

THE PROBLEM OF SIMULATED EVIL

Abstract: According to the simulation hypothesis, the world we live in is a computer simulation. According to longtermism, we should aim to bring about the best possible future. In this paper, I argue that there is a tension between the two: insofar as we have reason to think we are living in a computer simulation, we have reason to think the longtermist project will fail (or has already failed). I make my case by developing a novel version of the problem of evil. It says that there exist certain forms of evil in our world that longtermist philosophers have argued we should try to eradicate. But then, the existence of those evils give us reason to think our world is not a computer simulation run by a being aligned with longtermist values. In short, the more that longtermists think we could make the world a better place in the future, the more troubling they should find it that the world is not already a better place, given the assumption that we are living in a computer simulation.

According to the *simulation hypothesis*, the world we live in is a computer simulation. According to *longtermism*, we should aim to bring about the best possible future. In what follows, I argue that there is a tension between the two: insofar as we have reason to think we are living in a computer simulation, we have reason to think the longtermist project is doomed. Here is a formalization of my argument.

(P1): There exist certain forms of evil in our world.

(P2): If these forms of evil exist in our world, then our world is probably not a computer simulation run by longtermists.

(P3): If our world is probably not a computer simulation run by longtermists, then either the longtermist project is probably doomed or else the simulation hypothesis is false.

(C): Either the longtermist project is probably doomed or else the simulation hypothesis is false.

The argument is plainly valid and so the question will be whether its premises are true.¹

After an introductory section where I briefly review the simulation hypothesis and longtermism (§0), the paper is organized by devoting a different section to each premise (§§1-3), followed by a concluding discussion (§4).

¹ The logical form of the argument is: (P1) P, (P2) $P \rightarrow Q$, (3) $Q \rightarrow R \vee S$, (4) $R \vee S$. The term “probably” occurs in the consequent of (P2) and the antecedent of (P3), but the argument is still deductive. It is a valid deductive argument that is partly about probabilistic claims.

One reason the argument is of interest is that there are philosophers and others who are attracted both to the simulation hypothesis and to the longtermist project. The leading example is Nick Bostrom (see for instance Bostrom, 2003a and 2003b), and so in what follows I discuss Bostrom's views at length. Bostrom's views have been influential, and not just within the academy. Citing Bostrom as an inspiration, Elon Musk has both publicly defended the simulation hypothesis and also endorsed longtermism as a "close match" for his own philosophy (Godfrey-Smith, 2022). If my argument is successful, it will show that there is a tension in the beliefs of Bostrom, Musk, and others who follow their lead.

Another reason my argument is of interest is that it sets out a novel version of the *problem of evil*. The simulation hypothesis is sometimes compared to the theist view that our world was created by God, with some authors even suggesting it should be regarded as a special case of theism (for discussion, see Chalmers, 2022: Chapter 7). But the most familiar version of the problem of evil is generated from the assumptions that God is omniscient, omnipotent, and omnibenevolent, while in contrast there is no presumption that the (possibly human) creator of a computer simulation would need to possess any of those traits. As David Chalmers (2022: 135) puts it, "the problem of evil is no obstacle to a naturalistic simulator god... [since] a simulator need not be all good." I will be arguing that Chalmers is wrong about this. As I will explain, a version of the problem can arise even for a naturalistic, morally imperfect simulator. At various points along the way, we will see (sometimes surprising) ways in which points first familiar from the traditional problem of evil bear on my argument here.²

² Aside from Chalmers, there are discussions of how the simulation hypothesis and the problem of evil interact in Dainton (2002 and 2020), Johnson (2011), Shea (2017) and Crummett (2020). However, these

o. Preliminaries

In this section I briefly review the simulation argument and longtermism. On Bostrom's (2003a) canonical formulation of the simulation argument, there is good reason to think that at some point in the course of history, human beings will reach a "posthuman" stage in which they will be able to run computer simulations of entire worlds. This reason is largely empirical. Look at how much progress we have already made in developing sophisticated computer simulations of this and that—simulations of the weather, presidential elections, Super Bowls, and more—and then extrapolate into the future. At some point, whole-world simulations are likely to happen, says Bostrom.

Next, suppose that mental properties are *substrate-independent* (Bostrom, 2003a; Chalmers, 2022: 93), where this means that a computer running the right program could instantiate the very same mental properties that human beings do despite their physical differences from us. Substrate-independence is a fairly widely accepted doctrine in the philosophy of mind, familiar especially from functionalist views. At any rate, it is something I am willing to grant Bostrom to allow the simulation argument to get off the ground.

Given this setup, there is reason to think that over the course of the time there will be the opportunity to run an extraordinary number of computer simulations of whole worlds that are relevantly like ours in that they include intelligent, conscious beings much like us. But if so then what reason do we have to think that our world is ground-level, unsimulated reality, as opposed to being one of these many simulated

other authors do not consider the version of the problem I discuss in the present paper, where the existence of evil is taken to pose a problem for certain naturalistic simulator hypotheses.

worlds instead? Appealing to a principle of indifference over worlds (Bostrom, 2003a: 249; Chalmers, 2022: 99), it seems we should think that the odds are very high that we are living in one of the simulations. As Hans Moravec put it (in a discussion that supposes robots will be the ones running simulations), “in fact, the robots will re-create us any number of times, whereas the original version of our world exists, at most, only once. Therefore, statistically speaking, it’s much more likely we’re living in a vast simulation than in the original version” (Platt, 1995).

Moving on to longtermism, I want to acknowledge from the outset that there is a fair amount of indeterminacy as to how exactly to define the view, partly because it is so new. To get around this issue, my plan both in this section and throughout the paper is to make various claims about what longtermists *characteristically* hold, and then use these claims to define “longtermism” as I am using the word in this paper. There may be some self-described longtermists who reject the claims in question. If so, they fall outside the scope of my argument.

Following this strategy here, longtermists characteristically hold that time is comparable to space with respect to morality. Peter Singer (1972: 231-2) famously argues that spatial distance is morally irrelevant: we do not have less of a moral obligation to help needy strangers if they live thousands of miles away in a foreign country than we do if they live nearby. Analogously, longtermists hold that temporal distance is morally irrelevant: we do not have less of a moral obligation to help people live the best possible lives if they live thousands of years in the future than we do if they live in the present.

Longtermists also characteristically hold that we have the potential to make the world a much better place than it presently is by curing diseases and ending wars and

more. Indeed, part of their reason for emphasizing the distant future (part of their *anxiety* about the future, even) is that things could be *so much better* than they are right now if we play our cards right. When Bostrom (2000, 2005) imagines a future without aging or death, or when Musk envisions colonizing outer space with trillions of artificial people (Gallagher, 2023), these are not short-term plans to be executed over the weekend. No, they are long-term plans that we should start to think about today even if they cannot be fully executed for years and years to come. This point is important to my argument because the problem of simulated evil depends partly on the gap between how good the world presently is and how good it could be in the future.

Finally, for the purposes of the argument that follows, it is important to distinguish between the longtermist *moral view*, which says that positively influencing the long-term future should be a priority, and the longtermist *project*, which aims at successfully bringing about a good future. If an asteroid or a comet were to wipe out all life on Earth in 2030, this would constitute no objection to the longtermist moral view but it would mean that the longtermist project had failed, which is why Toby Ord (2020: 67-74) discusses the steps humanity has taken to safeguard ourselves from asteroids and comets in his longtermist book, *The Precipice*.

My argument in this paper is meant to be broadly in the spirit of an asteroid or comet. That is to say, I will not be questioning the longtermist moral view, I am willing to assume it is right, but instead will be focusing my fire on whether the longtermist project will be successful. I will be arguing that insofar as we have reason to believe that we are living in a computer simulation, we have reason to believe the longtermist project is *already* doomed to failure. This conclusion has practical implications, just as there would be practical implications to learning that Earth will be destroyed by an asteroid or

comet in 2030. If we have good reason to think the longtermist project is doomed, then it does not make sense to spend our time or money pursuing it, just as if we were to learn that an asteroid or comet is going to kill us all in a few years, it would not make sense to spend our remaining time thinking about ways to mitigate climate change.

1. (P1)

In this section, I discuss and defend the first premise of my argument:

(P1): There exist certain forms of evil in our world.

Which forms of evil do I have in mind? Those that self-described longtermist authors address in their writings and hope to remedy as part of their project to make the world a better place. So for example, consider deaths due to malaria. A number of philosophers, philanthropists, and other figures associated with the effective altruist movement and the longtermist project have worked hard to fight the disease, for instance by raising large sums of money to purchase insecticide treated bednets (MacAskill, 2015; Ord, 2020: 289, n. 6). Still, the World Health Organization's *World Malaria Report 2022* estimates that worldwide there were over 600,000 deaths per year due to the disease in both 2020 and 2021, a number that is strikingly high even while it is down compared to what it was decades ago (WHO, 2022).

Malaria deaths are a paradigmatic example of what is known within discussions of the problem of evil as a *natural evil*, meaning an evil brought about not as the result of human beings or other agents exercising their free will but rather due to the world operating according to the laws of nature. Natural evils occupy a special position within discussions of the traditional problem of evil because leading theist responses to the problem that might seem promising for cases of *moral evil*, or evil that results from free

will, seem less promising for natural evils. For the sake of my argument in this paper, I could rest my case for (P1) purely on the basis of malaria deaths. But it will be illuminating to expand the discussion by introducing further examples here, in particular examples that are pretty different from the cases typically considered in connection with the traditional problem of evil.

Toward this end, consider the topic of *human enhancement*, focusing especially on three radical forms of enhancement that Bostrom (2008) explicitly argues would be worthwhile. First, healthspan. Bostrom argues that our well-being would be significantly improved if we could engineer a way to live radically longer lives—lives that lasted hundreds or thousands of years, or perhaps even lives that approached immortality—and also radically healthier lives—lives that saw the eradication of all disease (including malaria of course). In connection, he argues that our lives would be much better if we could put an end to aging, that is, eradicate the sort of physical and mental decline that typically accompanies the aging process (see also Bostrom 2000 and 2005).

Second, cognitive enhancement. Bostrom argues that our well-being would be substantially improved if we could engineer a way to become more intelligent than we presently are, where such intelligence could be used to help us better understand the world (theoretical reason) and figure out ways to satisfy our desires (practical reason). Radically enhanced intelligence might help us make various scientific breakthroughs that we could use to further improve our lives, for instance.

Third, emotional enhancement. As Bostrom grants, this might take many forms, but one way it might work is that we could transform ourselves into beings who are much happier on balance and who experience longer and more intense pleasures as well

as other positive emotions. Imagine a future in which there is no more severe depression, for instance, and where people simply have more of a “zest for life.”

In discussions of the traditional problem of evil, it is not typical to include among the evils that God needs to answer for things like the fact that we experience the familiar indignities of aging (due to the lack of healthspan enhancement), or that we sometimes make mistakes when solving math problems (due to the lack of cognitive enhancement), or that we sometimes wake up in a bad mood (due to the lack of emotional enhancement). But for the purposes of the version of the problem of evil I am developing in this paper, these all count. In fact, for the purposes of my argument we can simply *define* “evil” as that which obtains at a world when that world falls short of what it could be with respect to well-being.³ Insofar as Bostrom is right and our lives would be much better if we could achieve various forms of radical human enhancement, then the unenhanced lives we presently live involve a kind of deficiency in well-being (evil), a deficiency that Bostrom contends we should try to correct by pursuing human enhancements.

To be sure, this proposed definition of “evil” does not perfectly capture how the term is typically used in other contexts. When I look in the bathroom mirror and see a stray grey hair on my head, I would never earnestly say, “This is evil, I am confronting evil.” But for the purposes of the argument I am constructing, this definition of “evil” can get us to the sought conclusion. Given this stipulative definition, (P1) is both very plausibly true and also something that longtermists should grant. For as noted in §0, longtermists characteristically deny that the world is *presently* such a good place that

³ There are competing philosophical views of well-being: hedonist views, desire-satisfaction views, objective list views. My proposed definition of “evil” can be flexible on this point, or it could say that evil is what obtains when a world falls short on *all* (or *most*) leading conceptions of well-being.

substantial upgrades to well-being in the future are impossible. Rather, they characteristically hold that we can make the world a much better place going forward. Or, to put the point suggestively, they characteristically hold that we do not *presently* live in the best of all possible worlds.

In §4, I will return to (P1) and consider a way of rejecting it that involves denying the existence of evil in our world. We will be in a better position to evaluate such an objection once my entire argument is on the table.

2. (P2)

The second premise of my argument:

(P2): If these forms of evil exist in our world, then our world is probably not a computer simulation run by longtermists.

The case for (P2) goes as follows. If at some point in the future, longtermists were to acquire the ability to run whole-world simulations, you would expect their simulated worlds not to have various evils that our world has. For example, if you knew you were going to be born into a computer simulation run by William MacAskill, you would not expect it to contain hundreds of thousands of malaria deaths per year. Given that MacAskill has spent so much time and energy trying to reduce the malaria deaths in our world down to zero, it would be very surprising if once he gained the God-like power to build a world of his own, he made sure to put all the malaria deaths back in. And as it goes for MacAskill, it also goes for longtermists in general. Conditional on the hypothesis that our world is a computer simulation run by longtermists, then, the number of malaria deaths we find in our world is very surprising. But in that case, the

number of malaria deaths that occur our world is *evidence against* the longtermist simulator hypothesis.

The same line of reasoning can be extended to the types of enhancement described in §1. Bostrom (2005) writes that “Searching for a cure for aging is not just a nice thing that we should perhaps one day get around to. It is an urgent, screaming moral imperative.” But then, if Bostrom himself got the chance to design a world, or if other longtermists persuaded by Bostrom’s argument did, why would they include the senescence of aging in their world? More generally, if longtermists are convinced that various forms of human enhancements would make our world a better place in the *future*, why not just include those enhancements from the get-go in the simulated worlds they create? The absence of radical enhancements in our world is surprising given the hypothesis that our world is a computer simulation run by longtermists. But in that case, the absence of radical human enhancements is evidence against the longtermist simulator hypothesis.

Some points of clarification are in order at this point. First, (P2) is phrased in terms in terms of a computer simulation “run by longtermists.” There are different ways this could work. It could be a human being like MacAskill or Bostrom running the simulation on a computer, but (in anticipation of §3’s discussion) it also could be an artificial superintelligence that is aligned with longtermist values. The idea is not that for any specific individuals (the current leading members of the longtermist movement, say), it is surprising that *they* would run a whole-world simulation that includes the evils that our world contains. Rather, the idea is that for any entity that is aligned with the values that characterize the longtermist view, it is surprising that such an entity would run a whole-world simulation that includes the evils we find.

Second, the consequent of (P2) says that our world is “probably” not a computer simulation run by longtermists. The line of argument I am offering here is not a version of the so-called *logical problem of evil* (Mackie, 1955; Plantinga, 1974). With the logical problem of evil, the idea is that the existence of evil in the world is logically inconsistent with its being created by an omnipotent, omniscient, and omnibenevolent being. As I noted in the introduction, there is no reason to expect a naturalistic creator of a simulated world to be omnipotent, omniscient, or omnibenevolent, and so there is no potential for a parallel logical problem of simulated evil. Instead, my argument is a version of the *evidential problem of evil* (Hume 1779/1990; Draper, 1989; Rowe, 1991). With the standard evidential problem of evil, the idea is that the existence of evil in the world provides strong but defeasible evidence against the theist hypothesis that God exists. On my variant of the problem, the idea is that the forms of evil we find in the world provide strong, defeasible evidence against the hypothesis that our world is a computer simulation run by longtermists.

Sometimes the evidential problem of evil is developed along Bayesian lines (Rowe, 1996; Plantinga, 1998; Climenhaga, forthcoming). Doing so here is clarifying. Here is a statement of Bayes’ rule:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

In the present case, let H be the hypothesis that our world is a computer simulation run by longtermists, and E a proposition describing the various forms of evil that exist in our world. Then within the Bayesian framework, E is evidence against H just in case $P(H|E) < P(H)$; that is, just in case learning E would lower the probability that a rational agent assigns to H .

I claim that this inequality obtains, with part of my case being that we should assign a very low numerical value to $P(E|H)$, or the likelihood that we would find the forms of evil that we do if our world were a computer simulation run by longtermists. My argument earlier in this section provide support for thinking this value should be very low: it is very surprising that our world contains so many deaths due to malaria, that it contains the mental and physical decline that goes with aging, and so on, given the assumption that we are living in a longtermist simulation. But further support for assigning $P(E|H)$ a very low value can be gained by saying more about the role that longtermists hope for computer simulations to play within their project.

Longtermists characteristically hope to bring about a future with a tremendous amount of well-being. In various ways, longtermist authors have noted, this goal is better served by the creation of conscious, artificial beings living in computer simulations than it is through the creation of future generations of biological human beings. For one thing, maximizing the sheer number of conscious beings in the future plausibly requires colonizing outer space since there is a finite carrying capacity to how many beings can live on Earth, and artificial beings figure to be better suited to space colonization than biological humans are (Sandberg, 2014; Greaves & MacAskill, 2021). It is harder for a biological human being to lead a good life on Neptune than it is to set up a computer on Neptune that runs a simulation in which an artificial being leads a good life, for example. For another thing, artificial beings plausibly will be more efficient than biological beings at converting resources into good lives (Bostrom, 2003b; Shiller, 2017; Greaves & MacAskill, 2021). You can get more bang for your buck—more happy, conscious beings—by pouring your finite resources into AI.

On the basis of considerations of this sort, Hilary Greaves and William MacAskill (2021: 8) estimate that the Milky Way could contain 10^{45} future conscious minds by relying on artificial intelligence, in contrast with “just” 10^{36} future conscious minds if we restrict the scenario to biological beings. The exact numbers are meant to be taken with a grain of salt, but the point is that the difference between these two estimates is immense—a billion artificial minds for every single biological mind, across the two scenarios (cf. Bostrom’s 2003b somewhat different but also astronomically large estimates).

And of course, the longtermist hope is not that the future will contain an extraordinary number of bad or middling digital lives, but that it will contain an extraordinary number of exceptionally *good* lives. Lives that are not cut short by diseases like malaria, for example. After all, for much the same reasons that longtermists hope to put an end to malaria on Earth, they should hope to avoid populating the Milky Way with computer simulations of trillions upon trillions of people dying from malaria. And also lives that contain the sorts of radical human enhancements that Bostrom envisions (§1). Insofar as such enhancements would add to the well-being of life on Earth, they presumably also would add to the well-being of simulated lives across the galaxy.

A striking way to put the longtermist goal here is that if their hope is realized, your life will be, statistically speaking, *one of the worst lives ever lived* over the history of the universe. I do not mean to pick on you—the same is true of my life and of everyone who has ever lived from the past up until the present. Given the potential for trillions and trillions and trillions of simulated lives to come, and the potential for these simulated lives to involve substantial upgrades with respect to well-being, your life and my life figure to be well within the bottom 1% of lives that ever will be lived. And in fact,

that's a very conservative estimate. Our lives might rank in the bottom trillionth of 1%, if things go well on the longtermist front.⁴

Extrapolating now from individual lives to entire worlds, I say that if longtermists were to take control of running whole-world simulations, then our world would figure to be one of the *worst worlds that will ever exist*. Just to be careful, my claim is not that we live in the unique “worst of all possible worlds,” as Arthur Schopenhauer (1859/2020: 1550) famously wrote, as an expression of his extreme pessimism. Rather, I am making the slightly less pessimistic claim that we live in one of the worst 1% worlds of all time, conditional on the assumption that it is longtermists who get to run the whole-world simulation process. Longtermists have hopes that they can do better than our world, and I am willing to go along and suppose they are right, at least 99% of the time. In connection, I propose that a credence of .01 is a reasonable conservative estimate for the value we should assign to $P(E|H)$.

In his *Dialogues Concerning Natural Religion*, David Hume's (1779/1990: 69) character of Philo at one point entertains the thought that “Many worlds might have been botched and bungled” by their creators as they worked to perfect their “art of world-making.” Along similar lines, my conservative .01 credence estimate is meant to build in some leeway to allow that longtermist simulators might botch and bungle some simulated worlds along the way. They might forget to turn off the “aging” setting on the simulation panel during their early efforts at world-making, for instance. And the .01 credence estimate also gives longtermist simulators some leeway to run some morally

⁴ You might be able to do better than the bottom 1% if you can survive until the arrival of some of these radical human enhancements. Maybe you will be able to live a life that approaches immortality, and in connection tap into various forms of human enhancement to come.

dubious simulations in which bad things happen to good sims, so that they can better understand the world scientifically. Maybe longtermists occasionally run simulations in which inhabitants die painful deaths to disease in order to better understand how pandemics work, for instance. Maybe there are yet other scenarios on which lesser simulated worlds are created. But the .01 figure puts an upper bound on how many lesser simulations longtermists are expected to run.

To be sure, the .01 number should be taken with a grain of salt. But, as Ord (2020: 166) notes while discussing his own subjective degrees of belief in various existential risk scenarios—Ord thinks we have a 1 in 10,000 chance of suffering existential collapse due to supervolcanic eruption during the next 100 years, a 1 in 1,000 chance due to nuclear war, and so on—working with specific numbers can be clarifying even when we do not take their precision overly seriously. In the case at hand, assigning a low figure like .01 to $P(E|H)$ virtually ensures that $P(H|E) < P(H)$, in which case the evil we find in the world does indeed qualify as evidence against the longtermist simulation hypothesis. If a longtermist (or anyone else) disagrees with this assessment, the .01 figure can help them clarify their disagreement. What percentage of simulated worlds do they expect longtermists to botch and bungle? Give a number: 2%, 5%, 10%, whatever. What percentage of simulated worlds are longtermists going to use to give people painful, avoidable, fatal diseases? How much more than 1% is it?

In assessing the present argument, it is important to note that there are alternative hypotheses on which the evils we find in our world are not nearly so surprising. One is the hypothesis that we do not live in a computer simulation at all but instead live in the real (unsimulated) world. A second is the hypothesis that we live in a computer simulation run by a superintelligent AI whose values are not aligned with our

values, and so who is indifferent to whether our lives could be made better (cf. Bostrom, 2014: 153). Conditional on various alternative hypotheses obtaining, the existence of the forms of evil that we observe in our world do not seem nearly so surprising. That is to say, the evil we find in our world is not evidence against these alternative hypotheses in the way that it is evidence against the longtermist simulator hypothesis.

Given our (provisional) assignment of the value .01 to $P(E|H)$, and given that there are alternative hypotheses that do not make E nearly so unexpected, I do not see a promising way of assigning values to $P(H)$ and $P(E)$ that avoids the result that $P(H|E)$ should be given a fairly low value; that is, that avoids the result that we probably are not living in a computer simulation run by longtermists, given the evils we find in our world. But in that case, (P2) is true.

3. (P3)

The third and final premise of my argument:

(P3): If our world is probably not a computer simulation run by longtermists, then either the longtermist project is probably doomed or else the simulation hypothesis is false.

The consequent of (P3) is a disjunction. Let us start by focusing on its first disjunct. Consider the scenario in which our world is a computer simulation, but not run by longtermists. I claim that this would be very bad news for the longtermist project, it would mean that the project is probably doomed. To defend this claim, I can rely entirely on arguments that longtermists themselves have endorsed in their discussions of artificial superintelligence and the *value alignment problem* (Bostrom, 2014; Ord, 2020: Chapter 5; MacAskill, 2022: Chapter 4).

Suppose that human beings are able to develop artificial superintelligence one day, as is plausibly required in order to run whole-world computer simulations in the first place (Bostrom, 2014). What will civilization look like after such a development? According to Bostrom (2006; 2014), the most likely scenario is that it will lead to a *singleton* or *unipolar* outcome, meaning that there is a single, unified, decision-making agency at the highest level. There are different ways this might unfold, but it will be clarifying to think through a particular example, so consider a scenario that involves Bostrom's (2003c) infamous paperclip maximizer.

Imagine that the first artificial superintelligence to be created is such a device, a being whose sole objective or final value is to produce the largest number of paperclips possible. Upon its creation, the paperclip maximizer will immediately have good instrumental reason to try to prevent the creation of any further superintelligences, since such systems are potential rivals that might interfere with the project of maximizing paperclips. If, say, an AI lab in Montreal is just a few months away from completing a superintelligent staple maximizer, then the already existing paperclip maximizer would have reason to prevent this from happening, perhaps killing the Montreal team and destroying their work based on the decision-theoretically rational calculation that this is the action available to it that maximizes the expected number of paperclips in existence.

Now, maybe the first artificial superintelligence won't be a paperclip maximizer, maybe it will instead be something better aligned with human values. At any rate, whatever form the first artificial superintelligence happens to take, it potentially will possess a major strategic advantage (Bostrom, 2014: Chapter 5), an advantage it can use to promote whatever values it does happen to hold, and to neutralize threats to those

values, including especially threats represented by later developed superintelligences with different values. In other words, it has the potential to use its strategic advantage to bring about a singleton or unipolar outcome.

Now, if this first artificial superintelligence were aligned with longtermist values, the resulting singleton or unipolar outcome should be greeted by longtermists as a very welcome result—nothing could be a greater boon to achieving the longtermist project. The AI could devote itself to creating and spreading across the universe computers that run whole-world simulations whose inhabitants live the sort of fantastic lives that longtermists want conscious agents to live. And the longtermist AI could use its strategic advantage to prevent the creation of artificial superintelligences that are misaligned with human values, superintelligences that would be indifferent to or even in favor of pain and suffering in conscious agents. It could prevent the creation of superintelligences that run computer simulations in which people die painful deaths from malaria, for example, or in which they grow frail from age.

In no small part because of the promises and perils of artificial superintelligence, longtermists often emphasize the unique world-historical importance of the era we find ourselves in. According to Ord (2020: 7), “we live during the most important era of human history.” According to Holden Karnofsky (2021), “the 21st century could be the most important century ever for humanity.” According to MacAskill (2022: 117), if we do manage to create artificial superintelligence, it will be “one of the most important developments in all of history.” And according to Bostrom (2014: v), the issue raised by superintelligence is “quite possibly the most important and daunting challenge humanity has ever faced.”

How will things play out in this most important century? Well, if we are living in a computer simulation, then it seems like the conclusion to draw is that the 21st century has probably *already* played out, and that they have played out badly for the longtermist project. For if the artificial intelligence that (we are supposing) runs our world is misaligned with longtermist values, as evidenced by the existence of evils like malaria deaths and aging, then longtermists have very likely already blown their chance at creating the future they hope for. There is probably no comeback from having the singleton superintelligence that runs the world misaligned with your values. “God is against us,” longtermists might ruefully truthfully say, if they accept the view that the simulation hypothesis is a special case of the theist view and so take the term “God” to denote the being who runs our simulation (Chalmers, 2022: Chapter 7).

One way to put the thought here is that the simulation hypothesis conflicts in a surprising way with the hope that we could make the world a better place in the future. The better you think the future could be compared to the present, the more of a problem it is for you that the world is not *already* better in that way, given the simulation hypothesis. If you were a time-sliced Leibnizian who was content that we already live in the best of all possible worlds + times, then nothing you see in the world around you is evidence of a superintelligent similar misaligned with your values. It is precisely because longtermists characteristically hope for a future vastly better than the present that the problem arises. The bigger the gap you think there is between what presently is the case and what could be the case in the future, the greater the degree of value misalignment there figures to be between you and the simulator who runs our world.

We might put the idea here as a mock equation: Karnofsky + Moravec = Longtermist Doom. That is, suppose with Karnofsky that the 21st century will probably

be the most important century that humanity ever faces. And suppose with Moravec (from §0) that statistically speaking, we are probably living in a computer simulation of the 21st century instead of the unsimulated version. Then, the evils we find in the world are evidence that the actual (unsimulated) 21st century did not go the way longtermists would hope, in which case the longtermist project is probably doomed.

This takes us to the second disjunct of the consequent of (P3), which says that the simulation hypothesis is false. Here, the thought is that the preceding line of argument falls apart entirely if we suppose we are not living in a computer simulation, if we reject the Moravec component of our equation. If we are living in the real (unsimulated) world, then the fact that longtermists have not *yet* achieved their long-term goals is not evidence that they *never will*, because by assumption there is no superintelligent AI running our world, and so the evils of our world are not evidence that the values of such an AI are misaligned with those of longtermists.

4. Conclusion

This leads me to the conclusion of my argument.

(C): Either the longtermist project is probably doomed or else the simulation hypothesis is false.

I take no stand on how to resolve the tension between the longtermist project and the simulation hypothesis. Perhaps we should adopt a pessimistic view of the project, or perhaps there is compelling reason to reject the simulation hypothesis, or perhaps we should be willing to live with uncertainty, embracing the disjunction that is (C) without committing ourselves to either disjunct.

Now that the full case for my argument is on the table, I want to return to the objection to (P1) that I alluded to back in §1, which denies the existence of evil in our world.⁵ One version of this objection endorses a form of *solipsism* saying that you are presently living in a computer simulation that contains only one conscious mind: your own. If everyone else in the world lacks phenomenal consciousness, then all of their apparent suffering is merely an illusion, and so no genuine evil is involved. Or instead of just a single mind, perhaps the world contains a few conscious minds along with a number of zombie imposters, provided that none of the genuine minds are suffering.

What about your memories of your own past suffering? The view I am inviting us to consider continues by saying this too is an illusion. If our world is a computer simulation created just a few minutes ago with false memories built in, then your memories of past suffering are mistaken. In this way, we can construct a scenario in which (P1) is false and our world contains no real evil. In a discussion exploring the connections between the simulation hypothesis and solipsism, Grace Helton (2021: 9) even suggests that this is what we should expect superintelligent AI aligned with human values to do, to run whole-world simulations with illusions of suffering and evil rather than the genuine article. Maybe this is the way to resolve the tension between the simulation hypothesis and the longtermist project that I have been arguing for.

Now, in the context of the traditional problem of evil regarding a supernatural God, the response of denying the premise that evil exists is widely regarded as one of the least promising options for theists (for discussion, see Mander, 2018). A solipsistic version of this response is especially uncommon, given that it would undermine core

⁵ Thanks to an anonymous referee for pressing me to consider this.

elements of standard religious beliefs concerned with the suffering of others. That said, Christian Science founder Mary Baker Eddy (1934) maintained that evil is an illusion—one bound up with the illusion that the material world is real. It would be ironic if denying the existence of evil united Christian Scientists—who often avoid medical treatments on the grounds that illness is an illusion—with longtermists—who are in favor of using biomedical means extremely widely, including as a tool to put an end to aging and alter other central features of human nature. Strange bedfellows.

The objection to (P1) we have been considering does not work if you are presently suffering. If, as you read these words, you are experiencing a splitting headache, this is enough for you to know that there exists a certain form of evil in the world, in which case (P1) is true. And given the expansive conception of “evil” I am working with (§1), this point will generalize widely. As I write these words, I feel a bit hungry, and in connection a little grumpy—I’m “hangry” (hungry-angry). This suffices to make (P1) true, since it represents a deficiency in well-being that could be avoided. If longtermists manage to simulate trillions and trillions of lives, we should expect comparatively few of those lives to ever experience hanger. In simulated utopias, light snacks should be abundant, and good moods universal. My hanger is therefore enough for the evidential problem of simulated evil to rear its head; it gives me reason to lower the probability I assign to the hypothesis I am living in a simulation run by longtermists.

Instead of leaning too hard into this point, however, my preferred response to the present objection is to grant that something along the lines of solipsism provides one way to block my argument, and then allow readers to gauge whether the theoretical costs of this response are worth paying. In his original discussion of the simulation hypothesis, Bostrom (2003: 254) suggests in passing that something like the solipsistic

scenario we have been discussing might provide a “far-fetched” solution to the traditional problem of evil for a supernatural God. But maybe this far-fetched option demands more attention from those proponents of the longtermist project that take the simulation hypothesis seriously, given the argument I have presented in this paper.

WORKS CITED

- Bostrom, N. (2000) The case against aging, [online], <https://nickbostrom.com/aging/aging>.
- Bostrom, N. (2003a) Are you living in a computer simulation? *Philosophical Quarterly*, 53 (211), pp. 243-255, DOI: 0.1111/1467-9213.00309
- Bostrom, N. (2003b) Astronomical waste, *Utilitas*, 15 (3), pp. 308-314, DOI: <https://doi.org/10.1017/Sop53820800004076>.
- Bostrom, N. (2003c) Ethical issues in advanced artificial intelligence, in Lasker, G. E., Smit, I., & Wallach, W. (eds.) *Cognitive, Emotion and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, pp. 12-17.
- Bostrom, N. (2005) The parable of the dragon-tyrant, *Journal of Medical Ethics*, 31 (5), pp. 273-277, DOI: 10.1136/jme.2004.009035.
- Bostrom, N. (2006) What is a singleton? *Linguistic and Philosophical Investigations*, 5 (2), pp. 48-54.
- Bostrom, N. (2008) Why I want to be posthuman when I grow up, in Gordijn, B. & Chadwick, R. (eds.) *Medical Enhancement and Posthumanity*, Springer, pp. 107-137, DOI: https://doi.org/10.1057/9781137349088_15.
- Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press.
- Chalmers, D. J. (2022) *Reality+: Virtual Worlds and the Problems of Philosophy*, W. W. Norton & Company.
- Climenhaga, N. (forthcoming) If we can't tell what theism predicts, we can't tell whether god exists: skeptical theism and bayesian arguments from evil, in Buchak, L. & Zimmerman, D. (eds.) *Oxford Studies in Philosophy of Religion, Volume 11*.

- Crummett, D. (2020). The real advantages of the simulation solution to the problem of natural evil, *Religious Studies*, 57 (4), pp. 1-16, DOI: <https://doi.org/10.1017.S003441251900726>.
- Dainton, B. (2002) Innocence lost: simulation scenarios: prospects and consequences, [online], <https://philarchive.org/rec/DAILLS>.
- Dainton, B. (2020) Natural evil: the simulation solution, *Religious Studies*, 56 (2), pp. 209-230, DOI: <https://doi.org/10.1017/S0034412518000392>.
- Draper, P. (1989) Pain and pleasure: an evidential problem for theists, *Noûs*, 24, pp. 331-350, DOI: [10.2307/2215486](https://doi.org/10.2307/2215486).
- Eddy, M. B. (1934) *Science and Health with Key to the Scriptures*. Christian Science Publishing Society, authorized edition.
- Gallagher, B. (2023) The race to colonize mars perpetuates a dangerous religion,” *Nautilus*, [online] <https://nautil.us/the-race-to-colonize-mars-perpetuates-a-dangerous-religion-298323/>.
- Godfrey-Smith, P. (2022) Is longtermism such a big deal? *Foreign Policy*, [online], <https://foreignpolicy.com/2022/11/12/longtermism-william-macaskill-book-elon-musk-philosophy-ethics/>.
- Gohd, C. (2017) Are we living in a computer simulation? Elon Musk thinks so, *Futurism*, [online], <https://futurism.com/are-we-living-in-a-computer-simulation-elon-musk-thinks-so>.
- Greaves, H. & MacAskill, W.. (2021) The case for strong longtermism, *GPI Working Paper*, No. 5-2021.
- Helton, G. (2021) Epistemological solipsism as a route to external world skepticism, *Philosophical Perspectives*, 35 (1), pp. 229-250.
- Hume, David. (1779/1990) *Dialogues Concerning Natural Religion*. Penguin Classics.
- Johnson, D. K. (2011) Natural evil and the simulation hypothesis,” *Philo*, 14 (2), pp. 161-175, DOI: [10.5840/Philo201114212](https://doi.org/10.5840/Philo201114212).
- Karnofsky, H. (2021) The ‘most important century’ blog post series, *Cold Takes*, [online], <https://www.cold-takes.com/most-important-century/#:~:text=The%20%22most%20important%20century%22%20series,people%20imagine%20to%20a%20deeply>.
- MacAskill, W (2015) *Doing Good Better: Effective Altruism and How You Can Make a Difference*. Random House.

- MacAskill, W. (2022) *What We Owe the Future*. Basic Books.
- Mackie, J. L. (1955) Evil and omnipotence, *Mind*, 64, pp. 200-212, DOI: 10.1093/mind/lxiv.254.200.
- Mander, W. J. (2018) The unreality of evil, *Sophia*, 57 (2), pp. 249-264, DOI: <https://doi.org/10.1007/s11841-017-0585-x>
- Ord, T. (2020) *The Precipice: Existential Risk and the Future of Humanity* Hachette Books.
- Plantinga, A. (1974) *The Nature of Necessity*, Oxford University Press.
- Plantinga, A. (1998) Degenerate evidence and Rowe's new evidential problem of evil, *Noûs*, 32 (4), pp. 531-544, DOI: 10.1111/0029-4624.00137.
- Platt, C. (1995) Superhumanism, *Wired*, [online], <https://www.wired.com/1995/10/moravec>.
- Rowe, W. L. (1991) Ruminations about evil, *Philosophical Perspectives*, 5, pp. 69-88.
- Rowe, W. L. (1996) The evidential problem of evil: a second look, in Howard-Snyder, D. (ed.), *The Evidential Problem of Evil*, Indiana University Press, pp. 262-285.
- Sandberg, A. (2014) Ethics of brain emulations, *Journal of Experimental and Theoretical Artificial Intelligence*, 26 (3), pp. 439-457, DOI: <https://doi.org/10.1080/0952813X.2014.895113>.
- Shiller, D. (2017). In defense of artificial replacement, *Bioethics*, 31 (2), pp. 393-399, DOI: 10.1111/bioe.12340
- Schopenhauer, A. (1859/2020) *The World as Will and Idea*, Haldane, R. B. & Kemp, J. (trans), Mosaic Books.
- Shea, B. (2017) The problem of evil in virtual worlds, in Silcox, M. (ed.) *Experience Machines: The Philosophy of Virtual Worlds*, Rowman & Littlefield, pp. 137-154.
- Singer, P. (1972) Famine, affluence, and morality, *Philosophy and Public Affairs*, 1 (3), pp. 229-243.
- Tegmark, M. (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence*, Alfred A. Knopf.
- WHO (2022) *World Malaria Report 2022*, Geneva: World Health Organization.,

License: CC BY-NC-SA, 3.0 IGO.