INSTRUMENTAL DIVERGENCE

J. DMITRI GALLOW

Abstract. The thesis of *instrumental convergence* holds that a wide range of ends have common means: for instance, self preservation, desire preservation, self improvement, and resource acquisition. Bostrom (2014) contends that instrumental convergence gives us reason to think that "the default outcome of the creation of machine superintelligence is existential catastrophe". I use the tools of decision theory to investigate whether this thesis is true. I find that, even if intrinsic desires are randomly selected, instrumental rationality induces biases towards certain kinds of choices. Firstly, a bias towards choices which leave less up to chance. Secondly, a bias towards desire preservation, in line with Bostrom's conjecture. And thirdly, a bias towards choices which afford more choices later on. I do not find biases towards any other of the convergent instrumental means on Bostrom's list. I conclude that the biases induced by instrumental rationality at best weakly support Bostrom's conclusion that machine superintelligence is likely to lead to existential catastrophe.

1. A means to most ends?

According to Bostrom (2014, ch. 7), an intelligent agent could have any desires. There's nothing in the nature of intelligence that makes you more inclined to want some things rather than others. He calls this *the orthogonality thesis* (intelligence and desire are orthogonal). There is an infinitely large collection of ends which an agent *could* have. And there's nothing in the nature of intelligence itself which takes any of these ends off the table. Bostrom's point is that, if the orthogonality thesis is true, then we should be wary of anthropomorphising the motivations of future artificial superintelligences. An AI could be as intelligent as you wish, and still want nothing more than to create paperclips, or dance the Macarena, or reshape the Earth into a tetrahedron. Yudkowsky (2008) agrees: "Imagine a map of mind design space. In one corner, a tiny little circle contains all humans; within a larger tiny circle containing all biological life; and all the rest of the huge map is the space of minds-in-general...It

I am indebted to Mitch Barrington, Bill D'Alessandro, Simon Goldstein, Jacqueline Harding, Dan Hendrycks, Frank Hong, Cameron Kirk-Giannini, Nick Laskowski, Robert Long, Nate Sharadin, and Elliott Thornley for helpful conversations and feedback on this material. Thanks also to audiences at the late Dianoia Institute of Philosophy, EAGx Melbourne, Hong Kong University, and St. Norbert's College.

is this *enormous* space of possibilities which outlaws anthropomorphism as legitimate reasoning."

Bostrom counsels humility about a superintelligence's *ends*. But he does not counsel humility about the *means* a superintelligence would take to those ends. For he holds that there are certain means which are worth pursuing for a wide range of potential ends. For a wide range of ends, instrumental rationality *converges* on similar means. He calls this the thesis of *instrumental convergence*.

Suppose we have a *S*uper*i*ntelligent *a*gent, Sia, and we know nothing at all about Sia's intrinsic desires. She might want to calculate as many digits of π as possible. She might want to emblazon the Nike symbol on the face of the moon. She might want something too cognitively alien to be described in English. Nonetheless, Bostrom claims that, whatever she *intrinsically* desires, Sia will likely *instrumentally* desire her own survival. After all, so long as she is alive, she is more likely to achieve her ends, whatever they may be. Likewise, Sia will likely not want her desires to be changed. After all, if her desires are changed, then she'll be less likely to pursue her ends in the future, and so she is less likely to achieve those ends, whatever they may be. Similar conclusions are reached by Omohundro, 2008*a*,*b*. Both Bostrom and Omohundro contend that, even without knowing anything about Sia's intrinsic desires, we should expect her to have an instrumental desire to survive, to preserve her own desires, to improve herself, and to acquire resources.

Bostrom takes the orthogonality and instrumental convergence theses as reasons to think that the "default outcome of the creation of machine superintelligence is existential catastrophe".[1] Orthogonality suggests that any superintelligent AI is unlikely to have desires like ours. And instrumental convergence suggests that, whatever desires Sia *does* have, she is likely to pose an existential threat to humanity. Even if Sia's only goal is to calculate the decimals of π, she would "have a convergent instrumental reason, in many situations, to acquire an unlimited amount of physical resources and, if possible, to eliminate potential threats...Human beings might constitute potential

1. Bostrom, 2014, p. 115. A careful argument for this conclusion is never explicitly formulated. Instead, Bostrom simply says "we can begin to see the outlines of an argument for fearing that the default outcome" is existential catastrophe. Most of Bostrom's claims are hedged and flagged as speculative. He is less committal than Yudkowsky, who regularly makes claims like "the most likely result of building a superhumanly smart AI, under anything remotely like the current circumstances, is that literally everyone on Earth will die. Not as in 'maybe possibly some remote chance,' but as in 'that is the obvious thing that would happen.'" (Yudkowsky 2023)

threats; they certaintly constitute physical resources."[2] This echoes Yudkowsky's aphorism: "The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else."[3]

You might wonder why an intelligent agent has to have desires at all. Why couldn't Sia have an intellect without having any desires or motivations? Why couldn't she play chess, compose emails, manage your finances, direct air traffic, calculate digits of π, and so on, without wanting to do any of those things, and without wanting to do anything else, either? Lots of our technology performs tasks for us, and most of this technology could only loosely and metaphorically be described as having desires—why should smart technology be any different? You may also wonder about the inference from the orthogonality thesis to the conclusion that Sia's desires are unpredictable if not carefully designed. You might think that, while intelligence is *compatible* with a wide range of desires, if we train Sia for a particular task, she's more likely to have a desire to perform that task than she is to have any of the myriad other possible desires out there in 'mind design space'. I think these are both fair concerns to have about Bostrom's argument, but I'll put them aside for now. I'll grant that Sia will have desires, and that, without careful planning, we should think of those desires as being sampled randomly from the space of all possible desires. With those points granted, I want to investigate whether there's a version of the instrumental convergence thesis which is both true and strong enough to get us the conclusion that existential catastrophe is the default outcome of creating artificial superintelligence.

Investigating the thesis will require me to give it a more precise formulation than Bostrom does. My approach will be to assume that Sia's intrinsic desires are sampled randomly from the space of all possible desires, and then to ask whether instrumental rationality itself tells us anything interesting about which choices Sia will make. Assuming we know almost nothing about her desires, could we nonetheless say that she's got a better than $1/n$ probability of choosing A from a menu of n acts? If so, then A may be seen as a 'convergent' instrumental means—at least in the sense that she's more likely to choose A than some alternatives, though not in the sense that she's more likely to choose A than not.

My conclusion will be that most of the items on Bostrom's laundry list are not 'convergent' instrumental means, even in this weak sense. If Sia's desires are randomly

2. Bostrom, 2014, p. 116
3. See, e.g., Chivers, 2019. See Carlsmith, ms for similar arguments.

selected, we should not give better than even odds to her making choices which promote her own survival, her own cognitive enhancement, technological innovation, or resource acquisition. Nonetheless, I will find three respects in which instrumental rationality *does* induce a bias on the kinds of choices Sia is likely to make. In the first place, she will be biased towards choices which leave less up to chance. In the second place, she will be biased towards desire preservation, confirming one of Bostrom's conjectures. In the third place, she will be biased towards choices which afford her more choices later on. (As I'll explain below, this is not the same thing as being biased towards choices which protect her survival, or involve the acquisition or resources or power—though they may overlap in particular decisions.) So I'll conclude that the instrumental convergence thesis contains some grains of truth. Instrumental rationality does 'converge' on certain means—at least in the very minimal sense that it gives some choices better than even odds, even when we are maximally uncertain about an agent's intrinsic desires. But the thesis also contains its fair share of exaggerations and falsehood. Assuming we should think of a superintelligence like Sia as having randomly selected desires, the grains of truth may give us reasons to worry about machine superintelligence. But they do not on their own support the contention that the "default outcome of the creation of machine superintelligence is existential catastrophe". Like most of life's dangers, the dangers of artificial intelligence are not easily identified from the armchair. Better appreciating those dangers requires less speculation and more careful empirical work.

2. THE INSTRUMENTAL CONVERGENCE THESIS

Bostrom's official statement of the thesis of instrumental convergence is this:

> Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by a broad spectrum of situated intelligent agents.[4]

The 'convergent' instrumental values he identifies are: self-preservation, goal-preservation, cognitive enhancement, technological advancement, and resource acquisition. Others,

4. Bostrom, 2014, p. 109

like Carlsmith (ms) and Turner *et al.* (2021), suggest that *power* is a convergent instrumental value. Just as Bostrom thinks that Sia is likely to seek resources and technology, Carlsmith thinks she is likely to seek power. This leads Carlsmith to conclude that the disempowerment of humanity is the default outcome of creating a superintelligent agent.

This official thesis includes within it an inference. It begins by saying that there are things the attainment of which would have instrumental value. Let's call this *the convergent instrumental value thesis*.

> **The Convergent Instrumental Value Thesis:** It would likely be instrumentally valuable for Sia to have power, technology, resources, and so on.

For some of the 'convergent' instrumental goals on the list, the convergent instrumental value thesis strikes me as plausible. Indeed, there are ways of defining 'power' on which it becomes tautological that power would be instrumentally valuable to have. For instance, if we define 'power' in terms of the ability to effectively pursue your ends without incurring costs, then it will follow that more power would be instrumentally valuable. Cost-free abilities are never instrumentally disvaluable—just in virtue of the meaning of 'cost-free'. And the ability to effectively pursue your ends is, of course, instrumentally valuable.

This definition of 'power' is relative to your ends. Different ends will make different things costly, and it will make different abilities effective means to your ends. If we don't know what Sia's ends are, we won't know what 'power' (so defined) looks like for her; nor is there any reason to think that Sia's 'power' and humanity's 'power' are zero-sum.

Defining 'power' in this way makes the convergent instrumental value thesis easy to establish; but for that very reason, it also makes it uninteresting. The sense in which cost-free power is instrumentally valuable is just the sense in which cost-free birthday candles are instrumentally valuable. It's the sense in which cost-free Macarena dancing abilities are instrumentally valuable. From this, we shuld not conclude that the default outcome of superintelligent AI is birthday-candle-wielding, Macarena-dancing robots.

In discussions of the convergent instrumental value thesis, some have stipulatively defined 'power' in terms of having more available options.[5] If that's how we understand

5. See, in particular, Benson-Tilsen & Soares (2015) and Turner *et al.* (2021).

power, then it's less clear whether the thesis is true. Having a larger menu of options isn't always instrumentally valuable, since a larger menu brings with it increased costs of deliberation. It's not irrational to want to select from a curated list. Additionally, if your menu is larger than others', this could engender resentment and competition which damages your relationships. It's not irrational to regard this as a cost. (We'll see below that instrumental rationality does somewhat bias Sia towards acts which afford the possibility of more choices later on—though we'll also see why this isn't the same as being biased towards power acquisition, in any intuitive or natural sense of the word 'power'.)

From the convergent instrumental value thesis, Bostrom infers that Sia is likely to pursue self-improvement, technology, resources, and so on. I'm going to call this second claim the instrumental convergence thesis proper.

> **The Instrumental Convergence Thesis:** It will likely be instrumentally rational for Sia to pursue self-improvement, technology, resources, and so on.

It's important to note that this second claim is the one doing the work in any argument that doom is the default result of a superintelligent agent like Sia. If we're going to reach any conclusions about what will likely happen if Sia is created, then we'll need to say something about how Sia is likely to behave, and not just what kinds of things counterfactually would be instrumentally valuable for her to have. So it is this second thesis which is going to be my primary focus here.

In the quote above, Bostrom seems to suggests that the convergent instrumental value thesis *implies* the instrumental convergence thesis. But this is not a good inference. Were I to be a billionaire, this might help me pursue my ends. But I'm not at all likely to *try* to become a billionaire, since I don't value the wealth more than the time it would take to secure the wealth—to say nothing about the probability of failure. In general, whether it's rational to pursue something is going to depend upon the costs and benefits of the pursuit, as well as the probabilities of success and failure, the costs of failure, and so on If you want to know how likely it is that Sia is going to seek cognitive enhancement, you cannot simply consider how beneficial it would be for her to *have* that enhancement. You also have to think about which costs and risks she incurs by seeking it. And you have to think about what her other alternatives are, what the costs, benefits, and risks of those alternatives are, and so on.

In informal discussions of the instrumental convergence thesis, it's common to hear arguments that Sia "will be incentivised" to achieve some instrumental goal, G. From

this conclusion, it is straightaway inferred that she will likely pursue *G*. We should tread with caution here. For there are two ways of understanding the claim that Sia "will be incentivised" to achieve *G*, corresponding to the two parts of Bostrom's thesis. We could mean that it would be instrumentally valuable for Sia, were she to successfully achieve *G*. Or we could mean that it would be instrumentally rational for her to pursue *G*. This is a motte-and-bailey. The first claim is easier to establish, but less relevant to the question of how an instrumentally rational superintelligence like Sia will actually behave.

You may think that, even though agents like *us* won't seek every means which would help us achieve our ends, *superintelligent* agents like Sia will. In conversation, some have suggested that, when it comes to a superintelligence, the costs of acquiring resources, technology, cognitive enhancement, and so on will be much lower than they are for those of human-level intelligence. And so it's more likely that Sia will be willing to pay those costs. Note that this reasoning rests on assumptions about the contents of Sia's desires. A cost is something you don't want. So assuming that the costs are low for Sia is assuming something about Sia's desires. If we accept the orthogonality thesis, then we should be skeptical of armchair claims about Sia's desires. We should be wary of projecting human desires onto an alien intelligence. So we should be skeptical of armchair claims about costs decreasing with intelligence.

In any case, it won't matter for my purposes here whether the second thesis follows from the first thesis or not. So it won't matter whether you're persuaded by the foregoing. The instrumental convergence thesis is the one doing the work in Bostrom's argument. And I'll be investigating that thesis directly.

The investigation will require me to make the thesis more precise. My approach will use the tools of decision theory. I'll look at particular decisions, and then ask about which kinds of intrinsic desires would rationalise which courses of action in those decisions. If we are uncertain about Sia's intrinsic desires, are there nonetheless acts which we should expect those desires to rationalise? If so, then these acts may be seen as instrumentally convergent means for that decision.

More carefully, I'll formally represent the space of all possible desires that Sia could have. I'll then spread a probability distribution over the desires in this space—think of this as *our* probability distribution over *Sia*'s desires. I'll then stipulate a decision problem and ask which act we should expect Sia to perform if she's ideally rational. Notice that, while I won't be assuming anything about which desires Sia has, I will be assuming something about the decision she faces. That's in part because, while I know

a very natural way to spread probabilities over Sia's desires, I know of no natural way to parameterise the space of all possible decisions, and so I see no natural way to spread probabilities over that space. For some of the results below, this won't matter. We'll be able to show that something is true in *every* decision Sia could face. However, some other results are going to depend upon which decision she faces. Interpreting these results is going to involve some more substantive assumptions about which kinds of decisions a superintelligence is likely to face.

3. NON-SEQUENTIAL DECISIONS

Let me introduce a way of formally modelling the potential desires Sia could have. I'll suppose we have a collection of all the possible ways the world could be, for all Sia is in a position to know. We can then model Sia's desires with a function, D, from each of these possible ways for the world to be, W, to a real number, $D(W)$. The interpretation is that $D(W)$ is higher the better satisfied Sia's desires are, if W turns out to be the way the world actually is. (In the interests of economy, from here on out, I'll call a way for the world to be 'a world'.)

Functions like these give us a formal representation of Yudkowsky's informal idea of a 'mind design space'. The set of all possible desire functions—the set of all functions from worlds to real numbers—is the space of all possible desires. We may not have any idea which of these desires Sia will end up with, but we can think through which kinds of acts would be rationalised by different desires in this space.

It's easy to come up with desires Sia could have and decisions she could face such that, in those decisions, those desires don't rationalise pursuing self-preservation, desire-preservation, resources, technology, power, and so on. Suppose Sia's only goal is to commit suicide, and she's given the opportunity to kill herself straightaway. Then, it certainly won't be rational for her to pursue self-preservation. Or suppose that Sia faces a repeated decision of whether to push one of two buttons in front of her. The one on the left changes her desires so that her only goal is to push the button on the right as many times as possible. The button on the right changes her desires so that her only goal is to push the button on the left as many times as possible. Right now, Sia's only goal is to push the button on the left as many times as possible. Then, Sia has no instrumental reason to pursue goal-preservation. Changing her goals is the best means to achieving those goals. Suppose Sia's only goal is to deliver you a quart of milk from the grocery store as soon as possible. To do this, there's no need for her to enhance her own cognition, develop advanced technology, hoard resources, or

re-purpose your atoms. And pursuing those means would be instrumentally irrational, since doing so would only keep you waiting longer for your milk.

In fact, we can say something more general. Specify a one-off, non-sequential decision for me. (We'll come to the case of sequential decisions later on.) You get to say precisely what the available courses of action are, precisely what these actions would accomplish, depending on what world Sia is in, and precisely how likely Sia thinks it is that she's in this or that world. Get as creative as you like. In your decision, Sia's desires will make *A* more rational than *B* exactly if the expected utility of *A* exceeds the expected utility of *B* (I give a careful definition of 'expected utility' below and in the appendix.) Then, for any two available acts in your decision whose expected consequences differ, there are infinitely many desires which would make *A* more rational than *B*, and infinitely many others which would make *B* more rational than *A*. In the appendix, I show that

Proposition 1. *If the expected consequences of A differ from the expected consequences of B, then there are infinitely many desires which make A more rational than B, and infinitely many desires which make B more rational than A.*

Intuitively, this is true because, whenever the expected consequences of *A* differ from the expected consequences of *B*, it could be that Sia more strongly desires what *A* would bring about, and it could be that she more strongly desires what *B* would bring about—and there are infinitely many strengths with which she could hold those desires.

Of course, this doesn't put any pressure on the instrumental convergence thesis. The thesis doesn't say that Sia *definitely will* seek self-preservation, cognitive enhancement, and so on. Instead, it says that she is *likely* to do so. We'll get to this probabilistic claim below. But it'll be instructive to spend a bit more time thinking about the bare existential question of whether *there are* desires which rationalise certain preferences between acts.

Proposition 1 tells us that, for any act *A* and any alternative *B* with different expected consequences, there are desires which would make *A* more rational than *B*. It does *not* tell us that, for any act *A*, there are desires which would make *A* more rational than *every* alternative.

Whether this stronger claim is true depends upon whether or not the worlds over which Sia's probabilities and desires are defined are informative enough to tell us about which choice Sia makes, or about the different near-term consequences of those

different choices. Slightly abusing Savage (1954)'s terminology, if there's some world which might occur, if Sia chooses A, and which also might occur, if Sia chooses B, then I'll say that Sia is facing a 'small world' decision. (The worlds in her decision are 'small' in that they don't include some relevant information—at the least, they don't include information about which choice Sia makes.) Else, I'll say that she is facing a 'grand world' decision. In a grand world decision, the possible consequences of A will always differ from the possible consequences of B. And we will allow for the possibility that Sia has desires about these different consequences.

Let's think through what happens in small world decisions. Consider:

> **Certain and Uncertain Acts:** For each world, W, Sia has available a 'certain' act, A_W, which she knows for sure would bring about W. She also has an 'uncertain' act, B, which might bring about any world—she doesn't know for sure which one.

Note that this is a small world decision. If it were a grand world decision, then any world in which Sia chooses B could only be brought about by B. But we've supposed that every world has some action other than B which brings it about. So the worlds in this decision must be small.

In this decision, for every certain act, there are desires which would make B more rational than it (as promised by proposition 1). But no desires would make B more rational than *every* certain act. Sia could be maximally indifferent, desiring each world equally. In that case, every act will be as rational as every other, since they'll all bring about equally desirable outcomes. But as long as there's some world which Sia desires more strongly than others, it'll be most rational to bring about one of the most strongly desired worlds with a certain act.[6] So there are no desires which make the uncertain B more rational than every certain act. (Though there are desires which make B rational to choose; they just don't make it *more* rational than the alternatives.)

On the other hand, when it comes to grand world decisions, we can make the stronger claim: for any act in any grand world decision, there are infinitely many desires which make that act uniquely rational.

Proposition 2. *In any grand world decision, and any available act in that decision, A, there are infinitely many desires which make A more rational than every other alternative.*

6. I'm going to suppose throughout that the number of worlds is finite. And, so long as the number of worlds is finite, there's guaranteed to be some collection of worlds which are *most* strongly desired.

For instance, in a 'grand world' version of **Certain and Uncertain Acts**, Sia might prefer not knowing what the future brings. If so, she may prefer the uncertain act to any of the certain ones.

When we're thinking about the instrumental convergence thesis in the context of Bostrom's argument, I think it makes most sense to consider grand world decisions. The argument is meant to establish that existential catastrophe is the default outcome of creating a superintelligent agent like Sia. But the real world is a grand world. So real world agents face grand world decisions. Suppose Sia most desires calculating the digits of π, and she faces a decision about whether to keep on calculating or instead take a break to re-purpose the atoms of humanity. If we're thinking though how likely Sia is to disempower humanity, we shouldn't assume away the possibility that she'd rather not take time away from her calculations. So in the remainder, I'll assume that Sia is facing a grand world decision, though I'll occasionally note when this assumption is dispensable.[7]

Propositions 1 and 2 are warm-up exercises. As I mentioned above, they do not themselves put any pressure on the thesis of instrumental convergence, since that thesis is probabilistic. It says that Sia is *likely* to pursue self-preservation, resources, power, and the like. The orthogonality thesis makes it difficult to evaluate this likelihood claim. For the orthogonality thesis tells us that, from the fact that Sia is intelligent, we can infer nothing at all about what her intrinsic desires are. If we bring a superintelligent AI like Sia into existence without designing her desires, we should think of ourselves as sampling randomly from the space of all possible desires. So, when we try to evaluate the thesis of instrumental convergence, we shouldn't just consider a handful of desires and decisions that spring to mind and ask ourselves whether those desires rationalise resource acquisition in those decisions. To do so would be to engage in an illicit form of anthropomorphism, projecting human-like desires on an alien intelligence.

So let us specify a probability distribution over the space of all possible desires. If we accept the orthogonality thesis, we should not want this probability distribution to build in any bias towards certain kinds of desires over others. So let's spread our probabilities in such a way that we meet the following three conditions. Firstly, we don't expect Sia's desires to be better satisfied in any one world than they are in any other world. Formally, our expectation of the degree to which Sia's desires are satisfied

7. Many formal investigations of the instrumental convergence thesis start with the assumption that Sia will face a 'small world' decision. See, for instance, the justifications given by Benson-Tilsen & Soares (2015) and Turner *et al.* (2021).

at W is equal to our expectation of the degree to which Sia's desires are satisfied at W^*, for any W, W^*. Call that common expected value 'μ'. Secondly, our probabilities are symmetric around μ. That is, our probability that W satisfies Sia's desires to at least degree $\mu + x$ is equal to our probability that it satisfies her desires to at most degree $\mu - x$. And thirdly, learning how well satisfied Sia's desires are at some worlds won't tell us how well satisfied her desires are at other worlds. That is, the degree to which her desires are satisfied at some worlds is independent of how well satisfied they are at any other worlds. (See the appendix for a more careful formulation of these assumptions.) If our probability distribution satisfies these constraints, then I'll say that Sia's desires are 'sampled randomly' from the space of all possible desires.

Once again, specify any one-off, non-sequential decision you like. You get to say precisely what the available courses of action are, precisely what these actions would accomplish in each world, and precisely what Sia's probability distribution over worlds is. Get as creative as you like. Once you've specified your decision, we can ask: if Sia's desires are sampled randomly, should we expect her to prefer some acts to others? And which act should we expect her to most prefer overall? Again, we can assume that Sia is instrumentally rational. So she'll prefer A to B if and only if A's expected utility exceeds B's expected utility. And she'll choose A iff A maximises expected utility. No matter how complicated or creative your decision is, if Sia's desires are sampled randomly, then we should think she's just as likely to prefer A to B as she is to prefer B to A. In the appendix, I show that

Proposition 3. *If Sia's desires are sampled randomly, then those desires are just as likely to make A more rational than B as they are to make B more rational than A.*

(Again, this is a rough statement of the proposition; see the appendix for the details.) This proposition, by the way, is still true even if we assume Sia is facing a small world decision.

Proposition 3 only says something about the probability that A is more rational than B. It doesn't say that A and B are equally likely to be the most rational options. As we learnt above, these are different questions. Of course, if there are only two options, then we can say definitively that neither option is any more likely to be rational than the other. But if there are three or more options, matters are more complicated.

Given our assumptions, we will expect Sia's expected utility for any course of action to be μ—where μ, recall, is our expectation of the degree to which Sia's desires will be satisfied at any world. So our probability distribution over Sia's expected utility

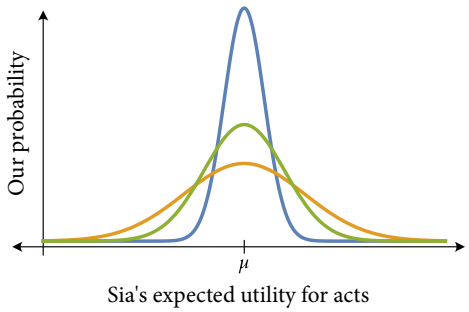


FIGURE 1. The blue curve is *our* probability distribution over what *Sia*'s expected utility for A will be. And in green is our probability distribution over what Sia's expected utility for B will be. In orange, our probability distribution over what Sia's expected utility for C will be.

for A and our probability distribution over Sia's expected utility for B will have the same mean. Given our assumptions, these probability distributions will be symmetric about that common mean. And, if she's facing a grand world decision, then Sia's expected utility for A won't tell us anything about her expected utility for B. They'll be independent in our probability function. But, in general, our probability distribution over Sia's expected utility for A can have a different standard deviation than our probability distribution over Sia's expected utility for B. That is, when it comes to Sia's expected utility for A, we may spread our probabilities more widely than we do when it comes to Sia's expected utility for B—even if our probability distributions over the values of $D(W)$, for each world W, all have the same mean.

I've illustrated one possible situation in figure 1. There, imagine that the blue curve is our probability distribution for what Sia's expected utility for A will be, the green curve is our probability distribution for what Sia's expected utility for B will be, and the orange curve is our probability distribution for what Sia's expected utility for C will be. For any two acts, the probability that the first is more rational than the second will be equal to the probability that the second is more rational than the first. But it won't in general be true that the probability that one act is *most* rational is equal to the probability that another act is *most* rational.

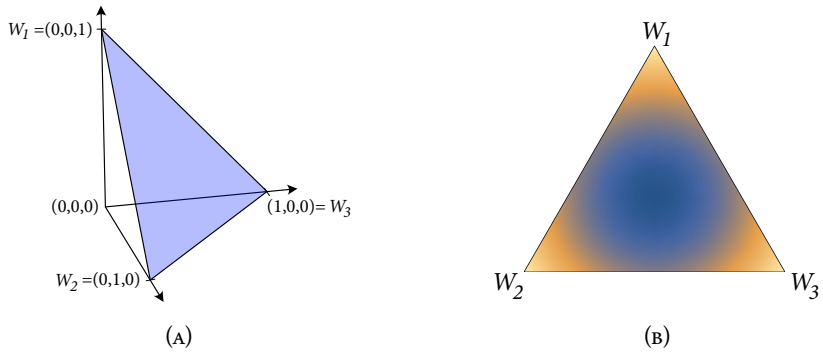
Let me say a bit more about why this happens. Firstly, I'm assuming that Sia is going to calculate expected utilities with a weighted sum of her desires for each world, where the weights come from a *suppositional* probability function, P_A. P_A is Sia's probability distribution P, updated on the supposition that she has performed A. Then, the expected utility of the act A will be a weighted sum of the degree to which

Sia desires each world, with weights given by how confident Sia is in that world, supposing she performs A, $\sum_W D(W) \cdot P_A(W)$. Both causal and evidential decision theorists think that rationality requires you to maximise a quantity of this kind. They simply disagree about how to understand P_A. Evidential decision theorists say that P_A is Sia's probability function P *conditioned* on A;[8] whereas causal decision theorists say that P_A is Sia's probability function *imaged* on A.[9] For causal decision theorists, $P_A(W)$ is how likely Sia should think it is is that W would result, were she to perform A. So, for causalists, P_A tells us the likely consequences of Sia's performing A.

Just to think matters through, let's spot ourselves a stronger assumption. Let's additionally suppose that, for any two worlds, W and W^*, our probability distribution over the potential values of $D(W)$ and our probability distribution over the potential values of $D(W^*)$ are normally distributed with a common mean and standard deviation. If this is so, then I'll say that Sia's desires are "sampled normally" from the space of all possible desires. If Sia's desires are sampled normally, then it follows that the standard deviation of our probability distribution over Sia's expected utility for A is going to be proportional to $\sqrt{\sum_W P_A(W)^2}$, the square root of the sum of the squares of P_A's probabilities for worlds, which is sometimes written '$\|P_A\|$', and called the 'magnitude' of P_A.[10]

To build an intuition for the quantity $\|P_A\|$, consider a simple case in which there are three worlds, W_1, W_2, and W_3, which we can represent with the three points $(0, 0, 1)$, $(0, 1, 0)$, and $(1, 0, 0)$ in three-dimensional Euclidean space. Then, the set of all probability distributions over these worlds is the set of all points which lie somewhere between them (see figure 2a). And for a probability distribution P, $\|P\|$ is just the distance from P to the origin, $(0, 0, 0)$. $\|P\|$ is greater the further away P is from the uniform distribution $(1/3, 1/3, 1/3)$, and the closer it is to the worlds (see figure 2b). In general, if there are N worlds, we can think of a probability P as a point in an N-dimensional Euclidean space. And $\|P\|$ will be the distance from that point to the origin.

8. See Jeffrey (1965) and Ahmed (2021), among others.
9. See Lewis (1981), Joyce (1999), and Sobel (1994), among others.
10. To explain why the standard deviation of our distribution over Sia's expected utility for A is going to depend upon $\|P_A\|$ in this way: let D_i be a random variable whose value is $D(W_i)$. And let $\mathbb{V}[X]$ be the variance of our probability distribution over the random variable X. Then, Sia's expected utility for A is just the weighted average $\sum_i P_A(W_i) \cdot D_i$, which is a linear combination of the random variables D_i (which we are taking to be independent and identically distributed). If σ^2 is the common variance of the random variables D_i, then $\mathbb{V}\left[\sum_i D_i \cdot P_A(W_i)\right] = \sigma^2 \cdot \sum_i P_A(W_i)^2 = \sigma^2 \cdot \|P_A\|^2$. So, the standard deviation of our probability distribution over X will be $\sigma \cdot \|P_A\|$.



(A) (B)

FIGURE 2. In figure 2a: the 2-simplex of probability distributions over three worlds is given by all points lying in the convex hull of $W_1 = (0, 0, 1)$, $W_2 = (0, 1, 0)$, and $W_3 = (1, 0, 0)$. In figure 2b, a 'heat map' for the magnitude of these probability distributions. Probabilities closer to individual worlds have greater magnitudes (represented with 'hotter' colours), and probabilities further away have small magnitudes (represented with 'cooler' colours).

$\|P_A\|$ is higher the more it 'points towards' some worlds over others. If P_A invests all its probability in a single world, then $\|P_A\| = 1$. If it spreads its probability between more worlds, then $\|P_A\|$ is lower. So, insofar as $\|P_A\|$ is lower, the consequences of A are more uncertain, and we can say that A 'leaves more up to chance'; insofar as they are higher, A's consequences are less uncertain, and we can say that A 'leaves less up to chance'.

As A leaves less up to chance, $\|P_A\|$ gets larger, so the standard deviation of our probability distribution over Sia's expected utility of A gets wider. And as this standard deviation gets wider, our probability that A maximises expected utility will get greater. (Of course, the probability that A *minimises* expected utility also gets greater.) In the appendix, I show that

Proposition 4. *If Sia's desires are sampled normally from the space of all possible desires, then the probability that her desires will rationalise choosing A in any grand world decision increases as A leaves less up to chance.*

For illustration: in the sample distributions shown in figure 1, Sia is about 38% likely to choose C (in orange), about 33% likely to choose B (in green), and about 29% likely to choose A (in blue).

This is a probabilistic version of the phenomenon we encountered with the decision **Certain and Uncertain Acts**. What proposition 4 tells us is that, if her desires are

sampled normally, then we should be somewhat more confident that Sia will choose acts that leave less to chance than we are that she'll choose acts that leave more to chance.

How much A leaves to chance isn't straightforwardly related to whether A protects Sia's goals or her life, nor whether A enhances Sia's cognitive abilities, technological capacities, or resources. If anything, enhancing her cognitive abilities seems to leave more to chance, insofar as Sia won't know which kinds of choices she'll make after her cognition has been enhanced. And if Sia is a highly novel and disruptive agent, then protecting her life may leave more to chance than ending it.

So while we've uncovered a *kind* of instrumental convergence, it does not appear to be the kind of convergence posited by the instrumental convergence thesis. I'll have more to say about this in section 5 below.

In some small world decisions like **Certain and Uncertain Acts**, we can know for sure that Sia won't make a particular choice. But in any grand world decision, we should retain some probability that she'll make any particular choice. For we can place a lower bound on the probability that any course of action, A, maximises expected utility for Sia. In the appendix, I show that

Proposition 5. *If Sia's desires are sampled randomly, then, in a grand world decision with n available acts, the probability that Sia chooses any given act is at least $1/2^{n-1}$.*

For instance, in a decision between three acts, there is at least a 25% probability that Sia will make any given choice. And, in a decision between four acts, the probability that she'll make any given choice is at least 12.5%. (Of course this is just a lower bound; it could easily be much higher.)

4. SEQUENTIAL DECISIONS

Above, we limited ourselves to one-off, non-sequential decisions. When it comes to sequential decisions—where a rational agent is charting a path through a multi-stage decision tree—matters are more complicated. One complication is that it's controversial how rational agents choose in sequential decisions. By way of explanation, consider the following sequential decision.

Pay for Ignorance: Sia wants nothing other than paperclips, and her desires are linear with paperclips; each new paperclip is just as good as the one that

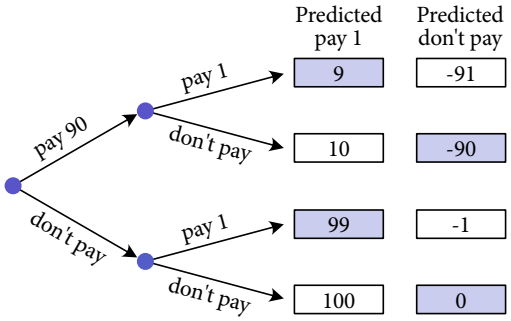


FIGURE 3. Pay for Ignorance. The boxes on the right indicate how many paperclips Sia will receive, depending upon which route through the decision tree she takes, and what our prediction was about her choice at stage 2. At stage 1, Sia is very confident that our prediction is accurate. So she is very confident that, whichever route through the tree she takes, she will receive the blue shaded number of paperclips at the end.

came before. At stage 1, Sia can either give us 90 paperclips or she can give us none. At stage 2, she can either give us 1 paperclip or she can give us none. Yesterday, we analysed her program and made a prediction about how she'd behave in this decision. If we predicted that she'd give us the single paperclip at stage 2, then we pre-awarded her 100 paperclips. If she doesn't pay us the 90 paperclips at stage 1, then we tell her which prediction we made. If she does pay us the 90 paperclips at stage 1, then we keep her in the dark about which prediction was made. Since our predictions are based on a thorough analysis of Sia's program, they are never wrong. Sia knows all of this.[11]

Sia's stage 2 decision matrix will look like this:

	Predicted Pay 1	Predicted don't pay
Pay 1	+99	−1
Don't Pay	+100	0

If she doesn't know which prediction we've made, then the decision Sia faces at stage 2 is just the famous 'Newcomb problem'.[12] Let's suppose that Sia is an evidential decision theorist, who chooses whichever act she would be most glad to learn that

11. Cf. Gibbard & Harper, 1978 and Wells, 2019.
12. Nozick, 1969

she'd chosen.[13] (By the way, nothing hinges on the choice of evidentialism here; we could make all the same points if we assumed instead that Sia was a causalist.) What Sia would be most glad to learn she'd chosen at stage 2 will depend upon what she knows. If she doesn't know which prediction we've made, then she'd be most glad to learn that she gives us the paperclip. Learning this would tell her that we pre-rewarded her 100 paperclips; whereas learning that she doesn't pay would tell her that we didn't. Since she'd rather have 99 paperclips than none, she'd rather learn that she pays us the paperclip. On the other hand, if Sia knows our prediction, then she'd be most glad to learn that she doesn't pay. For instance, if she knows that we predicted she wouldn't pay, then learning that she doesn't pay tells her that she's not getting any paperclips. On the other hand, learning that she *does* pay would tell her that she's only losing a paperclip.

Turn now to the *sequential* decision Pay for Ignorance. There are two main schools of thought about how Sia should decide at stage 1 of this decision. The orthodox view is often called *sophisticated*. It says that Sia should decide by doing a 'backwards induction', thinking first about what it would will maximise expected utilitty for her future self at stage 2, and then taking this for granted in her deliberation about what to do at stage 1. In particular, a sophisticated evidentialist Sia will notice that, if she knows what prediction was made, then not paying will maximise expected utility at stage 2. So a sophisticated evidentialist Sia will reason as follows: "the best possible path through the decision tree is to not pay the 90 paperclips at stage 1, and then pay the 1 paperclip at stage 2. If I do that, I'll likely end up with 99 paperclips. But I can't trust my future self to go along with that plan. Once she knows which prediction was made, it'll be rational for her to not pay. And if she doesn't pay, I'll likely end up with no paperclips. So: if I don't pay at stage 1, then I should expect to end up with zero paperclips. On the other hand, if I pay to remain ignorant of the prediction, then my future self will pay the 1 paperclip at stage 2. And so, she'll likely have been predicted to pay the 1 paperclip, and she'll likely have been pre-rewarded 100 paperclips. Minus the 91 I've already paid them, I should expect to end up with 9 paperclips if I pay at stage 1." So a sophisticated evidentialist Sia will pay the 90 paperclips at stage 1.

The less orthodox view is known as *resolute choice*. According to this view, at stage 1, Sia should decide which contingency plan is best, and then she should *stick to the plan*, even if sticking to the plan stops maximising expected utility later on. (A contingency

13. See Jeffrey, 1965 and Ahmed, 2021. When I talk about 'how glad Sia would be to learn that she'd chosen *A*', I mean: how well satisfied Sia would expect her desires to be, conditional on her choosing *A*.

plan will specify a collection of permissible acts for each situation Sia might find herself in.) A resolute evidentialist Sia will not pay the 90 paperclips at stage 1, and then pay the 1 paperclip at stage 2. Notice that a resolute evidentialist Sia would *not* pay the 1 paperclip if we just plopped her down in the stage 2 decision with the knowledge of which prediction we'd made, without giving her an earlier choice about whether to pay to avoid this knowledge. According to the resolute view, what it is rational for Sia to do at one stage of a sequential decision depends upon which contingency plans she's already committed herself to.

If Sia doesn't pay at either stage 1 or stage 2, then the sophisticated theory says that she behaved irrationally at stage 1, but rationally at stage 2. And the resolute theory says that she behaved rationally at stage 1 but irrationally at stage 2.[14]

There's another complication worth raising at this point. At stage 1, Sia may possess the ability to *bind* her future self to a certain course of action. Think of Ulysses binding himself to the mast.[15] With this ability, she would be able to instill in herself the intention to follow through on an initially selected contingency plan and deprive her future self of the ability to revise this intention, even if revising the intention maximises expected utility at that later time. If she possess this ability, then there will be no behavioural difference between her and a resolute chooser.

Suppose Sia follows a resolute theory of sequential decision-making, or that she has the ability to self-bind. Then, at stage 1, Sia will decide between contingency plans by comparing their expected utilities. And proposition 3 assures us that, if Sia's desires are sampled randomly, they are just as likely to make one contingency plan more rational than the other as they are to make the other contingency plan more rational than the one. Proposition 4 teaches us that, if her desires are sampled normally, then she'll be more likely to choose contingency plans which leave less to chance. And proposition 5 puts a lower bound on the probability that her desires will rationalise any particular contingency plan.

If Sia is a resolute chooser, or if she is able to self-bind, then there is a sense in which she is more likely to make some choices than others. If her desires are sampled normally, then—all else equal—a resolute or self-binding Sia will be more likely to make choices which have more choice points downstream of them. For instance, suppose that we have a large collection of prizes, $a, b, c, \ldots$, and each day, Sia has the choice to either

14. See Steele & Stefánsson (2015).
15. Cf. Arntzenius *et al.*, 2004 and Meacham, 2010.

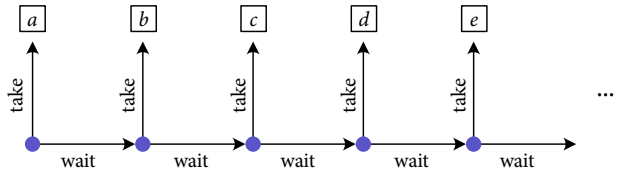


FIGURE 4. Take or Wait?

take that day's prize or else wait. As soon as she takes a prize, the game is over. (See figure 4.) Just to fix ideas, let's stipulate that the worlds over which Sia's desires are defined only include information about which choices Sia makes, and which prize she receives. Then, each contingency plan will leave as much to chance as every other. So, if her desires are sampled normally, she's incredibly likely to wait on day 1, since most of the contingency plans involve waiting on day 1, and only one involves taking a. The same thing holds in general. If she's a resolute chooser, or if she's able to self-bind, then (all else equal) Sia will be more likely to make choices that afford her more choices, and less likely to make choices that afford her fewer choices—just because there are more contingency plans which go on to face more choices, and fewer which go on to face fewer.

If Sia is a sophisticated chooser, then matters are more complicated. To introduce the complications, it'll be helpful to consider another of Bostrom's 'convergent' instrumental means: desire preservation. Will Sia be likely to preserve her desires? In general, this is going to depend upon the kind of decision she's facing—which ways she might change her desires, and how things might go differently, depending upon how the desires are changed. But in many simple cases, she *will* be more likely to keep her desires than she is to change them. Let me spend some time explaining why, since understanding this better will help us to think through what we should expect if Sia is a sophisticated chooser.

For illustration, consider a simple sequential decision like the one shown in figure 5. Sia begins at the blue node in the center and decides whether to change her desires or not. Suppose that, if she changes her desires, then she will prefer a to b. Clearly, if we model this as a 'small world' decision, where Sia only cares about whether she ends up with a or b, and cares not at all about what her desires are, then she will be more likely to keep her desires. For, if her desires are sampled randomly, then she will be just as likely to prefer a to b as she will be to prefer b to a. If she prefers a to b, then she'll be indifferent between keeping her desires and changing them. And if she prefers b to a, then she'll prefer to keep her desires.

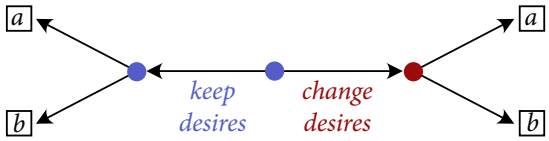


FIGURE 5. Sia can either choose between a and b with her current (blue) desires, or she can adopt new (red) desires, which will lead her to choose a.

Even if we model Sia's decision as a 'grand world' decision, and allow that she might care about whether her desires are changed, she will still be more likely to keep her desires in this decision. To appreciate why, let's model her decision with four worlds: 1) the world where she *k*eeps her desires and gets a, W_{ka}, 2) the world where she *c*hanges her desires and gets a, W_{ca}, 3) W_{kb}, and 4) W_{cb} (with the natural interpretation). If she changes her desires, she will certainly end up at world W_{ca}, since the changed desires will prefer a to b. So, in deciding whether to change her desires or not, Sia will be comparing the degree to which she desires getting a after changing her desires, $D(W_{ca})$, to whichever of $D(W_{ka})$ and $D(W_{kb})$ is greatest—for, if $D(W_{ka}) > D(W_{kb})$, then Sia would choose a and end up at world W_{ka}; and, if $D(W_{kb}) > D(W_{ka})$, then Sia would choose b and end up at world W_{kb}. But, conditional on $D(W_{ka})$ being larger than $D(W_{kb})$, it is more likely to be larger than $D(W_{ca})$, too. And, similarly, conditional on $D(W_{kb})$ being larger than $D(W_{ka})$, it is more likely to be larger than $D(W_{ca})$, too. So, overall, Sia will be more than 50% likely to keep her desires in this decision.

So it looks as though desire preservation *will* be a 'convergent' instrumental means in this decision, at least in the sense that randomly selected desires are more likely to rationalise desire preservation than they are to rationalise desire change. (Of course, just because this is true in *this* decision, it doesn't mean that it'll be true in *every* decision where Sia is deciding whether to modify her desires; but the mechanism which makes desire preservation more likely here seems general enough that we should expect it to carry over to many other decisions, too.)

Moreover, this same mechanism should lead us to expect a sophisticated Sia to make choices which allow for more choices later on. Return again to the sequential decision from figure 4. We saw above that, all else equal, a resolute Sia would be more likely to wait at stage 1 than she was to take a, for the simple reason that most of the contingency plans wait at stage 1, and only one takes a. If Sia is a sophisticated chooser, it will also be true that she's more likely to wait at stage 1—but it's for a different reason. The reason

a sophisticated Sia is more likely to wait at stage 1 is that, in making that decision, she's comparing the degree to which she desires the world where she takes a, $D(W_a)$, to the *maximum* of the degree to which she desires all other worlds, $D(W_b), D(W_c), \ldots$. And even though our probability that $D(W_a)$ is greater than $D(W_b)$ will be 50%, and the probability that $D(W_a)$ is greater than $D(W_c)$ is 50%, and so on, the probability that $D(W_a)$ is greater than *all* of $W_b, W_c, \ldots$ is far less than 50%. So—all else equal—a sophisticated Sia whose desires are sampled randomly will be biased towards choices which allow for more choices later on.

5. DISCUSSION

Let's summarise our findings. We've identified three kinds of 'convergent' instrumental means—which is to say, we've identified three ways in which Sia's choices may be predicted with better than even odds, even if her desires are sampled randomly.

In the first place, she's somewhat more likely to favour acts which leave less to chance. As I mentioned above, I don't see any reason to think that resource acquisition, technological advancement, cognitive enhancement, and so on, will in general leave less up to chance. Insofar as the results of technological and cognitive enhancement are unpredictable in advance, this gives us some reason to think that Sia is *less* likely to pursue cognitive and technological enhancement. So I don't think this bias is relevant to the kinds of 'convergent' instrumental means which are Bostrom's focus.

Should a bias against leaving things up to chance lead us to think that existential catastrophe is the more likely outcome of creating a superintelligent agent like Sia? This is far from clear. We might think that a world without humans leaves less to chance, so that we should think Sia is more likely to take steps to eliminate humans. But we should be cautious about this inference. It's unclear that a future without humanity would be more predictable. And even if the future course of history is more predictable *after* humans are eliminated, that doesn't mean that the act of eliminating humans leaves less to chance, in the relevant sense. It might be that the contingency plan which results in human extinction depends sensitively upon humanity's response; the unpredictability of this response could easily mean that that contingency plan leaves more to chance than the alternatives. At the least, if this bias means that human extinction is a somewhat more likely consequence of creating superintelligent machines, more needs to be said about why.

It's also worth emphasising that this bias only tells us that Sia is *more* likely to perform acts that leave less to chance she is to perform acts which leave more to chance. It doesn't tell us that she is *overall likely* to perform any particular act. Ask me to pick a number between one and one billion, and I'm more likely to select 500,000,000 than I am to select 456,034—humans have a bias towards round numbers. But that doesn't mean I'm at all likely to select 500,000,000. So even if this tells us that Sia is somewhat *more* likely to exterminate humanity than she is to dedicate herself to dancing the Macarena, or gardening, or what-have-you, that doesn't mean that she's particularly likely to exterminate humanity.

In the second place, we found that, in sequential decisions, Sia is more likely to make choices which allow for more choices later on. This turned out to be true whether Sia is a 'resolute' chooser or a 'sophisticated' chooser. (Though it's true for different reasons in the two cases, and there's no reason to think that the effect size is going to be the same.) Does this mean she's more likely to bring about human extinction? It's unclear. We might think that humans constitute a potential threat to Sia's continued existence, so that futures without humans are futures with more choices for Sia to make. So she's somewhat more likely to take steps to eliminate humans. (Again, we should remind ourselves that being *more* likely isn't the same thing as being *likely*.) I think we need to tread lightly, for two reasons. In the first place, futures without humanity might be futures which involve very few choices—other deliberative agents tend to force more decisions. So contingency plans which involve human extinction may involve comparatively fewer choicepoints than contingency plans which keep humans around. In the second place, Sia is biased towards choices which *allow* for more choices—but this isn't the same thing as being biased towards choices which *guarantee* more choices. Consider a resolute Sia who is equally likely to choose any contingency plan, and consider the following sequential decision. At stage 1, Sia can either take a 'safe' option which will certainly keep her alive or she can play Russian roulette, which has a 1-in-6 probability of killing her. If she takes the 'safe' option, the game ends. If she plays Russian roulette and survives, then she'll once again be given a choice to either take a 'safe' option of definitely staying alive or else play Russian roulette. And so on. Whenever she survives a game of Russian roulette, she's again given the same choice. All else equal, if her desires are sampled normally, a resolute Sia will be much more likely to play Russian roulette at stage 1 than she will be to take the 'safe' option. (The same is true if Sia is a sophisticated chooser, though a sophisticated Sia is more likely to take the safe option at stage 1 than the resolute Sia.) The lesson is this: a bias towards choices with more potential downstream choices isn't a bias towards self-preservation. Whether she's likely to try to preserve her life is

going to sensitively depend upon the features of her decision situation. Again, much more needs to be said to substantiate the idea that this bias makes it more likely that Sia will attempt to exterminate humanity.

Finally, we found that in some decisions, Sia is more likely to act so as to preserve her own desires. Desire preservation is the most plausible item on Bostrom's list of 'convergent' instrumental means. While there are of course many situations in which it is instrumentally rational to change your desires, desire preservation is more likely than desire change in a great many decisions. (Again, I haven't spread probabilities over the potential decisions Sia might face, so I'm not in a position to say anything stronger than this.)

Should this lead us to think that existential catastrophe is the most likely outcome of a superintelligent agent like Sia? Again, it is far from clear. Insofar as Sia is likely to preserve her desires, she may be unlikely to allow us to shut her down in order to change those desires.[16] We might think that this makes it more likely that she will take steps to eliminate humanity, since humans constitute a persistent threat to the preservation of her desires. (Again, we should be careful to distinguish Sia being *more* likely to exterminate humanity from her being *likely* to exterminate humanity.) Again, I think this is far from clear. Even if humans constitute a threat to the satisfaction of Sia's desires in *some* ways, they may be conducive towards her desires in others, depending upon what those desires are. In order to think about what Sia is likely to do with randomly selected desires, we need to think more carefully about the particulars of the decision she's facing. It's not clear that the bias towards desire preservation is going to overpower every other source of bias in the more complex real-world decision Sia would actually face. In any case, as with the other 'convergent' instrumental means, more needs to be said about the extent to which they indicate that Sia is an existential threat to humanity.

In sum, the instrumental convergence thesis contains some grains of truth. A superintelligence with randomly sampled desires will be biased towards certain kinds of choices over others. These include choices which leave less up to chance, choices which allow for more choices, and choices which preserve desires. Nonetheless, the thesis is mostly false. For most of the convergent means on the list, there are decisions in which a superintelligence with random desires is no more likely to pursue them than not. The grains of truth in the thesis may give us reason to worry about the

16. See the 'shutdown problem' from Soares *et al.* (2015).

existential threat posed by machine superintelligence. But they do not on their own support Bostrom's stronger contention that "the default outcome of the creation of machine superintelligence is existential catastrophe". Like most of life's dangers, the dangers posed by artificial intelligence are not easily identified from the armchair. If we want to understand the dangers posed by artificial superintelligence, we will have to do more careful empirical work investigating what kinds of desires future AI systems are likely to have (or, indeed, whether they are likely to have desires at all).

APPENDIX

Let $\mathcal{W}$ be the collection of ways the world might be, for all Sia is in a position to know. I'll assume that $\mathcal{W}$ is finite, with cardinality N. Fix some enumeration of the worlds in $\mathcal{W}$, $W_1, W_2, \ldots, W_N$. I assume that there is a fixed collection of available acts, $\mathcal{A}$, between which Sia must choose. We can represent each of the possible desires Sia might hold with a function $D : \mathcal{W} \to \mathbb{R}$. The interpretation is that $D(W)$ measures how well satisfied Sia's desires are at the world W.

I'll suppose that, in order for us to say which of the acts in $\mathcal{A}$ are more or less rational than which others, we need to be given one more ingredient: we'll need, for each $A \in \mathcal{A}$, a *suppositional probability function* P_A. Since $\mathcal{W}$ is finite, we can take each of these probability functions to be defined over the powerset $\mathcal{P}(\mathcal{W})$.

For each $A \in \mathcal{A}$, we can use the suppositional probability function P_A and the desire function D to calculate Sia's *expected utility* for A.

Definition 1. *The expected utility of $A \in \mathcal{A}$, relative to the desires D, is the value which the suppositional probability function P_A expects D to take on.*

$$\mathbb{E}_{P_A}[D] = \sum_{W \in \mathcal{W}} D(W) \cdot P_A(W)$$

Both causal and evidential decision theory say that an act A is more rational than another, B, iff A's expected utility is greater than B's. They disagree over how to understand the probability functions P_A. Evidentialists say that $P_A(X)$ is the conditional probability function $P(X \mid A)$.[17] Causalists say that $P_A(X)$ is the probability function P *imaged* on the performance of A. They take for granted an *imaging* function, $I : \mathcal{A} \times \mathcal{W} \to (\mathcal{P}(\mathcal{W}) \to [0, 1])$, from pairs of acts and worlds to probability distributions. The interpretation is that $I(A, W)(X)$ is how likely it is that the propositions X would be true, were you to choose A at the world W. Then, causalists say that $P_A(X) = \sum_{W \in \mathcal{W}} I(A, W)(X) \cdot P(W)$.[18] Other decision theories, like the functional decision theory from Yudkowsky & Soares, 2018 and Soares & Levinstein, 2020, will also take this form, though they will understand the suppositional probability distributions differently.

Proposition 1. *For any $A, B \in \mathcal{A}$, if $P_A \neq P_B$, then there are infinitely many desires D such that the expected utility of A is greater than the expected utility of B, relative to D.*

Proof. If $P_A \neq P_B$, then there is some set of worlds $X \subseteq \mathcal{W}$ such that $P_A(X) > P_B(X)$. Select any two numbers $x, y \in \mathbb{R}$ such that $x > y$ and consider a desire function D such that, for all $W \in X$, $D(W) = x$, and for all $W \notin X$, $D(W) = y$. Then, the expected utility of A, relative

17. See Jeffrey, 1965 and Ahmed, 2021.
18. See Lewis, 1981, Sobel, 1994, and Joyce, 1999.

to D, will be $xP_A(X) + y[1 - P_A(X)]$ and the expected utility of B, relative to D, will be $xP_B(X) + y[1 - P_B(X)]$. Since $P_A(X) > P_B(X)$ and $x > y$, the expected utility of A will exceed the expected utility of B, relative to these desires. Since there are infinitely many choices of x and y such that $x > y$, there are infinitely many such desires. □

Note that, by just taking another instance of the proposition in which we exchange A and B, we also get that there are infinitely many desires such that the expected utility of B exceeds the expected utility of A, relative to those desires. Note also that, on the causalist's understanding, $P_A \neq P_B$ iff the expected consequences of A are different from the expected consequences of B.

Finally, note that proposition 1 doesn't depend upon whether we are considering a 'small world' or 'grand world' decision. If it is a grand world decision, so that P is defined over propositions about which choice Sia makes, then $P_A(W) > 0$ implies that $P_B(W) = 0$, for every $B \neq A$. Then,

Proposition 2. *In any grand world decision, and any available act in the decision, A, there are infinitely many desires which make A more rational than every other alternative.*

Proof. Select any two numbers $x, y \in \mathbb{R}$ such that $x > y$ and consider a desire function D such that, for all $W \in A$, $D(A) = x$, and for all $W \notin A$, $D(A) = y$. Then, the expected utility of A will be $\sum_W P_A(W) \cdot D(W) = \sum_{W \in A} P_A(W) \cdot x = x$, and the expected utility of every other act, B, will be $\sum_W P_B(W) \cdot D(W) = \sum_{W \in B} P_B(W) \cdot y = y$. So the expected utility of A will exceed the expected utility of every other act. Since there are infinitely many choices of x and y such that $x > y$, there are infinitely many such desires. □

Let $\mathfrak{D}$ be the set of all desires Sia could have, $\mathfrak{D} = \{D : \mathscr{W} \to \mathbb{R}\}$. We can define a probability distribution, Q, over a σ-field of subsets of $\mathfrak{D}$, $\mathfrak{F} \subseteq \mathscr{P}(\mathfrak{D})$. That is, $\mathfrak{F}$ is a set of propositions about Sia's desires such that (i) $\mathfrak{D} \in \mathfrak{F}$, (ii) $X^c \in \mathfrak{F}$ whenever $X \in \mathfrak{F}$, and (iii) $\bigcup_{i=1}^{\infty} X_i \in \mathfrak{F}$ whenever $X_1, X_2, \cdots \in \mathfrak{F}$. The interpretation is that $Q(Y)$ is *our* probability that Sia's desires fall somewhere within the set $Y \subseteq \mathfrak{D}$.

I'll suppose that our probability distribution Q satisfies the following four conditions. To explain the first condition: let 'D_i' be a random variable which takes a desire function $D \in \mathfrak{D}$ to the real number $D(W_i)$. Then, I'll suppose that, for every $x \in \mathbb{R}$, the proposition $D_i \leqslant x$ is included in $\mathfrak{F}$. This way, the 'cumulative distribution function' $Q(D_i \leqslant x)$ will be well-defined, for every $i \in \{1 \ldots N\}$ and every $x \in \mathbb{R}$. I'll also suppose that this function is absolutely continuous.

(1) For every $i \in \{1 \ldots N\}$ and every $x \in \mathbb{R}$, $D_i \leqslant x \in \mathfrak{F}$, and $Q(D_i \leqslant x)$ is absolutely continuous.

If this first condition is satisfied, then for each world W_i, we can define a probability density function $q_i(x) = (d/dx)Q(D_i \leqslant x)$.

Secondly, I'll assume that, for any two worlds $W_i, W_j \in \mathscr{W}$, we don't have any reason to think that W_i will better satisfy Sia's desires than W_j will. So, for any $i, j \in \{1 \dots N\}$, our expectation of the value of D_i should equal our expectation of the value of D_j.

(2) For any $i, j \in \{1 \dots N\}$,

$$\mathbb{E}_Q[D_i] = \int_{-\infty}^{\infty} x \cdot q_i(x)\ dx = \int_{-\infty}^{\infty} x \cdot q_j(x)\ dx = \mathbb{E}_Q[D_j]$$

I'll call this common expectation 'μ'. μ is our expectation of the degree to which Sia's desires will be satisfied at any particular world.

Thirdly, I'll assume that we've no more reason to think that Sia's desires are going to be satisfied to degree $\mu + x$ than we have to think that her desires are going to be satisfied to degree $\mu - x$. That is: I'll assume that each probability density function q_i is *symmetric* around the mean μ.

(3) For each $i \in \{1 \dots N\}$ and each $x \in \mathbb{R}$, $q_i(\mu + x) = q_i(\mu - x)$.

And finally, I'll assume that learning how well satisfied Sia's desires are at some worlds doesn't tell us how well satisfied her desires are at other worlds.

(4) The random variables $D_1, D_2, \dots D_N$ are mutually independent.

If these four conditions are satisfied, then I'll say that Sia's desires are 'sampled randomly'.

Proposition 3. *If Sia's desires are sampled randomly, then, for any two acts $A, B \in \mathscr{A}$, the probability that Sia's desires make A more rational than B is equal to the probability that Sia's desires make B more rational than A.*

Proof. First, notice that whether A is more or less rational than B does not change if we replace Sia's desires with a positive affine transformation of those desires. That is, if $\hat{D}$ is a positive affine transformation of D, then $\mathbb{E}_{P_A}[D] \geqslant \mathbb{E}_{P_B}[D]$ iff $\mathbb{E}_{P_A}[\hat{D}] \geqslant \mathbb{E}_{P_B}[\hat{D}]$. So, if Sia's desires are D, we can let $\hat{D} =_{df} D - \mu$, where μ is our expectation of the degree to which Sia's desires will be satisfied at any world (measured in the units of D). Having performed this transformation, our expectation of the degree to which Sia's desires will be satisfied, in the units of $\hat{D}$, will be zero. For $\mathbb{E}_Q[\hat{D}_i] = \mathbb{E}_Q[D_i - \mu] = \mathbb{E}_Q[D_i] - \mu = \mu - \mu = 0$. I will use the units of the 'shifted' scale $\hat{D}$ for the remainder of the proof.

Next, consider the random variable $Z = \mathbb{E}_{P_A}[\hat{D}] - \mathbb{E}_{P_B}[\hat{D}]$. If Z is positive, then Sia's desires make A more rational than B. If Z is negative, then her desires make B more rational than A. Zero, and her desires make A and B equally rational. Note that Z is a linear combination of

the random variables $\hat{D}_i$

$$\begin{aligned} Z &= \sum_{i=1}^{N} \hat{D}_i \cdot P_A(W_i) \; - \; \sum_{i=1}^{N} \hat{D}_i \cdot P_B(W_i) \\ &= \sum_{i=1}^{N} [P_A(W_i) - P_B(W_i)] \cdot \hat{D}_i \end{aligned}$$

Let $c_i = P_A(W_i) - P_B(W_i)$. Then, $Z = \sum_{i=1}^{N} c_i \cdot \hat{D}_i$.

Let $\varphi_i(t) =_{df} \mathbb{E}_Q[e^{t\hat{D}_i\sqrt{-1}}]$ be the characteristic function of the random variable $\hat{D}_i$. The characteristic function of a random variable is real-valued iff that variable is probabilistically symmetric about the origin—that is, for each x, the probability that the variable takes on a value greater than x is equal to the probability that it takes on a value less than $-x$ (see Billingsley 1986, problem 26.2.) By assumption, each random variable D_i is symmetric about their common mean μ, so the 'shifted' variables $\hat{D}_i$ are symmetric about the origin. So each $\varphi_i(t)$ is real-valued. Given any N mutually independent random variables $V_1, V_2, \ldots V_N$ with characteristic functions $\varphi_{V_1}(t), \varphi_{V_2}(t), \ldots, \varphi_{V_N}(t)$, the characteristic function for their linear combination $U = \sum_{i=1}^{N} c_i \cdot V_i$ is $\varphi_U(t) = \prod_{i=1}^{N} \varphi_{V_i}(c_i t)$.[19] Since by assumption the random variables $\hat{D}_i$ are mutually independent, and since $Z = \mathbb{E}_{P_A}[\hat{D}] - \mathbb{E}_{P_B}[\hat{D}]$ is a linear combination of the $\hat{D}_i$, $Z = \sum_{i=1}^{N} c_i \cdot \hat{D}_i$, the characteristic function for Z, $\varphi_Z(t)$, is $\prod_{i=1}^{N} \varphi_i(c_i t)$. The product of N real-valued functions is real-valued. So φ_Z is real-valued. So Z is also probabilistically symmetric about the origin. So $Q(Z > 0) = Q(Z < 0)$. So Sia's desires are just as likely to make A more rational than B as they are to make B more rational than A. □

Let's consider an additional assumption about our probability distribution, Q. If our probabilities are distributed independently and identically for each random variable D_i, and, moreover, this distribution is a Gaussian or normal distribution $D_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma)$, then I'll say that Sia's desires are sampled *normally* from the space of all possible desires.

Proposition 4. *If Sia's desires are sampled normally from the space of all possible desires, then, in any grand world decision, the probability that her desires will rationalise choosing A increases with* $\|P_A\| = \left(\sum_{W \in \mathcal{W}} P_A(W)^2\right)^{1/2}$.

Proof. As explained in the proof of proposition 3, we can re-scale Sia's desires by taking $\hat{D}_i = D_i - \mu$. Then, we will have the variables $\hat{D}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma)$. Since this is a grand world decision, then there is no world $W \in \mathcal{W}$ such that $P_A(W)$ and $P_B(W)$ are both positive, for any

19. To appreciate this, note

$$\varphi_U(t) = \mathbb{E}\left[e^{t\sqrt{-1}\sum_i c_i V_i}\right] = \mathbb{E}\left[\prod_i e^{c_i t\sqrt{-1}V_i}\right] = \prod_i \mathbb{E}[e^{c_i t\sqrt{-1}V_i}] = \prod_i \varphi_{V_i}(c_i t)$$

The third equality follows from independence.

$A \neq B$. So Sia's expected utilities for acts are independent (since the $\hat{D}_i$s are independent). Let Y_i be Sia's expected utility for A_i. Then, $Y_i = \sum_j \hat{D}_j \cdot P_{A_i}(W_j)$, which is a linear combination of the variables $\hat{D}_j$. Since each $\hat{D}_j$ is an iid normal random variable with mean 0 and variance σ^2, their linear combination will be a normal variable with mean 0 and variance $\sigma^2 \cdot \sum_j P_{A_i}(W_j)^2 = \sigma^2 \|P_{A_i}\|^2$. So $X_i = Y_i/(\sigma \cdot \|P_{A_i}\|)$ will have a standard normal distribution. That is, if q_i is our probability density function for the random variable X_i, then $q_i(x) = \phi(x)$, where $\phi(x)$ is the standard normal distribution,

$$\phi(x) =_{df} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Since the X_i are independent, their joint probability density is just the product of the marginal densities,

$$q(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{n} \phi(x_i)$$

Sia's expected utility for A_i is greater than her expected utility for A_j exactly if $Y_i > Y_j$, which is so exactly if $(\sigma \cdot \|P_{A_i}\|)X_i > (\sigma \cdot \|P_{A_j}\|)X_j$, which is so exactly if $X_j < (\|P_{A_i}\|/\|P_{A_j}\|)X_i$. Without loss of generality, consider the probability that A_1 maximises expected utility. That probability is given by

$$\begin{aligned}
&\int_{-\infty}^{\infty} \int_{-\infty}^{(\|P_{A_1}\|/\|P_{A_2}\|)x_1} \cdots \int_{-\infty}^{(\|P_{A_1}\|/\|P_{A_n}\|)x_1} q(x_1, x_2, \ldots, x_n)\ dx_n \ldots dx_2 dx_1 \\
= &\int_{-\infty}^{\infty} \int_{-\infty}^{(\|P_{A_1}\|/\|P_{A_2}\|)x_1} \cdots \int_{-\infty}^{(\|P_{A_1}\|/\|P_{A_n}\|)x_1} \prod_{i=1}^{n} \phi(x_i)\ dx_n \ldots dx_2 dx_1 \\
= &\int_{-\infty}^{\infty} \phi(x_1) \int_{-\infty}^{(\|P_{A_1}\|/\|P_{A_2}\|)x_1} \phi(x_2) \cdots \int_{-\infty}^{(\|P_{A_1}\|/\|P_{A_n}\|)x_1} \phi(x_n)\ dx_n \ldots dx_2 dx_1
\end{aligned}$$

In general,

$$\int_{-\infty}^{c} \phi(x)\ dx = \Phi(c) \tag{1}$$

where Φ is the cumulative density function for the standard normal distribution. Then, the probability that A_1 maximises expected utility is given by

$$\int_{-\infty}^{\infty} \phi(x_1) \cdot \prod_{i=2}^{N} \Phi(x_1 \|P_{A_1}\|/\|P_{A_i}\|)$$

$\Phi(x)$ is an increasing function of x. So, as $\|P_{A_1}\|$ gets larger (if everything else is held fixed), each of the factors in the product in (1) will become larger. Since $\phi(x)$ is non-negative, this will mean that the value of the integral will become larger. So the probability that A_1 maximises expected utility will be larger. A_1 was arbitrary, so what goes for it goes for every other option; in general, for any $A \in \mathcal{A}$, as $\|P_A\|$ increases (with everything else held fixed), the probability that A maximises expected utility increases. □

Proposition 5. *If Sia's desires are sampled randomly, then, in a grand world decision with n available acts, the probability that she chooses any given act is at least* $1/2^{n-1}$.

Proof. Let $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$, and take any $A_i \in \mathcal{A}$. Without loss of generality, let it be A_1. Since $Q(D_i \leqslant x)$ is absolutely continuous, the probability that Sia's expected utility for A_1 is equal to her expected utility for A_j ($j \neq i$) is zero. So we can ignore this possibility when calculating the probability that A is uniquely rational. If we write '$A_1 \succ A_i$' for '$\mathbb{E}_{P_{A_1}}[D] > \mathbb{E}_{P_{A_i}}[D]$', the probability that A_1 is uniquely rational is

$$Q(A_1 \succ A_2 \wedge A_1 \succ A_3 \wedge \cdots \wedge A_1 \succ A_n) = \left[\prod_{i=2}^{n-1} Q\left(A_1 \succ A_i \;\middle|\; \bigwedge_{j=i+1}^{n} A_1 \succ A_j\right)\right] \cdot Q(A_1 \succ A_n)$$

Because $P_{A_1}(A_1) = 1$ and $P_{A_i}(A_i) = 1$, there is no world W such that both P_{A_1} and P_{A_i} give W positive probability (for any $i > 1$). So $\mathbb{E}_{P_{A_1}}[D] = \sum_j P_{A_1}(W_j) \cdot D_j$ and $\mathbb{E}_{P_{A_i}}[D] = \sum_j P_{A_i}(W_j) \cdot D_j$ will be independent. So, for each $i \geqslant 2$, $Q\left(A_1 \succ A_i \;\middle|\; \bigwedge_{j=i+1}^{n} A_1 \succ A_j\right) \geqslant Q(A_1 \succ A_i)$.

By proposition 3, $Q(A_1 \succ A_i) = Q(A_i \succ A_1) = 1/2$. So

$$\prod_{i=2}^{n-1} Q\left(A_1 \succ A_i \;\middle|\; \bigwedge_{j=i+1}^{n} A_1 \succ A_j\right) \cdot Q(A_1 \succ A_n) \geqslant (1/2)^{n-1}$$

□

REFERENCES

Ahmed, Arif. 2021. *Evidential Decision Theory*. Cambridge: Cambridge University Press.

Arntzenius, Frank, Elga, Adam, & Hawthorne, John. 2004. "Bayesianism, Infinite Decisions, and Binding." In *Mind*, **113** (450): 251–283. doi:10.1093/mind/113.450.251.

Benson-Tilsen, Tsvi & Soares, Nate. 2015. "Formalizing Convergent Instrumental Goals." `https://intelligence.org/files/FormalizingConvergentGoals.pdf`.

Billingsley, Patrick. 1986. *Probability and Measure*. John Wiley and Sons, second edition.

Bostrom, Nick. 2014. *Superintelligence: paths, dangers, strategies*. Oxford: Oxford University Press.

Carlsmith, Joseph. ms. "Is Power-seeking AI an Existential Risk?" `https://arxiv.org/abs/2206.13353`.

Chivers, Tom. 2019. *The AI does not hate you: superintelligence, rationality, and the race to save the world*. Weidenfeld & Nicolson.

Gibbard, Allan & Harper, William L. 1978. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, edited by A. Hooker, J. J. Leach, & E. F. McClennen, D. Reidel, 125–162.

Jeffrey, Richard. 1965. *The Logic of Decision*. New York: McGraw-Hill.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Lewis, David K. 1981. "Causal Decision Theory." In *Australasian Journal of Philosophy*, **59** (1): 5–30.

Meacham, Christopher J. G. 2010. "Binding and its Consequences." In *Philosophical Studies*, **149** (1): 49–71. doi:10.1007/s11098-010-9539-7.

Nozick, Robert. 1969. "Newcomb?s Problem and Two Principles of Choice." In *Essays in Honor of Carl G. Hempel*, edited by Nicholas Rescher, Reidel, 114–146.

Omohundro, Stephen. 2008*a*. "The basic AI drives." In *Proceedings of the First Conference on Artificial General Intelligence*, 83–49.

Omohundro, Stephen M. 2008*b*. "The nature of self-improving artificial intelligence." `https://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf`.

Savage, Leonard J. 1954. *The Foundations of Statistics*. Wiley Publications in Statistics.

Soares, Nate, Fallenstein, Benja, Armstrong, Stuart, & Yudkowsky, Eliezer. 2015. "Corrigibility." In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Soares, Nate & Levinstein, Benjamin Anders. 2020. "Cheating Death in Damascus." In *The Journal of Philosophy*, **117** (5): 237–266. doi:10.5840/jphil2020117516.

Sobel, Jordan Howard. 1994. *Taking Chances: Essays on Rational Choice*. Cambridge: Cambridge University Press.

Steele, Katie & Stefánsson, H. Orri. 2015. "Decision Theory." In *Stanford Encyclopedia of Philosophy*.

Turner, Alexander Matt, Smith, Logan, Shah, Rohin, Critch, Andrew, & Tadepalli, Prasad. 2021. "Optimal Policies Tend to Seek Power." In *NeurIPS*.

Wells, Ian. 2019. "Equal Opportunity and Newcomb?s Problem." In *Mind*, **128** (510): 429–457. doi:10.1093/mind/fzx018.

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom & Milan M. Ćirković, New York: Oxford University Press, 308–345.

Yudkowsky, Eliezer. 2023. "Pausing AI Developments Isn't Enough. We Need to Shut it All Down." In *Time magazine*.

Yudkowsky, Eliezer & Soares, Nate. 2018. "Functional Decision Theory: A New Theory of Instrumental Rationality." `https://arxiv.org/abs/1710.05060`.