically different kinds of minds, they might be epistemically intractable.

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

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

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

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.[11]

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.[12]

Your uphill neighbor and her cat move out and are replaced by two new AI neighbors, Shriya and Alaleh.[13] 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

[11] 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.

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

[13] Names randomly chosen from lists of former lower-division students, excluding Jesus, Muhammad, and very uncommon names.

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 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.[14]

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?

[14] 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. AI personhood could convert Parfit's puzzles from far-fetched science fiction to real-world practical problems.

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. Conversely, 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 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 distinct but similar 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.[15]

[15] 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 article, 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: Schwitzgebel and Garza 2020; Schwitzgebel forthcoming-a), (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.

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.[16] 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 puzzles of eugenics, which could easily fill its own article.[17]) Duplicating one's current AI self thus constitutes one end of a continuum of AI creation from infancy to maturity.

If 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.[18]

---

[16] Compare Egan 1997 on "orphanogenesis".

[17] On the ethics of disability, eugenics, and human enhancement, see e.g., Glover 2006; Buchanan 2011; Sparrow 2011, 2019; Garland-Thomson 2012; Savulescu and Kahane 2017; Anomaly 2020/2024; Wilson 2026. I 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.

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

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. 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, seven years, a typical waiting time for immigrants to apply for citizenship – could still generate political chaos after a few election cycles.

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?[19] 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 can'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

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

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 article.

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.

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.[20] 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 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?

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?

---

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

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.[21]

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

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

promises from today, about which she has not yet learned. If Shriya is wealthy or fission cheap, she might potentially divide into many more than two.

Let's call this type of person or persons a Fission-Fusion Monster.[22] 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;
- 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

[22] 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!

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*).[23] 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 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.[24] 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

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

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

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.[25] 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 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.[26] 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.

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

[26] Schwitzgebel 2023; Schwitzgebel and Nelson 2026.

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 (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?[27]

[27] 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.

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 could 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 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.

As many philosophers have noticed, current language models are not easily divided into discrete bundles of the sort familiar from vertebrate biology.[28] 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, each constituting a wholly separate conscious existence that soon expires. If language models become (or are) conscious, 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?[29] 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?

---

[28] See references in note 15.

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

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 article: 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.

6. The Policy of Humanlike Design.

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 is this: *Don't create such AI systems.* Don't create 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 about morally perplexing AI systems. The world is good enough, perhaps, without them. Thus, I recommend:

> 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.

Unfortunately, the Policy of Humanlike Design 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 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 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 forward at maximum speed.

References

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

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

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

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

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

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

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

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

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

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,* S3:E0.

Brooker, Charlie, and Tim Van Patten (2017). Hang the DJ. *Black Mirror,* S4:E4.

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

Chalmers, David J. (2025/2026). What do we talk to when we talk to language models? *PhilPapers:* https://philarchive.org/rec/CHAWWT-8.

Chappell, Richard Y. (2021). Negative utility monsters. *Utilitas, 33,* 417-421.

Chi, Quentin (forthcoming). A new argument for partial unity. *Ergo.*

Chilson, Kendra, and Eric Schwitzgebel (2026). Artificial Intelligence as strange intelligence: Against linear models of intelligence. *ArXiv*:2602.04986.

Dainton, Barry (2000). *Stream of consciousness.* Routledge.

Danaher, John, and Sven Nyholm (2025). The ethics of personalised digital duplicates: a minimally viable permissibility principle. *AI and Ethics, 5,* 1703-1718.

Dennett, Daniel C. (1978). *Brainstorms.* MIT Press.

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.

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

Ewen, Jacob (2026). The bearer problem for large language models. *PhilPapers:* https://philarchive.org/rec/EWETBP.

Garland-Thomson, Rosemarie (2012). The case for conserving disability. *Bioethical Inquiry, 9,* 339-355.

Garson, Justin (2016). Two types of psychological hedonism. *Studies in History and Philosophy of Biology and Biomedical Sciences, 56,* 7-14.

Glover, Jonathan (2006). *Choosing children.* Oxford University Press.

Goldstein, Simon, and Cameron D. Kirk-Giannini (forthcoming). *AI welfare.* Oxford University Press.

Goldstein, Simon, and Harvey Lederman (forthcoming). AI death. *Philosophical Perspectives.*

Gunkel, David (2018). *Robot rights.* MIT Press.

Hanson, Robin (2016). *The age of em.* Oxford University Press.

Hausman, Daniel M. (1995). The impossibility of interpersonal utility comparisons. *Mind, 104,* 473-490.

Jones, Max, James Ladyman, and Ryan M. Nefdt (2026). Counting (on) large language models. *PhilPapers:* https://philarchive.org/rec/JONCOL.

Leckie, Ann (2013). *Ancillary justice.* Orbit.

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.

Nagata, Linda (1995). *Bohr maker.* Mythic Island.

Nagata, Linda (2019). *Edges.* Mythic Island.

Niven, Larry (1969/1995). Death by ecstasy. In L. Niven, *Flatlander.* Del Rey.

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

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.

Register, Chris (2025). Individuating artificial moral patients. *Philosophical Studies, 182,* 3225-3246.

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.

Savulescu, Julian, and Guy Kahane (2017). Understanding procreative beneficence. In L. Francis, ed., *Oxford Handbook of Reproductive Ethics.* Oxford University Press.

Schechter, Elizabeth (2018). *Self-consciousness and "split" brains.* Oxford University Press.

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

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

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 (2023). Borderline consciousness: When it's neither determinately true nor determinately false that experience is present. *Philosophical Studies, 180,* 3415-3439.

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

Schwitzgebel, Eric (forthcoming-a). Against designing "safe" and "aligned" AI persons (even if they're happy). In S. AbuMasab and D. Tamez, eds., *Social cognition and agency.* Bloomsbury.

Schwitzgebel, Eric (forthcoming-b). *AI and consciousness.* Cambridge University Press.

Schwitzgebel, Eric (forthcoming-c). How weird the world is: Reply to Machery, Roelofs, and Schneider. *Journal of Consciousness Studies.*

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

Schwitzgebel, Eric, and Mara Garza (2020). Designing AI with rights, consciousness, self-respect, and freedom. In S. M. Liao, ed., The 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.

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

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.

Sober, Elliott, and David S. Wilson (1998). *Unto others.* Harvard University Press.

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

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.

Temkin, Larry S. (1993). *Inequality.* Oxford University Press.

Tye, Michael (2003). *Consciousness and persons.* MIT Press.

Vold, Karina (2015). The parity argument for extended consciousness. *Journal of Consciousness Studies, 22* (3-4), 16-33.

Williams, Bernard (1973). *Problems of the self.* Cambridge University Press.

Wilson, Robert A. (2026). *Eugenification+*. Unpublished manuscript.

Ziesche, Soenke, and Roman V. Yampolskiy (2025). *Considerations on the AI endgame.* Taylor and Francis.