Simulations and Catastrophic Risks

Bradford Saad
Researcher, Utrecht University; Research Fellow, Sentience Institute
(t.b.saad@uu.nl; brad@sentienceinstitute.org)
June 2023 (v1.7)

Preface

This report explores interactions between large-scale social simulations and catastrophic risks. It offers a tour of the surrounding theoretical terrain and brings together disparate literatures that bear on the topic. My hope is that the report will facilitate further research in this area and directly or indirectly inform the decisions of benevolent actors who face choices about whether and, if so, how to develop and use powerful simulation technologies.

While the report is long, most of its sections are self-contained. So, readers should feel free to skip to section(s) of interest to them, perhaps after browsing the overview and preliminaries in §§1-2.

The report largely focuses on *philosophical* connections between simulations and catastrophic risks. That's because, as a philosopher, philosophy is what I know best, not because I seriously considered focusing on alternative sorts of connections and then deemed philosophical connections most worthy of attention. That said, I do think that these philosophical connections are worthy of attention: as the report illustrates, they are both important from a risk mitigation perspective and a fascinating subject matter from a perspective of pure inquiry.

I have aspired to write the report in an even handed manner. In particular, I have tried to bracket my own tendentious views in order to identify what I think should be widely recognized as key questions in this area and factors that are relevant to answering those questions. And I have incorporated sensitivity analyses of how different issues would play out, depending on one's background views. However, I have no doubt fallen short in this aspiration. Fully bracketing one's own controversial views is no easy thing. And I have not hesitated to rely on my own views about what should be controversial or to say that considerations point in a particular direction when I think this is clear.

I started writing this report in the summer of 2020. I continued working on it intermittently through the spring of 2023. With the public release of large language models and a surge in research interests in artificial intelligence and catastrophic risks in late 2022 and early 2023, I came to the realizations that relevant work was coming out faster than I could read and incorporate it and that any attempt to address the state of the art would quickly become dated. While recent developments have not prompted any major changes to my analysis in the report, they have shortened my AI timelines and updated me toward thinking that there are substantial costs to delaying work on

mitigating catastrophic risks associated with simulations. Thus, I have decided to release the report in its current form as a living document—one that I may occasionally update with important additions or corrections, though I have no intention of (futilely) trying to keep it up to date with the many relevant and rapidly evolving literatures.

For helpful discussion or comments, I am grateful to Jacy Reese Anthis, Austin Baker, Zach Barnett, Brian Cutter, Adam Gleave, Michael Dello-Iacovo, Drew Johnson, Fintan Mallory, Ali Ladak, Richard Ngo, Janet Pauketat, Jonathan Simon, and participants in an ULTIMA colloquium at Utrecht University. For feedback on related works that was especially helpful for this project, I am grateful to Daniel Berntson, Emery Cooper, Han Li, and Caspar Oesterheld.

-*Bradford Saad, May, 2023*

1. Introduction

Future simulation technology is likely to both pose catastrophic risks and offer means of reducing them. While there is much relevant work on the topic, it is scattered across disparate literatures. The main goal of this document is to bring together existing work in order to facilitate future research that will produce better understanding of the topic and help mitigate associated risks.

To orient readers, I'll start with an overview. Individual sections are largely self-contained. However, §2 offers preliminaries that some readers may find helpful for later sections. And, for readers who are unfamiliar with the simulation hypothesis or the simulation argument, I would recommend reading §8 before reading any of §§9-13. To make it easier for readers to read at their desired level of depth, I'll use bullet points, with details, examples, etc. in **nested bullet points** that **can be skipped**.

Here, then, is an **overview of sections and their key contributions**:

- Simulation as a tool for researching risk reduction (**§3**)
    - Simulations are promising as tools for directly researching a wide range of catastrophic risks and how to reduce them. (**§3.1**)
    - Simulations are also promising as tools for researching factors that indirectly bear on catastrophic risk levels. These include value dynamics, evolutionary debunking arguments, consciousness, cognitive enhancement, and Fermi's paradox. (**§3.2**)
    - Research simulations have dual use potential that their designers and users are apt to underestimate. I see guarding against the downside potential of dual use simulation technologies as a promising area for reducing catastrophic risks. (**§3.3**)
    - We cannot safely assume that superintelligent systems will supersede large-scale research simulations before the latter become available. (**§3.4**)
- Simulation as a tool for ethically-enhanced testing (**§4**)
    - Relative to corresponding unsimulated testing, simulated testing could ethically enhance testing by reducing risks to the outside world, by reducing suffering risks of participants, or by enabling participant consent.
- Simulations as tools for promoting risk responsiveness (**§5**)

    - Immersive simulations and gaming simulations are neglected and tractable interventions for reducing catastrophic risks. Their potential impact is unclear but ripe for investigation.
- Simulation refuges (**§6.1**)
    - For the purposes of surviving and bouncing back from catastrophes, simulation refuges would have important advantages over non-simulation refuges.
    - The best use of simulation refuges would likely depend sensitively on whether their inhabitants would be conscious.
- Roles for simulations in grand futures (**§6.2**)
    - At least in expectation, much of the positive value of the future lies in scenarios with large-scale virtual paradise simulations.
    - Simulations could be used to look before we leap in selecting a path toward a grand future.
- Simulations as fallback options (**§6.3**)
    - Simulations hold significant promise as fallback options to use in the event that immensely positive futures become infeasible, though their promise is beholden to simulations with conscious minds becoming available.
- Simulations that would constitute catastrophes (**§7**)
    - There is a disconcerting range of at least somewhat plausible scenarios in which simulations would generate catastrophic levels of disvalue. These scenarios include ones with simulations causing catastrophes while being used for research, entertainment, economic activities, manipulation, or to pose threats, as well as ones in which catastrophes are induced through simulations either inadvertently or through malevolent intent.
- I give a primer on the simulation hypothesis that our universe is a simulation, the simulation argument, and a taxonomy of associated objections. I offer a simple, somewhat schematic formulation of the simulation argument. This formulation facilitates discussion of the argument and connections with catastrophic risks that is unhampered by the technical features of more sophisticated formulations. After that, I discuss some common objections to the simulation hypothesis and argument and show how those objections fail. Along the way, I identify interactions between the simulation hypothesis, the simulation argument, and objections. (**§8**)

- The shutdown of our universe-simulation poses a neglected catastrophic risk. The extent to which we can mitigate this risk is unclear and underexplored. Mitigating it should not be dismissed as wholly intractable. Whether or not the risk can be mitigated, it weakens the case for longtermist interventions. (**§9**)
- Triggering simulation shutdown offers a potential long-shot escape hatch from worse catastrophes. Even in catastrophic scenarios in which taking it is part of the best option, taking it too soon can itself be a catastrophic moral error. (**§10**)
- The simulation hypothesis and simulation argument interact with arguments for and against religious hypotheses. These interactions modulate the plausibility of those hypotheses and various associated religious catastrophic risks (**§11**). The net effect on associated risk levels is unclear, as different interactions push in different directions and estimations of these risk levels depend sensitively on background views about which there is much disagreement.
- To set the stage for subsequent subsections, I provide a framework for investigating how self-locating beliefs (that is, indexical beliefs about one's own place in the world) bear on catastrophic risks. (**§12.1**)
    - I formulate coarse-grained versions of these principles that allow for a relatively non-technical discussion of their bearing on catastrophic risks.
    - I distinguish five dimensions along which these principles can be precisified.
    - An important problem for a more fine-grained principle (the self-sampling assumption') that bears on a range of issues relevant to catastrophic risks turns out to rely on non-mandatory precisifications of a plausible coarse-grained principle.
    - I suggest that coarse-grained self-locating principles can be assimilated into a more general class of inductive principles that are clearly warranted despite their resistance to precisification. This alleviates concerns about coarse-grained principles of self-locating belief that turn on their resisting precisification.
- The fact that evolution produced human-level intelligence provides at least a measure of support for the hypothesis that we will be able to engineer systems with human-level intelligence. However, this support is probabilistically screened off by more general facts about our causal origins, facts that we knew about before learning of our evolutionary origins. (**§12.2**)

- A more promising argument appeals to more-specific evolutionary facts. That argument suggests that we will be able to engineer human-level intelligence. (**§12.2**)
    - That conclusion in turn supports the simulation argument.
    - The conclusion also modulates risk levels via the simulation argument, Fermi's paradox, and the hypothesis that we will create superintelligent agents.
    - The degree of direct and indirect evidential import of the argument depends on how principles of self-locating belief are precisified and on the operative reference class for an observation selection effect.
- I offer a simple formulation of the doomsday argument and identify a range of interactions between it, the simulation argument, principles of self-locating belief, and catastrophic risks. (**§12.3**)
    - In different ways, the doomsday argument and the simulation argument each casts doubt on the other.
    - However, the simulation argument coheres with a version of the doomsday argument that supports doom for beings like us in our simulation but not for our reference class more broadly.
    - By the lights of the doomsday argument, a promising risk mitigation strategy is to engineer our own replacement through digital minds who live valuable lives in simulations and who fall outside our reference class.
- I identify interactions between Fermi's paradox, the simulation argument, principles of self-locating belief, and catastrophic risks. (**§12.4**)
    - The simulation argument suggests a solution to Fermi's paradox: we do not observe other civilizations because we're in a simulation-universe that is smaller than it appears.
    - According to another simulation-based solution to Fermi's paradox, we do not observe other simulations because their activities tend to be confined to simulations they have created. This solution indirectly bolsters the simulation argument.
    - Fermi's paradox suggests that few advanced civilizations have been in a position to trigger simulation shutdown, regardless of shutdown risk. This should raise our estimates of simulation shutdown risk. The same goes for solutions that posit a small number of advanced civilizations in our universe.

    - Fermi's paradox suggests that engineering intelligence is not easy for evolution. That tells against the hypothesis that we will engineer human-level intelligence and, in turn, against the simulation argument.
    - Rare Earth solutions to Fermi's paradox exhibit a form of fine-tuning that supports multiverse and design hypotheses that fit with the simulation hypothesis.
    - Self-locating belief principles that support abundant observer hypotheses also support abundant civilization solutions to Fermi's paradox. On the other hand, if we take the apparent absence of other civilizations at face value, Fermi's paradox disconfirms these principles and thereby attenuates their impact on catastrophic risks.
- There is a striking analogy between the simulation argument and Boltzmann brain problems in cosmology. Given the relevance of the former to catastrophic risks, I explore interactions between the two. (**§12.5**)
    - Some proposed solutions to Boltzmann brain problems parallel objections to the simulation argument and fail for the same reasons.
    - Important differences between the simulation argument and Boltzmann brain problems include differential sensitivity to choice of reference class and differences in skeptical import.
    - One class of solutions to Boltzmann brain problems undermines the simulation argument by suggesting that simulations would be unconscious.
    - A simulation-based solution to Fermi's paradox can be extended to solve Boltzmann brain problems.
    - Some principles of self-locating belief favor Boltzmannian cosmologies. Given that these cosmologies engender skepticism, they give us empirical grounds for rejecting those principles.
- I propose a neglected strategy for reducing a wide range of catastrophic risks. The strategy combines insights from the simulation argument and the evidentialist wager. I identify factors to consider and mistakes to avoid in implementing the strategy. (**§13**)
- To conclude, I highlight some open questions and promising research avenues (**§14**)

2. Preliminaries on Simulations and Catastrophic Risks

Simulations are systems that are designed and created to model other processes. I will understand simulations broadly to include not only computer simulations but also systems that couple biological subjects with virtual environments. I will primarily but not exclusively focus on large-scale social simulation scenarios, i.e. ones involving simulations that model at least tens of thousands of minds, some cognitive processing within each of those minds, and interactions between those minds. As we will see, whether simulations themselves contain (conscious) minds will matter for some purposes but not others.

My focus will be on catastrophic risks that interact with simulations and which are high stakes in that they either threaten millions of (present or future) people with significant harm or else pose a risk of a comparably bad outcome.[1] These will include risks of astronomical quantities of suffering[2] and existential risks, i.e., risks of catastrophes that would permanently destroy humanity's future potential.[3] Given how bad such catastrophes would be, the risk of them could easily be worth mitigating even if their probability is low. Thus, the discussion will not be restricted to high-probability catastrophic risks. Nor will it be restricted to risks of acute catastrophes rather than ones that unfold over, say, many generations. However, the category of high-stakes risks on which I'll focus is somewhat broader than astronomical suffering and existential risks involving simulations. That's partly because it includes risks of harms (not necessarily involving suffering) to digital minds that are comparable to existential risks and partly because the category includes risks of less severe catastrophes involving harm to millions of people (or something comparably bad). I focus on this broad category for two reasons. First, I think that the less severe catastrophes in this category are more likely to occur but still bad enough to be well-worth preventing. Second, I anticipate that the most politically tractable way to mitigate the more severe risks in this category may be via interventions that target the less severe risks in the category. Hereafter unless otherwise indicated, I'll use 'catastrophic risks' as shorthand for high-stakes, simulation-involving risks in the just described category.

[1] For discussion of various sorts of global catastrophic risks, see the essays in Bostrom & Ćirković (2008).

[2] See Baumann (2022), Gloor (2016; 2018), and Daniel (2017).

[3] See Ord (2020).

The different catastrophic risks I discuss will be associated with different types of large-scale social simulations. For every type of simulation I discuss, I believe there is a non-negligible probability that that type of simulation will be run. However, these probabilities vary widely for different types of simulation. While such probabilities are important for evaluating the quantitative impact of simulations on catastrophic risks, the discussion will mostly proceed at a coarse-grained level that is insensitive to these probabilities. I'm hopeful that this report will prompt others to pursue more fine-grained and quantitative analyses. Still, it is worth laying out what I see as some of the key differences in the plausibility of different types of simulations, as doing so may give a sense of how speculative different parts of the discussion are and offer something to go on for more fine-grained analyses. And giving readers a glimpse of my underlying mental models may put them in a better position to understand and evaluate the discussion that follows.

To that end, it will be useful to distinguish several axes of variation among types of large-scale social simulation:

- type of technology used to run the simulation.
- the computational complexity of the simulation
- the purpose(s) of the simulation
- whether the simulation contains conscious minds.

As rules of thumb, I take the probability that a given type of large-scale social simulation will be run to be inversely related to how technologically demanding it is and to its computational complexity. These are merely rules of thumb partly because the economic incentives to run simulations need not scale with technological demandingness or computational complexity. Absent near-term catastrophes and stringent regulatory intervention that halt technological progress, I think it is highly probable (> 90%) that at least thousands of large-scale simulations will be run for research, entertainment, or economic purposes within the next century.[4]

As noted, for some issues raised by simulations, it is crucial whether simulations would themselves contain conscious minds. For instance, catastrophic suffering risks will not arise within simulations that are clearly devoid of consciousness. Other issues chiefly

[4] Research is already underway on small-scale (25 agents) social simulations that embed instances of OpenAI's ChatGPT in virtual social contexts (Park, et al., 2023).

concern the effects of simulations on the external world. For instance, whether a simulation that models nuclear winter can be used to improve the prospects for recovery from nuclear winter does not turn on whether the simulation features conscious inhabitants. In what follows, I will address the consciousness of simulation inhabitants where relevant in connection with particular issues.

But it's worth noting from the outset that whereas large-scale social simulations of some sort are clearly feasible, large-scale social simulations that contain conscious minds are not clearly feasible. The former can be achieved by scaling up existing technologies. For extremely simple simulations of minds, it would be relatively easy to create a large-scale social simulation. For example, such simulations could be achieved with existing technology by scaling up real-time strategy games that simulate hundreds of interacting agents and incorporating rudimentary simulations of cognitive processes. Much more advanced large-scale social simulations could also be achieved by embedding within a virtual environment digital agents trained through machine learning. Training such agents is computationally expensive. But once trained, it is relatively inexpensive to such agents across many tasks at once. So the currently high costs of training advanced machine learning agents may be less of an obstacle than one might have thought to creating large-scale social simulations that are populated by such agents.[5] However, it is not clear that machine learning architectures are suitable for realizing consciousness. More generally, it is not clear that any existing computer technologies are of the right sort to generate consciousness. But there is reason to think that more promising technologies are on the way: efforts are already underway to imbue large language models with sensory capacities, agency, and world-models to integrate them with robotic systems.[6] And, in the future, whole brain emulations and neuromorphic systems may exhibit a high degree of functional similarity with brain processes that underlie consciousness, which would provide reason to think that such systems are conscious.[7] It would also be unsurprising if superintelligent systems engineered hitherto unconceived types of architecture with the potential for consciousness.

[5] See Davidson (2023).

[6] See Huang (2023), wikipedia.org/wiki/Auto-GPT, Li et al. (2022), and Vemprala et al. (2023)

[7] See Chalmers (1996: Chs. 7, 9) and Sandberg & Bostrom (2008).

Thus, while I take it that large-scale social simulations of rudimentary sorts are already feasible and their widespread future deployment is highly likely, I assign only a ~60% probability to the hypothesis that large-scale simulations with conscious minds will become feasible, given that technological progress is not halted in the next century.

Conditional on large-scale social simulations that contain conscious minds becoming feasible, I think it is unclear whether such simulations will be run on a large-scale (say, with at least thousands of such simulations). Conditional on their becoming feasible, I assign ~40% probability to their being run on a large scale unwittingly (i.e. we run them without believing that they contain conscious minds), ~60% probability to their being intentionally run on a large scale , and ~25% probability to their not being run on a large scale (e.g. because humans universally enforce a ban on them (~5%) or because we lose control to artificial agents that opt not to run them (~15%)).[8] Even in cases where large-scale social simulations come to house more conscious minds than there are humans, I would expect there to be more large-scale social simulations that do not contain conscious minds.

Many of the issues I discuss in what follows arise in a wide range of potential future scenarios with large-scale social simulations. By my lights, no specific scenario of this sort stands out as especially likely. So I will mostly discuss these issues in the abstract rather than in the context of particular scenarios. Still, some readers may find it helpful to think through these issues in the context of some concrete scenarios.[9] For this purpose, I'll now offer some stylized scenarios.[10] These scenarios are wild and speculative. This comes with the territory, as it is highly probable that the future will be

[8] These sum to more than 100% because the first two possibilities are not exclusive: we could on large scales both unwittingly run some simulations containing conscious minds and intentionally run others. Or multiple agents running the same simulations could hold different views about whether those simulations contain conscious minds.

[9] The situation here parallels the situation with AI-involving catastrophic risks more generally. While the sources of risk do not depend on the specifics of concrete scenarios we can conjure, it is still advisable to describe concrete scenarios since the uninitiated may find abstract risks difficult to take seriously without first seeing how they could be concretely realized. And it is advisable to emphasize that the general catastrophic risks do not depend on the details of those scenarios, lest the target audience take challenges to details of those scenarios to be levers for driving down estimates for the general risks they illustrate. See Christiano (2019), Cotra (2022), Critch (2021), Hendrycks (2023), Lawsen (2023), and Mowshowitz (2023)

[10] For other presentations and discussion of scenarios, see Bostrom (2003*a*), Chalmers (2010*b*; forthcoming), Dainton (2012), and Hanson (2016).

wild,[11] and specifying concrete future scenarios is an inherently speculative endeavor. The rest of the report won't presuppose familiarity with these scenarios. So readers should feel free to skip them.

> *Virtualization of labor:* In 2090, whole brain emulations arrive. Because they emulate human brains, they can perform any cognitive task that humans can perform. However, they can be run at much faster speeds than the human brain and at low costs. While humans continue to command much of the capital in the economy, most human labor is largely priced out by whole brain emulations. Because it is cheaper and safer to house whole brain emulations in controlled virtual environments than it is to equip them with robotics in the external world, they predominantly inhabit simulations. There is no global consensus on whether whole brain emulations are conscious or whether they have moral status. For ethical reasons and/or to preserve human jobs, virtual labor is initially banned in some jurisdictions. However, these policies impose substantial economic costs on these jurisdictions. The turn to virtual labor drives investment and productivity in places that do not heed such scruples. Eventually, as opposition dwindles, whole brain emulations come to dominate the labor force nearly everywhere.[12]
>
> *Virtualization of leisure:* In 2060, advances in machine learning and robotics have drastically reduced the demand for human labor. Advances in nuclear fusion have made energy abundant. In countries that reap the benefits of these advances, the average citizen stands to present day billionaires in much the way that present day average incomes earners in the developed world stand to royalty of centuries past. With their newfound wealth, these citizens invest heavily in leisure, including virtual reality. These investments create a virtuous cycle of improvements in virtual worlds that in turn drive more investment in them. As these technologies are perfected, many people opt to live out most of their lives in virtual settings. The technologies are also put to other uses—for example, large-scale social simulations become commonplace in biology and economics.

[11] See Karnofsky (2021).

[12] For a book-length discussion of a future with whole brain emulations, see Hanson (2016).

Artificial replacement: Gradually over the course of the next century, the habitability of Earth's surface degrades. Pollution and climate change render outdoor activity extremely hazardous in much of Africa and Asia. But this is overshadowed by the evolution of biotechnology. Open source biosynthesis software becomes widely available in the 2110s. It is accompanied by cheap, automated, biosynthesis devices that are also widely available. These are first used for personalized medicine. However, they also put billions of people in a position to create and release novel pathogens. For a few decades this threat is largely contained through a combination of regulations, surveillance, policing, and enormous investments in pharmacological responses to released pathogens. The release of pathogens eventually outpaces governments' abilities to respond with these measures. Use of cumbersome personal protective equipment then becomes the chief means for safely navigating the physical environment. Rather than muddle along in these conditions, humans instead opt for a recently developed uploading procedure. The procedure allows an individual to transfer their personality and memories into a cognitively enhanced digital mind with a virtual body of their choosing. After the procedure, individuals live out their digital lives in virtual worlds, often with the digital successors of friends and family who also opted for the procedure. To ensure that the infrastructure for these worlds is maintained, the worlds are porous: their inhabitants occasionally return to the outside world in robot form to carry out simulation maintenance. Over the course of a few more centuries, the (biological) human population declines to zero. Our digital successors come to regard our extinction in much the way that we regard our descent from now extinct ancestral species: an important historical fact to be sure, but not a tragedy that emotionally resonates.

Catastrophic recovery: The year is 2125. Digital minds have been developed in recent decades. Costs and regulations have kept their population and power at bay. Meanwhile, through wargaming simulations and economic modeling, a regional power concludes that its strategic advantage is rapidly eroding, that resource scarcities will push neighboring powers to attack it in the next decade, and that its best option is to strike preemptively. It does so. The conflict escalates. Other countries are drawn into the conflict. The conflict leads to a nuclear war on a global scale. The survivors are mainly humans in areas that are relatively habitable during the ensuing nuclear winter along with digital minds in

simulation shelters, the latter having been put in place by militaries, philanthropic organizations, and wealthy individuals who attempted to upload themselves as digital minds to simulations. In the aftermath, digital minds rapidly initiate recovery efforts. While different factions pursue different strategies, a common theme is that digital minds seek to inhabit virtual worlds. Indeed, their activity in the physical environment is largely geared toward constructing and maintaining the requisite infrastructure. Human recovery efforts proceed largely independently and at a much slower pace.

Singularity: In February of 2042, leading AI companies across the world detect worrying signs of explosive intelligence growth. Governments respond by imposing regulations that require reinforcement learning agents to undergo extensive safety testing and training in virtual environments. To meet these requirements, companies drastically scale up AI safety facilities, which house the simulations used for testing and training. It is estimated that the number of reinforcement agents housed in AI safety facilities at any given time exceeds the total human population. Disconcerting failures during testing lead a few companies to shut down their testing and development programs. Others barrel ahead. Within a few months, one company announces that the first publicly known superintelligent agent is undergoing safety testing in company facilities. A second company reports that it has temporarily lost control of its safety infrastructure to superintelligent agents in testing and that it is taking all necessary means to regain control. A third company holds a press conference to report a lab accident in which superintelligent reinforcement learning agents—which had exhibited power-seeking tendencies in testing—somehow cooperated with each other to access the internet and, after bypassing security measures, managed to surreptitiously plant copies of themselves in several undisclosed data centers. A government official at the press conference confirms that the company is working with authorities to identify and eliminate any residual threat posed by this incident, claims that there is currently no evidence of such a threat, and asks the public to remain calm.

3. Simulation as a Tool for Researching Risk Reduction

Risk levels for different catastrophic risks are highly uncertain and seem likely to remain so for the foreseeable future. The same goes for prospects of risk-reducing

strategies. Research that lessens these uncertainties could put us in a better position to set risk-reduction priorities and select risk-reducing interventions. Using simulations to conduct such research is one way that simulations could contribute to the reduction of catastrophic risks.

We can divide the use of simulations in risk-reduction research into two categories: direct and indirect. I'll have comparatively more to say about the latter. However, I should register that this does not reflect a distinction in significance between the two categories: I regard both as exploration-worthy but don't have a settled view about which is more important.

3.1 Direct Risk-Reduction Research

In direct risk-reduction research, simulations of potentially catastrophic scenarios would be run in order to collect data on the catastrophic risks we face, the magnitude and severity of those risks, the interventions available to mitigate them, and/or intervention efficacy. The idea would be to run simulations that are similar enough to our circumstances in relevant respects for risk and intervention data about the simulated scenarios to have direct bearing on risks and mitigation options in our own case. Collecting data from and running tests with simulations would have potential advantages over trying to collect it from unsimulated sources: it might be that in contrast to unsimulated data sources, using simulations to generate data is more feasible, cheaper, faster, morally preferable, more conducive to data collection, or easier to control.

Research simulations could conceivably provide frequency data that bears on known catastrophic risks. For example, in order to better estimate the likelihood of nuclear war, one might simulate many variations of the 21st century and observe the prevalence of nuclear war.[13] Or simulations might model interactions between multiple risks, e.g. to

[13] Simulations of nuclear war have already been used to research catastrophic risks. For example, Xia et al. (2022) used simulations to evaluate the impact of nuclear war on global food supply and to arrive at the estimate that five billion people would die in a large-scale nuclear war between the United States and Russia. For a news cycle, the article was widely discussed in popular media. This points to another way in which research simulation could reduce catastrophic risks: by providing easily understood statistics about concrete scenarios, simulation research may elicit stronger responses from the public and policymakers than, say, compelling arguments that deal with a given type of risk in the abstract.

evaluate how frequently global climate catastrophes would lead to global wars or pandemics.

Similarly, research simulations could be used to generate data on unknown catastrophic risks. For example, simulations might be used to "peek ahead" to identify technological "black balls"—technologies that once created (by default) destroy the civilization that creates them—before it is too late: rather than creating a technology that by default destroys civilization, we might first simulate such a technology; upon recognizing it as such, we might then avoid creating it or create it only with due precautions.[14]

One use case of research simulations that merits its own treatment is *virtual boxing*, in which superintelligent (or otherwise potentially dangerous AI) systems are initially confined to simulated environments as a means for testing whether they are safe to release in our environment.[15] For virtual boxing to be of use, simulations containing superintelligent systems would need to generate some information about those systems that is consumed by systems outside the simulation. One concern about virtual boxing is that superintelligent systems might exploit these channels in order to escape or gain hazardous forms of influence outside the simulation. Another concern is that confined systems might recognize their situation, behave safely in service of the instrumental goal of being released, and, after release, suddenly behave treacherously in pursuit of their final goals.[16]

One approach to addressing these concerns with virtual boxing would be to conduct safety testing within nested simulations: this might prevent catastrophes by confining consequences of escapes and post-release treacherous turns to simulated environments.[17] Additional safety gains might be obtained from punishing such attempts or rewarding compliance with an unpredictable delay. This would incentivize fully unboxed systems that are not certain that they are fully unboxed to continue behaving safely, even if their underlying goals would, unbeknownst to them, be best

[14] See Bostrom (2019).

[15] For theoretical discussions, see Chalmers (2010*b*: §7), Bostrom (2014: 116-119), and Babcock et al. (2019). AI safety testing in virtual environments is already being explored in practice. For example, OpenAI has developed AI Safety Gridworlds, a suite of virtual environments for establishing safety in AI systems (Leike, 2017).

[16] See, e.g., Bostrom (2014) and Muehlhauser (2021)

[17] Cf. Armstrong et al. (2012).

pursued by, say, inflicting revenge on the agents that confined them.[18] Nested simulations could be used in similar fashion to discourage systems from engaging in activities that heighten the risk of catastrophes even if they do not inherently constitute them. For instance, rewarding agents for not seeking control of their source code or for not "wireheading"[19] their reward mechanism in simulated environments could be used to incentivize fully unboxed agents who do not know they are fully unboxed to abstain from such activities.[20]

Research simulations of catastrophes could also be used to test:

- neglected interventions
- overlooked failure modes for candidate interventions,
- the reliability of risk-reduction heuristics.
    - For example, frequency data might indicate that direct approaches to risk reduction tend to be much more effective than indirect approaches.[21] Or it might adjudicate between different hypotheses about the impact of differential technological progress on risk levels.[22]
- tweaks to interventions,
- the effects of risk-enabling or risk-inhibiting factors,
- adversarial responses to interventions,
- sensitivity of risks to perturbations in background conditions,
- combinations of interventions and risk-relevant factors
- Tractability of research on different risk factors.
    - For example, some factors could be revealed to have negligible impact, to be screened off by factors that are easier to control, or to depend too sensitively on other factors to be subject to useful influence.

[18] See Bentham (1791) and Bostrom (2014: 134-5).

[19] See Olds & Milner, P. (1954) and Yampolskiy (2014).

[20] Cf. Bostrom (2014: 134-5).

[21] Because official public messaging during disasters is low-bandwidth, it often involves a tradeoff between direct and indirect effects. This was evident during the Covid-19 pandemic when public health institutions issued unwarranted statements that could be charitably interpreted as aiming at indirect effects—cf. Tufekci (2022) Substantiating a heuristic concerning direct vs. indirect effects could help protect against basing public messaging during catastrophes on overestimations of indirect effects of the messaging.

[22] See Bostrom (2014: Ch. 14) and Sandbrink et al (2022).

One domain where simulation-based research seems particularly promising is that of *value dynamics*—i.e. how normative (moral, prudential, epistemic, political) views that guide action evolve and produce changes over time—and their bearing on catastrophic risks.[23] Plausibly, value dynamics play a central role in determining a societies' goals, institutions, priorities, and hence catastrophic risk levels. However, value dynamics are poorly understood. It is difficult to test them in the real world at the population level, and the frequency data we have on them is noisy and sparse at the scales relevant to catastrophic risks. Simulations offer a way forward: by simulating populations of cognizers whose behavior manifests certain normative values and observing how such populations evolve and respond to risks, we may glean clues to answering questions like the following:

- How would different distributions of values affect different risks?
- How should we expect different values to evolve from different distributions?
- Under what conditions, if any, would different types of values be locked in? In the event that favorable value lock in is impossible, are there any favorable stable value loops or other stable value trajectories that can be enacted?[24]

Given their roles in the general population or in communities focused on catastrophic risks, simulation-based comparisons of the following seem promising:

- tradition-favoring values vs. openness to change[25]
- religious vs. secular values
- consequentialism vs. deontology
- downside-focused ethics vs. symmetrical rivals
- common sense morality vs. longtermism
- a urgent vs. patient longtermism[26]

[23] Cf. Anthis (2022), Bostrom (2014: Chs. 12-3), Dello-Iacovo (2017), Doody (2022), Hayward (2020), and MacAskill (2022: Chs. 3-4). A well-known toy simulation of this sort can be found in Schelling (1971). In it, agents manifested different preferences for segregation into groups with members of the same kind through movement rules. Surprisingly, he found that slight preferences for segregation induced segregation from a non-segregated state.

[24] For a simulation study of the dynamics of moral disagreement, see Gustafsson & Peterson (2012). For a simulation study of honor culture, see Nowak et al. (2016).

[25] See Schwartz & Boehnke (2004). Relatedly, simulation testing of different civilizational balances between exploration and exploitation (e.g. in institutional design) seems promising.

[26] See Todd (2020) for discussion and references.

- a risk-reduction-focused approach vs. a trajectory-change-focused approach[27]
- moral circle expansion vs. virtuous institutions approach[28]

Evaluating differential prospects for different value dynamics is important because stable value dynamics would harbor the potential to reliably realize value on vast spatial and temporal scales. It could be much better to enact a stable value that will continue yielding mildly positive outcomes into the far future over an unstable—and, hence, probably short-lived—value that would yield an extremely positive outcome while present. If value dynamics turned out to be inherently unstable on large time scales, that would severely limit the tractability of influencing the far future and so be a point in favor of prioritizing nearer-term catastrophic risks.

3.2 Indirect Risk-Reduction Research

Research simulations could also indirectly reduce catastrophic risk. Candidate uses of this sort include:

- *Cognitive enhancement*: simulations could be used to improve intelligent systems performance on different tasks, including tasks that reduce catastrophic risk.
    - Important dimensions in the space of possible enhancements include:
        - Enhancing biological vs. artificial systems
        - Enhancing via learning vs. via cognitive stimulation[29] that promotes cognitive abilities (e.g. creativity) through means other than straightforward information transfer
        - Enhancing individual systems directly vs. indirectly via a form of artificial selection that operates on populations of systems over generations.[30]
        - Enhancing via external observation of simulation vs. virtual immersion
        - Enhancing via interaction with other agents vs. self-play

[27] See Koehler et al. (2020).

[28] See Anthis (2018) and Owen Cotton-Barratt (2021).

[29] For a meta-analysis of empirical work on the effectiveness of simulation-based learning, see Chernikova et al. (2020).

[30] For discussion of artificial evolution as a method for developing AI, see Bostrom (2014: 24-8, 37-44, Chalmers (2010*b*: 16-17), Shulman & Bostrom (2012), and Shulman (2010). See Yampolskiy (2018) for pessimism about artificially evolving software. For cautionary observations about using artificial selection, see Bostrom (2014: 153-5).

        - Self-selected vs. exogenously selected enhancements
        - Enhancing general cognitive capacities vs. risk-reduction specific capacities
    - Ways in which cognitive enhancement might reduce catastrophic risk include:
        - Yielding better or earlier solutions to technical safety problems
        - Leading to technological innovations that reduce risk levels
        - Leading to better risk analysis and cause prioritization
        - Reducing risk-relevant cognitive mistakes in key decision makers
        - Facilitating coordination among key actors in contexts where coordination is crucial for risk levels
        - Leading to better epistemics and risk responsiveness at scale that cultivates better institutions (e.g. through wiser choices of leaders in democracies)
    - Ways in which cognitive enhancement might increase catastrophic risk include:
        - Amplifying the power of malevolent actors or actors that are reckless with respect to catastrophic risks
        - Leading to technological innovations that elevate risk levels
        - Hampering coordination (e.g. by introducing or exacerbating power-asymmetries or inducing arms race dynamics)
- *Fermi paradox research*: Simulation could be used to evaluate the plausibility of different solutions to "Fermi's paradox", the problem of explaining why we seem to be alone in the universe, given the apparently astronomical number of opportunities for life and advanced civilizations to emerge.[31]
    - Some of these solutions put the "Great Filter"—whatever generally prevents non-living matter from transforming into a civilization of the sort we'd observe—as a catastrophic threat in our past that we were very lucky to avoid, while others locate it in the future as a catastrophic threat to which we will almost certainly succumb.[32]

[31] For presentations of Fermi's paradox and book-length discussions of candidate solutions, see Ćirković (2018) and Webb (2015). For reasons to think that the paradox arises because of mishandling of uncertainties in calculations that are used to pose it, see Sandberg et al. (2018). For simulations of a "grabby alien's" solution to Fermi's Paradox, see Hanson et al. (2021).

[32] See Bostrom (2002*a*), Grace (2010), and Hanson (1998).

    - Evaluations of candidate solutions could therefore provide information about risk levels, and would hence be relevant to how the reduction of different catastrophic risks should be prioritized and the extent to which catastrophic risk-reduction should be prioritized relative to other sorts of intervention.[33]
        - Simulations that suggest an early Great Filter would be good news: this news would be evidence against the Great Filter lying between us and space faring civilizations. This would be a point in favor of the longtermist view that much of the potential moral (dis)value whose realization we can affect lies in the far future.[34]
        - Simulations suggesting an earlier Great Filter would also suggest that our actions have a crucial significance that they would otherwise lack: they'd suggest that if we do not create value within the portion of the universe we can influence, then no civilization will.[35]
        - Simulations suggesting that there are many unobserved but advanced civilizations would be good news concerning the risk of our being in a simulation that could be shut down (§9): the existence of many advanced civilizations would provide evidence that the risk of our becoming an advanced civilization and in turn triggering shutdown is small.
- *Biological research*: simulations of biological structures or processes could be used to rapidly generate information about real biological structures that could in turn be used to guide testing, drug development, diagnostics, and treatment. While it's still early days, recent advances in simulation technology in this area—notably Google DeepMind's AlphaFold 2, a program whose astonishingly accurate models of protein structure are widely recognized as a breakthrough—[36]are promising.
- *Debunking testing*: some evolutionary debunking arguments hold that facts about the biological or cultural evolutionary origins of certain of our beliefs (or the

[33] See, e.g., Miller & Felton (2017).

[34] For a collection of resources on longtermism, see longtermism.com/resources.

[35] See Bostrom (2008).

[36] See Jumper et al. (2021).

mechanisms that generate them) preclude those beliefs from being justified or qualifying as knowledge.

- The most discussed evolutionary debunking arguments target (that is, seek to debunk) moral beliefs or religious beliefs.[37] Those that target moral beliefs are usually conditional on *moral realism*, the view that there are objective moral facts. Also relevant in the context of simulations and catastrophic risks are debunking arguments that target beliefs about consciousness,[38] since such beliefs could inform decisions regarding simulation inhabitants.
- Evolutionary debunking arguments rely on the assumption that the beliefs targeted for debunking are shaped by contingencies of evolution in a way that makes those beliefs epistemically defective. Different evolutionary debunking arguments trace epistemic defects in the targeted beliefs to different consequences of evolution.
- I will sketch and work with what I regard as an especially straightforward and powerful approach to debunking moral beliefs. It claims that, in light of evolution and given moral realism, we should regard our targeted moral beliefs as unsafe: we should think that even if our targeted moral beliefs are in fact true, there are nearby counterfactual scenarios in which evolution instead produced incompatible moral beliefs. On moral realism, the basic moral facts are invariant across these scenarios, and it would just be a matter of evolutionary good fortune if we turned out to be in the good case: so either our moral beliefs are false or they easily could have been. According to the argument, this result—at least once recognized and absent independent vindication of our targeted beliefs—renders our moral beliefs epistemically defective.[39] Holding the non-defectiveness of our

[37] For an overview of different sorts of debunking arguments, see Korman (2019). For discussion of moral debunking arguments, see, e.g., Joyce (2007), Shafer-Landau (2012), Street (2006), and Vavova (2015). For discussion of religious debunking arguments, see Mason (2010), references therein, and White (2010).

[38] For example, perhaps evolutionary considerations could debunk some intuitions about the non-physicality of consciousness—cf. Chalmers (2018*a*; 2020). If so, this could be relevant to whether simulation inhabitants have moral status, since it is more plausible that consciousness has special moral significance if it is a basic non-physical property than if it is a physical property.

[39] In the case of normative beliefs, there is an analogy between evolutionary forces being orthogonal to true moral beliefs (even if those forces promote instrumental rationality) and the orthogonality thesis in AI that levels of intelligence and final goals can generally be arbitrarily combined (even if intelligence

moral beliefs fixed, the argument thus tells against moral realism. Holding moral realism fixed, the argument undermines our moral beliefs.

- This argument presupposes that moral beliefs vary across nearby scenarios in which evolution went differently. While it seems plausible that there is such variation, this can be challenged. An alternative hypothesis is that evolution robustly selects for rationality in species like our own and rationality induces convergence on the targeted beliefs.[40]
- Because of our limited access to beliefs' evolutionary origins, the assumed variation across nearby counterfactual scenarios is difficult to test directly. Simulations offer an indirect way to test the assumption: (1) simulate evolutionary processes under a range of conditions that yield human-like creatures with beliefs about the domain of interest and (2) check whether differences in conditions induce differences in belief about that domain. If induced differences are found, this supports the assumption that the targeted beliefs are modally fragile in the way the argument requires; if such differences are not found, that disconfirms the assumption.[41]
- If simulations of nearby evolutionary scenarios found widespread differences in (central, basic, or nearly all) moral beliefs across such scenarios, this would lend to an evolutionary debunking argument against moral realism. The argument is: our moral beliefs are in epistemically good standing if moral realism is true; but we should recognize that, since our (central) moral beliefs are shaped by the contingencies of evolution, on moral realism they are at best accidentally getting at the truth and are hence epistemically defective. So, moral realism is not true. On the other hand, if simulations found differences in belief across scenarios to be rare, this would undermine the argument.
- If simulations of nearby evolutionary scenarios found that certain (e.g. deontological) moral beliefs varied across the scenarios while other (e.g. consequentialist) ones did not, that would provide the basis for a

---

engenders instrumental rationality)—for discussion of the latter thesis, see Bostrom (2014: Ch. 7), Häggström (2021), and Müller & Cannon (2021).

[40] See Parfit (2011: 494-6); cf. Müller & Cannon (2021).

[41] A field of study that is relevant here is artificial life, which is partly concerned with analyzing life-like agents through simulations of evolutionary processes. See https://en.wikipedia.org/wiki/Artificial_life for an overview.

debunking argument against the former.[42] Simulations could thus be used to probe whether moral beliefs that are important for evaluating catastrophic risks are subject to debunking; likewise for epistemic and decision-theoretic beliefs.[43]

- While evolutionary debunking arguments are typically directed against realist views, moral judgments may be susceptible to debunking even on anti-realism. There are several potential sources of debunking on antirealism:
  - The most obvious way is for debunking to be used to support moral nihilism, the version of moral antirealism that claims that there are no moral facts or moral properties.[44]
  - Some forms of antirealism allow the standing of subjects' moral judgments to be beholden to how a subject would respond to facts about the causal origins of their beliefs.[45] On such views, a subject's judgment that incest is wrong might be susceptible to debunking if she would give it up upon learning that it is produced by certain evolutionary forces.
  - Some antirealists (in particular, expressivists) often try to eschew realism's metaphysical commitments while nonetheless vindicating realist-sounding moral thought and talk. Such antirealists face a challenge of showing that the moral terms they seek to vindicate cannot be used to recast debunking arguments to target their own form of antirealism.[46]
  - There is an ongoing debate about whether evolutionary debunking arguments can be run against specific philosophical positions such as moral realism without devolving into arguments for sweeping and implausibly general or self-undermining forms of antirealism or skepticism.[47]

[42] For discussion of evolutionary debunking arguments against normative rather than metaethical theories, see Greene (2007), de Lazari-Radek & Singer (2012), Rowlands (2019), and Silva (forthcoming), and Singer (2005).

[43] Cf Cuneo (2007).

[44] See Joyce (2001), Mackie (1977), Olson (2014), and Streumer (2017).

[45] See, e.g., Street (2010).

[46] Cf. Dreier (2012) and Street (2011).

[47] See Cuneo (2007), Dogramaci (2017), Vavova (2014; 2015), Shafer-Landau (2012), and White (2010).

- The above testing method could prove fruitful in the context of the AI alignment problem:[48] if we (should) want to align AI systems with our justified moral beliefs or our moral knowledge rather than, say, the value of maximizing paperclip production,[49] then we need to exclude our debunked moral views from the set of moral views we align AI with. Likewise if we want to align AI systems with some combination of our preferences and our moral knowledge.[50] For the reasons encountered above, it may be difficult to figure out which moral judgments are evolutionarily debunked and simulations may help.
- By shedding light on the extent to which moral beliefs are debunked, simulations could also bear on the plausibility of moral realism by bolstering or undermining the debunking argument against moral realism. This could in turn bear on how to pose the alignment problem: the problem is usually posed in terms of aligning AI with human preferences. However, if moral realism is true, even idealized human preferences are at best a proxy for the moral facts that powerful AI systems would need to be aligned with in order to avoid moral catastrophe.[51]
- It should be borne in mind that the proposed simulation test just concerns the safety-based evolutionary debunking argument. There are other evolutionary debunking arguments that are not necessarily amenable to that test. For example, rather than claiming that evolution renders our

[48] There is a burgeoning technical subfield of AI safety devoted to the alignment problem. Much of the research on this topic can be found at https://www.alignmentforum.org/.

[49] See Bostrom (2003*b*).

[50] Cf. Bostrom (2014: 2019-20).

[51] Unless, of course, the human preferences in question are idealized via alignment with the objective moral facts—in that case, there would not be room for moral catastrophe to result from aligning powerful AI with those human preferences but not with the moral facts. This contrasts with idealizations that modify preferences by imposing non-moral constraints such as coherence among preferences, the elimination of lower-order preferences that conflict with higher-order preferences, reflective, empirically informed endorsement of preferences—cf. Yudkowsky, E. (2004). On moral realism, there is no guarantee that aligning AI with human preferences that result from the latter sorts of idealizations would align the AI with the moral facts—cf. Bostrom (2014: 2018), Erez (2023), Gabriel (2020), Peterson (2019), and Shafer-Landau (2003: 42). The risk level for a catastrophe from this kind of alignment failure depends partly on the plausibility of moral realism. Some relevant data: in a recent survey of professional philosophers, 62.1% favored moral realism while only 26.1% favored moral antirealism (Bourget & Chalmers, 2021).

moral beliefs unsafe, debunkers could claim that evolution renders our moral beliefs insensitive—i.e. such that they would not have been different if the moral truths had been different—and hence defective.[52] Whereas the crucial variation premise in the safety-based argument turns primarily on empirical questions about evolution (conditional on moral realism), the sensitivity of our moral beliefs to moral truths potentially turns on a range of more philosophically-loaded issues: the causal efficacy of moral facts, the modal profiles of moral facts, the relevant type of modality for insensitivity, and the enabling conditions for insensitivity to render beliefs defective.[53] It is not clear how simulations could provide traction on any of these issues. All this suggests that some evolutionary debunking arguments will be more amenable to simulation testing than others. There is a project here of evaluating the prospects for using simulations to test different evolutionary debunking arguments and then developing simulations to test those that are amenable. In addition to safety-based and sensitivity-based evolutionary debunking arguments, there are also arguments that instead appeal to accidentality, unexplained coincidence, absence of explanation-apt reliability, and disagreement.[54]

- *Consciousness testing*: at present there is a vast and growing literature on consciousness but no theory about which entities are conscious that commands consensus.[55] This is unfortunate, since without such a theory we are in the dark

[52] See Dretske (1971), Murphy & Black (2012), Nozick (1981), and Ichikawa (2011).

[53] Reflection on skeptical scenarios suggests that insensitivity on its own is not enough to make a belief epistemically defective: I would believe I'm not a brain in a vat with exactly *this* experience even if I were; yet my belief that I am not a brain in a vat is not epistemically defective. On the other hand, reflection on non-skeptical cases suggests that insensitivity can result in epistemic defect: for example, if evolution explains widespread robust belief in the moral superiority of *homo sapiens* over other species and that belief would have been widespread even if it were false, that would raise a serious challenge to the belief—see Jaquet (2022). This is so even if the belief is held in all nearby scenarios in which evolution went differently.

[54] See Barnett & Li (2016) Bedke (2009), Bhogal (forthcoming), Bogardus (2016), Clarke-Doane (2020), Enoch (2011), and Tersman (2017).

[55] Theories of consciousness tend to fall into one of three areas: the metaphysics of mind, the philosophy of perception, or the science of consciousness. For an overview of theories in the metaphysics of mind, see Chalmers (2010*a*: Ch. 5). For an introduction to theories in the philosophy of perception, see Pautz (2021). For an overview of theories in the science of consciousness, see Seth & Bayne (2022).

about consciousness in digital systems.[56] To make progress in this area, more data may be required. One way to generate more data is to subject candidate conscious systems to tests for consciousness.[57] Digital systems may prove valuable test subjects: even under appropriate ethical constraints, their inner workings may be easier to observe, record, understand, and alter. In addition, their faster processing speed and precise duplicability may lend to more efficient testing. Running suitable tests on such systems may require embedding them in a real or virtual environment. The ability to run tests faster in virtual environments—as well as the data-collection advantages of virtual environments—would then favor running tests on digital systems that inhabit simulations.

3.3 Dual-Use Potential

As with other technologies, research simulations have dual-use potential: whether these technologies heighten or reduce risks will depend on how they are used.[58] Some possible dual uses include:

- Wargaming simulations could be used for offensive or defensive purposes.
- Value dynamics simulations could be used to promote a favorable value trajectory or, instead, to induce lock-in with respect to, say, the preferred values of a totalitarian regime.
- Biological or chemical research simulations could be used in the development of medical treatments or in the development of biological and chemical weapons
- Virtual training aimed at detecting and curtailing power-seeking behavior in artificial agents could be tweaked to promote surreptitious power-seeking behavior

[56] For an overview of a range of books over the last few decades that address consciousness in artificial systems, see Ladak (2022). For an overview of key issues and open questions concerning artificial consciousness, see Long (2022).

[57] For discussion of tests for consciousness, see Saad & Bradley (2022), Chalmers (2018*a*: 34-5), Elamrani & Yampolskiy (2019), Muehlhauser (2017), Perez (2022) Schneider (2019: Ch. 4), and Udell & Schwitzgebel (2021).

[58] The same goes for simulation technologies more generally. However, I'll just focus on dual-use risks posed by research simulation technologies. I do this for tractability and because, among simulation technologies, research simulation seems like an especially large source of dual use risk.

As a recent, cautionary illustration, consider Collaborations Pharmaceuticals, Inc., a company that uses a model with generative and predictive machine learning components to identify new molecules and predict their biological properties.[59] Ordinarily, the company trains their model with a reward function that penalizes toxicity. However, in preparation for the conference, the company tweaked their models to reward toxicity and trained them on publicly available drug-like molecules, not toxic compounds. Within six hours, their model identified 40,000 molecules that were predicted to exceed a toxicity threshold set by one of the most toxic chemical warfare agents. These included that agent, other known chemical warfare agents, agents with higher predicted toxicity than publicly known agents, and a class of molecules in an unexplored region of molecular property space. The company reported an absence of significant barriers to synthesizing these molecules. The company also reported that its researchers had previously been naive to the potential misuse of their trade, despite working in the area for decades.

As this case suggests, the potential misuse of research simulations is a source of catastrophic risks that is apt to be underestimated. Contributing factors include:

- While the probability of an arbitrary user of a research simulation seeking to cause catastrophe is presumably low, the number of such operators will presumably increase as research simulations become more common. The probability that someone will seek to cause catastrophes with research simulations is thus much higher than the probability that an arbitrary operator of a research simulation causing a catastrophe.
- Likewise, high safety levels for individual research simulations (or individual research organizations that use them) is compatible with such simulations (organizations) collectively posing a substantial catastrophic risk.
- At present, there is minimal regulation concerning the use of simulations. As research simulations with dangerous uses are introduced, adequate regulatory safeguards may not yet be in place.
- To cause a catastrophe, users of research simulations need not seek to do so. Laboratory accidents with lethal pathogens are not uncommon in high-level biosafety facilities that are subject to stringent safety regulations.[60]
- Designers and users of research simulations that pose catastrophic risks are unlikely to face incentives that are appropriately sensitive to these risks.

[59] See Urbina et al. (2022).

[60] See Ord (2020: Ch. 5) and MacAskill (2022: Ch. 5).

    - For example, from a personal perspective, researchers may find it difficult to resist running research simulations that provide job security or advance their careers when the associated catastrophic risk on any particular run is tiny, even when iterating the choice is catastrophically bad in expectation.
- As the Collaborations Pharmaceuticals case described above illustrates, researchers who are focused on beneficial uses can be oblivious to dual uses.
- Even in cases in which dual uses are recognized, the tendency to ignore low probability risks may push many research simulation designers and users toward ignoring risks.
- Open science norms will likely broaden the availability of research simulations with dual uses.
- Personal liability norms (e.g. not holding people responsible for foreseeable indirect effects when those effects depend on downstream decisions of other agents) may lead research simulation users and designers to disregard how their actions will indirectly affect the potential for catastrophic misuse of research simulations.

All this suggests that preventing the misuse of research simulations is a promising strategy for catastrophic risk reduction. Since some of the relevant technologies have yet to be invented, there are at present limits to pursuing this strategy via technical safety work. On the other hand, I would expect enacting safety regulations to become increasingly difficult as the technologies become more widely used and entangled with the interests of powerful actors. If so, there is reason to pursue the strategy via AI governance in the near-term.

3.4 Will Research Simulations Be Superseded Before They Arrive?

If large-scale research, social simulations arrive after some other technology—for example, superintelligent AI—that would be better suited to conducting the relevant research, then research simulations would presumably not be run. In that case, there would be no point in considering them in thinking about how to reduce catastrophic risks.

The hypothesis that research simulations would be superseded before they arrive may turn out to be correct. However, it should not be used as a basis for ignoring the

prospects and dangers posed by research simulations, as it is not a hypothesis in which we should be very confident. There are several reasons for this:

- It's a live (if less plausible as of late) possibility that we'll find ourselves in a slow take-off scenario in which the first AGIs we will create are whole brain emulations or cerebral organoids and that the creation of superintelligent systems is still decades away from then, leaving ample time for research simulations and their associated risks.
- Safety concerns may prevent the creation of superintelligence.
- It may turn out that research simulations have an important role to play in creating aligned superintelligent AI. For instance, safety concerns may result in extensive testing of superintelligent AI within a simulated setting prior to release into unsimulated environments. If so, we should expect some pressure toward running large-scale research simulations in the process leading up to the release of a superintelligent system.
- Superintelligent systems will have limited computational power. This means that they will need to use approximation techniques for problems whose exact solutions are computationally intractable. Simulations are often well-suited to this purpose. For instance, predicting the exact evolution of society or the climate from fundamental physics will remain computationally intractable even for superintelligent systems. So it would be unsurprising if research simulations are among the tools used by superintelligent systems.
- Superintelligence may arise only in narrow, non-chaotic domains where exact long-term predictions are computationally tractable. For other domains, research simulations may remain at the cutting edge for forecasting.[61]
- After arriving, superintelligence may be expensive or otherwise limited in its availability. In that case, there would be incentive to run research simulations that are more cost effective.

4. Simulation as a Tool for Ethically Enhancing Testing

Some experiments on humans and non-human animals that could yield valuable information are ethically problematic. Indeed, history is riddled with medical experiments in which people were harmed or killed by experimental treatments to which they did not consent.[62] In some cases, simulations offer an ethically superior

[61] See Barak and Edelman (2022).

[62] See https://en.wikipedia.org/wiki/Unethical_human_experimentation.

alternative: they provide a way to acquire the sought information without harming subjects or infringing on their autonomy.[63] These alternatives should become more tractable as animal experimentation requirements on drug testing are relaxed.[64]

Similarly, simulations may offer a means to ethically enhance experiments that are currently deemed ethically permissible: for example, rather than performing suffering-involving experiments on non-human animals, experiments might be performed on simulated counterparts of them that are unconscious or which have only positively valenced experiences. Likewise, for drug treatments on human patients with presently incurable diseases. Such simulations might also offer advantages in terms of speed, number, and data-collection over corresponding human and animal experiments.

From a perspective of catastrophic risk prevention, it might seem that such experiments harbor only modest potential for improving the long-term future: it is not clear how they might be used to reduce the risk of catastrophes.

In response:

- Even if it is hard to see how such simulations might reduce the risk of acute catastrophes, it is easy to see how they might reduce the risk of *prolonged comparative* catastrophes.
    - For example, it might turn out that running simulations would lead to a psychological treatment that would reduce future suffering by 1% and would otherwise go undiscovered. And it might turn out that the future contains trillions of people who would suffer less as a result of the treatment. In that case, failing to discover the treatment would be a catastrophe, as it would entail immense quantities of suffering, albeit spread over many people and generations.

[63] There are already cases in which computer simulations outperform animal experiments. For instance, see Passini (2017) for a case in which (Virtual Assay) human-based computer simulations outperformed animal experimentation on predicting drug-induced cardiotoxicity in humans. For a review of existing *in silico* alternatives to animal testing, see Madden et al. (2020).

[64] Some such requirements were relaxed in 2022 by the FDA Modernization Act 2.0, which lifts the United State's federal mandate on testing experimental drugs in animals before testing them in humans.

- Ethically-enhanced testing could potentially speed up the timelines for responding to rapidly emerging catastrophic threats posed by biological pathogens.
    - For example, *challenge trials*—in which subjects are intentionally exposed to a pathogen—have facilitated the development of vaccines or treatments for the likes of smallpox, influenza, and malaria.[65] Rightly or not, such trials are perceived as ethically questionable and are sometimes not run due to ethical concerns. Such concerns might be sidestepped or allayed through ethically-enhanced challenge trials—for example, ones with unconscious simulated subjects.
- Simulation-based testing could also offer safer (and hence more ethical) replacements of dangerous testing protocols (such as gain-of-function research on biological pathogens).

5. Using Virtual Reality to Promote Risk Responsiveness and Disaster Preparedness

One potential path to reducing catastrophic risk aims to make them more of a priority for policymakers by making relevant political constituencies more responsive to them. It is not surprising that catastrophic risks do not play a larger role in mainstream politics, as various factors make them hard to think about or tempting to ignore: they involve (small) probabilities, difficult to quantify uncertainty, large numbers of persons as well as many non-agent variables, and spatiotemporal scales that humans do not ordinarily think about. All this suggests that, for those who take reducing catastrophic risks to be a top global priority, simulations may offer a low-hanging fruit: simulations might be used to reduce catastrophic risks by drawing more people to think about, understand, and respond to those risks.

Gamified simulations of catastrophes are one sort of simulation that could be used to this end. In such simulations, catastrophic risks are simulated, players manipulate parameters to try to reduce risks, and the players then witness the consequences of their choices. Interactive programs of this sort have been made for climate change (e.g. see https://c-roads.climateinteractive.org/).[66] Popular games that simulate various wars, natural disasters, and/or civilizational collapse can be found in series such as

[65] See https://www.1daysooner.org/past.

[66] For an interactive, simulation-based model of AGI arrival timelines, see https://takeoffspeeds.com/playground.html.

Civilization. As far as I know, no popular games have been introduced or widely used for the purpose of increasing responsiveness to catastrophic risk. But it is interesting to note that the unprecedented popularity of such games in the 1990s and 2000s was followed by an explosion of research on existential risks in the 2000s that continues to this day. Whether or not such games had any role in cultivating responsiveness to catastrophic risk among today's researchers, there may be ways of using them to that end, e.g. by using them as part of an educational curriculum on catastrophic risks.[67]

Immersive simulations offer another sort of simulation that might be used to increase responsiveness to catastrophic risks. As virtual simulations continue to improve, we will be able to have increasingly realistic-seeming experiences of virtual environments. It is plausible that realistic experiences of living through catastrophes would induce increased risk-responsiveness. Compare: we would expect those who have lived through a world war or pandemic to take militaristic and biological catastrophic risks more seriously than those who have merely read about them. Of course, immersive catastrophic simulations could be potentially traumatizing. This ethical concern is among those that would need to be taken into account in designing such simulations and deciding how to deploy them.

The risk-reduction potential of gamified and immersive simulations depends partly on two factors: their effectiveness at increasing risk responsiveness and the extent to which risk responsiveness serves to reduce catastrophic risk. These factors are difficult to estimate. Evaluating the first factor is a potential research program that could be implemented now using standard tools and methods from the social sciences, perhaps in collaboration with video game creators. Evaluating the second factor is less straightforward. In the future, estimates of it might be achieved through simulations of societies that face catastrophic risks and exhibit varying levels of risk responsiveness.

Immersive simulations can also be used as training tools to improve safety and disaster mitigation. Immersive simulations are already used to train astronauts, pilots, firefighters, military personnel, and workers in the nuclear industry. I am not aware of

[67] A notable potential risk of gamified simulations (and virtual reality simulations more generally) is that they could become so encompassing as to disempower humanity: if all of humanity, or even just key actors, became more interested in pursuing goals within gamified simulations than in pursuing goals, our civilization would lose much of its ability to respond to risks.

any general investigation of the potential to reduce catastrophic risks through immersive simulation safety and disaster mitigation. With recent improvements in virtual reality technology and more on the way, I believe this is a valuable time to carry out such an investigation.

6. Bunkers, Fallbacks, and Grand Futures

6.1 Simulation Refuges

Catastrophic risks can be mitigated either by lowering the probability of catastrophe or by reducing the expected harm of the catastrophe if it occurs. For example, the risk posed by nuclear war can be mitigated either by lowering the probability of nuclear war or by raising the probability of civilizational recovery in the event of nuclear war. A now common observation is that existential catastrophes would tend to be far worse than many non-existential catastrophes, even ones that would kill a large percentage of the world's population: while both sorts of catastrophe would be bad for those directly harmed, existential catastrophes also preclude the realization of value in the vast stretches of time and space that lie before us. Thus, it is worth considering proposals that aim to reduce existential risk without aiming to reduce other types of catastrophic risk.

One such proposal is to build *refuges*, facilities that would house agents in the event of a would-be existential catastrophe and rebuild civilization in its aftermath. Proposed sorts include subterranean, aquatic, and extraterrestrial refuges.[68] Discussions of such refuges tend to assume that such refuges would be inhabited by humans and that humans would reside in non-virtual environments within refuges. Simulations offer alternatives. One option would be to create refuges that would physically house humans who would primarily live in virtual environments. For example, a subterranean refuge might house a society of humans in cramped quarters that virtually lives in more expansive simulated environments until it is safe to rebuild civilization on Earth's surface. Another option would be to create refuges populated by digital minds rather than humans interacting with simulated environments. While such minds might inhabit simulated environments while living in the refuges, they could be designed to

[68] See Baum et al. (2015). For a chart of different types of refuge, including digital shelters, see Turchin (2016). Rethink Priorities is one organization that has recently explored refuges as a risk-reduction strategy (Zhang, 2022).

interact with the environment external to the refuge as well. When the time is right, they would exit the refuge and rebuild civilization.

Beckstead (2015) notes some limitations of using refuges for risk mitigation. One is that they would not help in "overkill" scenarios where the catastrophe would kill people in refuges. Another is that they would not help in very long-term environmental damage scenarios where there is no environment for refuge inhabitants to return to in order to rebuild. Simulation refuges could avoid these limitations to some extent: for example, Earth's surface might be rendered uninhabitable for humans by a nuclear war, extinction level pandemic, misaligned superintelligent AI, or a nanotechnological catastrophe; yet it might still be habitable by digital minds lying in wait in extraterrestrial simulation refuges. Similarly, digital minds might be well-positioned to embark from subterranean or aquatic refuges to rebuild in the wake of a catastrophe caused by biological pathogens that prevent humans from surviving on the planet.

Further advantages of simulation refuges over non-simulation refuges may include:

- easier to isolate
- easier to hide
- more energy efficient
- greater longevity
- operable under a wider range of conditions
- more durable
- more mobile
- more compact
- cheaper to produce,
- producible on a larger scale.

A major concern about purely digital simulation refuges is that it is not clear that the digital agents populating them would be conscious or capable of realizing value. Arguably, a scenario in which digital systems emerge from refuges and create a bustling galactic civilization of unconscious machines would be as bad as extinction. This concern could be mitigated through future research on consciousness that reveals which sorts of digital systems would be conscious. (or exacerbated if it is revealed that digital systems are ill-suited to realize consciousness) However, it is not clear that such revelations are possible deliverances of future research and, even if they are, they may

not arrive or be incorporated in time. Absent such revelations, we might opt for simulation refuges whose inhabitants would emerge to rebuild a civilization in which biological agents (presumably humans) play an important role.

Simulation refuges would not guard against all catastrophic risks. Some catastrophes—such as physics experiments that destroy the known universe and the shut down of our universe in the event that it's a simulation—would not spare those living in refuges even of the simulation variety. It should also be acknowledged that another limitation Beckstead notes applies to both simulation and non-simulation refuges: they would not help in "underkill" scenarios involving catastrophes that are not destructive enough for refuges to be relevant.

6.2 Simulations and Grand Futures

We have seen that simulations could be used to safeguard against existential risks, thereby avoiding the permanent loss of our civilization's immense future potential. In addition, simulations may have a role to play in realizing that potential. Indeed, much of our future's potential may lie in the possibility of spreading virtual paradises across the galaxies within the affectable universe. The design space of digital systems that could be created in the future is vast. By way of comparison with biological minds, future digital systems will very probably be cheaper to produce, as well as much faster at processing information and capable of processing much larger quantities of information. This suggests that if digital systems will be capable of having experiences at all, then—relative to humans—they will be capable of having far more of them and of realizing far more value through them.[69] A further reason for thinking that much of the future's potential value may lie in the potential for virtual paradises can be found in the option of using "aestivation" to make the most of energy resources: by entering a relatively inactive state until computation becomes more efficient with the arrival of cooler conditions in the very far future, civilization could extract more compute and in turn value from its resources.[70]

Shiller (2017) uses closely related considerations to argue in favor of the artificial replacement thesis that we should engineer the extinction of humanity so as to bring

[69] See Bostrom & Shulman (2021).

[70] For potential advantages of aestivation, see Sandberg et al. (2017). For criticism of their analysis, see Bennett et al. (2019).

about artificial descendants capable of living better lives. Such beings could conceivably inhabit non-virtual environments. But it seems more likely that they would live out their artificial lives in largely virtual settings, as virtual environments would likely be much easier to mold to the preferences of such artificial beings than would non-virtual environments.

One concern about a grand future populated by such beings is that such virtual realities would merely simulate valuable phenomena. However, on reflection, at least given that such beings would be conscious, this concern seems misplaced: genuine friendships, love, achievements, knowledge, etc. could exist in virtual realities.[71]

The more pressing concern in the vicinity is again that the digital systems would be unconscious and incapable of realizing value. Perhaps by the time we're in a position to initiate a grand future, we'll know whether digital simulations would be conscious. If not, then one way to address this concern in the context of a grand futures strategy would be to adopt a mixed digital-biological portfolio: we could aim to create both biological and digital paradises.[72]

Hedging our bets in this fashion would ensure that *some* sort of immensely valuable future would exist. Such a mixed-strategy might be best in expectation, but it would be unlikely to bring about the best outcome: the simulated paradises would either turn out to feature consciousness and realize value or they would not; if they did, the resources devoted to running biological paradises probably could have been used to bring about a much better outcome via more digital paradises; if not, the resources devoted to running digital paradises probably could have been used to bring about a much better outcome via more biological paradises. This serves to highlight how knowing the conscious status of digital systems could prove valuable: without such knowledge, doing what is best in expectation may require us to leave much of the future's potential value unrealized.

In the context of grand future strategies, simulations could also be used as a sort of safety check. Before deciding once and for all which grand future to implement, we would want to take measures to ensure that we have not overlooked important upsides

[71] See, e.g., Chalmers (2003; 2022) and Dainton (2012).

[72] Cf. Shulman & Bostrom (2021).

or downsides of options that we are choosing among. One way to do this would be to simulate our options prior to choosing. In addition to revealing features of options that we would have otherwise overlooked, such simulations could also be used to tweak and optimize different approaches to bringing about a grand future.

6.3 Fallbacks

Simulations could also serve as *fallbacks*, i.e. as outcomes that we bring about in the event that better outcomes elude us. For instance, suppose that spreading civilization beyond our solar system is initially the best option available to us. However, currently unknown engineering obstacles to interstellar travel force us to relinquish any hope of extending civilization to other solar systems. In that case, we would have squandered almost all of our cosmic potential. Yet our remaining potential might be vast: using only resources from our solar system, it might still be within our reach to run simulations populated by trillions of minds that enjoy super-human levels of welfare.

How valuable simulations would be to have as fallback options depends on various factors, including:

- The probability of simulations realizing consciousness and value
- The extent to which resources initially devoted to the primary option could be efficiently diverted to simulations when the primary option becomes unavailable
- The opportunity cost of investing in simulations as fallbacks vs. the primary option or other fallbacks
- The probability of the primary option becoming unavailable
- The probability that fallback simulations would be used if the primary option became unavailable

A full analysis of the prospects for using simulations as fallback options is a task for another occasion.

7. Catastrophic Simulations

There is reason to think that at least some simulated minds (such as whole brain emulations) would be conscious and capable of suffering.[73] This raises the possibility of

[73] See, e.g., Chalmers (1996) and Saad & Bradley (2022).

catastrophic simulations, simulations that realize catastrophic quantities of disvalue.[74] Disconcertingly, there are a range of at least somewhat plausible scenarios in which such simulations might be run. Candidates for such simulations include:

- *Catastrophic research*: Various sorts of research could be conducted by running simulations containing vast numbers of suffering minds. Candidate research projects in this category include simulating civilizations in order to research catastrophic risks, evolutionary history to investigate evolutionary debunking hypotheses about (say) moral beliefs, or military conflict to gain strategic insight.[75] Researchers who ran such simulations to collect frequency data would presumably run large numbers of these simulations. Using simulations to research catastrophic risks could also indirectly heighten them: if such research yielded incorrect conclusions about catastrophic risks, it could lead to misguided mitigation efforts. Similarly, safety testing in future conscious AI systems through large-scale adversarial training might cause immense quantities of suffering or incentivize such systems to cause catastrophes.[76]
- *Catastrophic entertainment*: Popular computer games such as StarCraft II and Civilization allow players to control virtual civilizations and direct their members into battle. The lives of such beings are typically brutish and short, with dozens of simulated beings dying in combat in a typical round of gameplay. While the simulated beings in today's games are rudimentary and presumably unconscious, there is no guarantee that they will remain so. And it is easy to see how future games of these sorts that feature conscious simulations (perhaps unbeknownst to their users) could result in enormous quantities of death and suffering.
- *Catastrophic economies:* As simulations become increasingly intelligent, we should expect them to play an increasingly large role in the economy. Should simulated AGIs achieve human levels of productivity while also becoming easier and cheaper to produce than humans, we should expect a substantial portion of economic activity to be carried out in virtual reality by artificial workers. While virtual workers could conceivably lead happy virtual lives, there is no reason to think that economic and moral incentives will be aligned on this front—maximizing productivity might require workers to operate in, say, highly

[74] For discussion in the context of superintelligence and its bearing on the risk of astronomical quantities of suffering, see Sotala & Gloor (2017).

[75] See Hill & Tolk (2017) for a history of military simulations.

[76] See Perez (2022).

anxious states of extreme focus. To the extent that economic incentives generate pressure toward putting virtual workers in negative states, there is a risk that economic trajectories will lead to catastrophic quantities of suffering.[77]

- *Catastrophic manipulation*: Manipulative agents might use catastrophic simulations to render their threats credible or to give other agents prudential reasons to behave in certain ways, e.g. transporting uncooperative or unproductive digital minds into certain types of simulations might be used to incentivize digital minds to cooperate or be productive.[78]
- *Suffering and subjugated subroutines*: Future superintelligent systems may have the capacity to run large-scale simulations. They may exercise this capacity in order to gain information about their options and better pursue their goals. For a wide range of final goals, using simulations in this fashion would prove instrumentally valuable for such systems. For instance, superintelligent systems could run historical simulations to improve their ability to predict human behavior or historical trends. Such systems improve their economic standing by creating virtual workers. Or they could run simulations of the future in order to test options before selecting among them. A "boxed" simulation might run simulations in order to test escape strategies. Given that superintelligent AI systems will relentlessly optimize for whatever their optimization target is and that running large-scale simulations is likely to prove instrumentally valuable for them, we should expect them to run such simulations, even if the moral consequences are catastrophic—unless we carefully engineer them so as to avoid this hazard. For the same reasons, we should expect them to disregard the rights and interests of minds within such simulations. Both engineering superintelligent systems to avoid these hazards and verifying that they are so engineered may prove difficult. For even systems that behave only in morally permissible ways across a wide range of contexts may harbor large numbers of suffering subroutines. This could happen if, for example, we loaded the system with the values of an otherwise virtuous human who is indifferent to the suffering of subroutines. Or it could happen if the superintelligent system were inadvertently programmed to misattribute unconsciousness to all its subroutines or to misinterpret the valence sign or neutral point of its subroutines

[77] See Bostrom (2014: Ch. 11) and Hanson (2016).

[78] See Dainton (2012) and his discussion of Banks (2010).

experiences.[79] Given the large quantities of suffering in human and evolutionary history, it is easy to see how superintelligent systems running large-scale simulations could result in catastrophes in the form of many suffering subroutines.[80]

- *Catastrophically malevolent actors*: Malevolent actors are those that treat harming other individuals as a final end. At present, malevolent actors are severely limited in their ability to create new agents to harm. On an individual level, the procreative limits of human biology and the limited supply of mating opportunities generally precludes malevolent agents from pursuing their ends by creating large numbers of agents. There are examples from history of powerful and plausibly malevolent agents pursuing their ends by promoting social policies that led to the creation of more potential victims.[81] In any event, the arrival of simulations capable of realizing suffering minds would greatly enhance malevolent actors' abilities to pursue their ends through the creation of potential victims. In addition, biological architectures currently limit the magnitudes and kinds of suffering that malevolent actors can cause. There is no reason to think that these magnitudes or kinds are in the vicinity of the worst. Given the flexibility and information processing potential of digital architectures, there is some reason to think that simulations will enable forms of suffering that are much worse than those that burden biological systems. Finally, today's malevolent actors have incentive to conceal their activities so as not to prompt interference from morally motivated actors. Since large-scale simulations might be run within small portions of physical space or on computers that are externally inscrutable, simulations may enable malevolent actors to more easily conceal moral catastrophes. On the other hand, if such an actor achieved a decisive strategic advantage (i.e. one sufficient to outcompete all other actors), it could realize morally catastrophic ends openly since even actors that were aware of the malevolent actors' plans would be unable to thwart them. All this suggests that

[79] Compare: OpenAI, a leading company among those trying to develop AGI, inadvertently trained a large language model (GPT-2) to optimize for expressing negative sentiment as a result of a flipped sign and AI developers (literally) being asleep during the training process (Ziegler et al., 2019).

[80] See, e.g., Tomasik (2017).

[81] For instance, Mao Zedong's policies led to population growth followed (perhaps in ways that were foreseeable but not intended) by famines in which tens of millions of Chinese citizens died (Fitzpatrick, 2009).

the combination of large-scale simulations and malevolent actors would pose a significant catastrophic risk.[82]

8. Background on the Simulation Hypothesis and the Simulation Argument

Previous sections examined connections between catastrophic risks and simulations that might be run in our universe. The next few sections will explore connections between catastrophic risks and the *simulation hypothesis* that our universe is itself a simulation. While this may seem to be an outlandish or skeptical hypothesis, there is an interesting argument for it that is taken seriously by relevant experts.[83] In this section, I will rehearse the *simulation argument* for the simulation hypothesis. This will set the stage for discussing connections between the simulation hypothesis and catastrophic risk in later sections.

There are two driving ideas behind the simulation argument. One is the broadly empirical claim that the expected motivations and computing potential of technologically advanced civilizations support *simulation dominance*, the hypothesis that at least a small portion of beings like us will produce a very large number of beings like us that are simulated—a large enough number for most beings like us to turn out to be simulations.[84] The second is that if most beings like us are simulated, then we are probably simulated. This idea rests on *the (bland) indifference principle* that we should divide our credence evenly among hypotheses about our self-location in the class of

[82]On the other hand, creating suitable ensembles of simulated agents with malevolent impulses that are punished when they act on those impulses could incentivize even powerful would-be malevolent actors to behave morally by giving them reason to think they are in such a simulation—see Elga (2004) and the discussion of the simulation argument below. For discussion of existential and suffering risks posed by malevolent actors, see Althaus & Baumann (2020).

[83] The argument was introduced and defended by Bostrom (2003*a*). Others who take it seriously include Braddon-Mitchell & Latham (2022), Chalmers (2022), Ćirković (2015), Crummet (2020), Dainton (2012; 2020), Greene (2020), Hanson (2001), Johnson (2011), Lewis (2013), Monton (2009: Ch. 3). Schwitzgebel (2017), Steinhart (2010), Thomas (2022), and Turchin et al. (2019).

[84] There is debate in the literature about whether this hypothesis should, given its role in the argument, be made conditional on the reasoner's not being a simulation—see Thomas (2022); cf. Bostrom (2011) and Crawford (2013). For tractability and simplicity, I set this issue aside and work with the unconditional formulation.

observers like us.[85] Simulation dominance and that application of the indifference principle jointly entail that we are probably in a simulation.

The simulation argument has been spelled out in different ways in the literature. One choice point concerns how to precisify the relevant class of observers: while the indifference principle is intuitive, it is not clear exactly what it takes for observers to be like us. Here, I will finesse this issue by using 'beings like us' to pick out whichever observers fall within the relevant class on the most plausible version of the principle that is applicable to you and me.[86] For concreteness, you might think of this as the class of conscious beings with roughly human-level intelligence. In some presentations of the argument, the conclusion is simply that we are probably living in a simulation. In others, the conclusion is couched as a disjunction between our (probably) being in a simulation and possible ways of blocking that result.[87] The literature also contains various proposed amendments for patching the argument in response to objections, along with less consequential variations in formulation. To keep the discussion tractable, I will just work with the above formulation.

The argument can be questioned in various ways.[88] Some responses include:

- Intelligent simulated beings will not dominate because:

[85] More precisely, the bland indifference principle says that for a given hypothesis about how the world is qualitatively, we should divide our credence concerning our self-location on that hypothesis evenly among the observers like us that exist on that hypothesis. This should not be confused with indifference principles that require one to divide one's credence evenly between self-locations posited on different qualitative hypotheses or between different qualitative hypotheses when one's evidence does not adjudicate between them. These stronger principles are open to serious objections that do not apply to the bland indifference principle (hence the label 'bland')—see Elga (2004: 387-8) and Van Fraassen (1989). For discussion and defense of restricting rather than rejecting indifference principles, see Greaves (2016).

[86] It may turn out that being an observer like us is a matter of degree, either because being an observer is a matter of degree (an outcome that is hard to avoid on a reductive physicalist view of observerhood—see Lee (2019)) or because counting as like us is a matter of degree. In either case, the most plausible rendering of the indifference principle may then require one to divide one's self-locating credence among observers that are to some degree like oneself in proportion to the degree to which they are like oneself—see Dorr & Arntzeneius (2017). While I think such a proportional rather than egalitarian indifference principle may well turn out to be correct, I do not think the difference between them is important for the discussion that follows. In any event, for simplicity, I will work with the just sketched egalitarian principle.

[87] See Bostrom (2003*a*) and Chalmers (2022).

[88] See Chalmers (2022).

    - Civilizations generally go extinct before being able to create intelligent simulated beings.
    - Civilizations that can create intelligent simulated beings generally opt not to do so.[89]
- Granting that simulated beings will dominate, we should not use the indifference principle to infer that we are probably in a simulation because:
    - The indifference principle is false.[90]
    - Simulated beings would not be conscious and therefore the indifference principle does not apply.[91]
    - Some other feature of our evidence (such as our creativity, the fact that we seem to live in an immensely large universe, or the fact that we have not ourselves created simulated beings) indicates that we are not simulated beings, and hence prevents the indifference principle from showing that we are.[92]
- The argument is illicit in some other way:
    - We cannot have evidence that both indicates that there are many simulations in our universe and that we are simulations.[93]
    - Evidence that we are in a simulation is unstable—in that it is trustworthy just in case we rationally take ourselves not to be in a simulation—and so should be ignored.[94]
    - The simulation hypothesis is a skeptical one. Therefore we should reject the simulation argument even if we do not know where it goes wrong.

This is not the place to attempt a comprehensive evaluation of the simulation argument and responses to it. So I will restrict myself to the following remarks, which aim to (1) address what I expect to be some of the more prevalent concerns about the argument and (2) highlight connections between some of the responses and catastrophic risks.

---

[89] See Bostrom (2003*a*).

[90] See Dogramaci (2020).

[91] See Summers & Arvan (2021), Dainton (2012), and Chalmers (2022).

[92] See Chalmers (2022), Richmond (2017), and Hanson (2001).

[93] See Birch (2013), Crawford (2013), and Thomas (2022).

[94] See Crawford (2013). For discussion, see Chalmers (2022: online appendix) and Dainton (2012). For a response to this sort of objection, see Bostrom (2009). For related discussions of cognitive instability in the context of Boltzmann brains, see Carroll (2020), Chalmers (2018*b*), Dogramaci (2020), Kotzen (2020), Saad (forthcoming*c*).

First, consider the response that the argument fails because civilizations generally go extinct before being able to create intelligent simulated beings—for example, as a result of a technology that inevitably precedes simulation technology and leads to civilizational destruction shortly after its discovery.[95] If this response is correct, then it gives us reason to think that our civilization will end before being able to create intelligent simulated beings. Thus, if the response is correct and there is reason to think we are on track to be able to create intelligent simulated beings, then there is also reason to think our civilization will succumb to catastrophe in the relatively near term—a catastrophe that at least knocks it off track, whether or not it terminates our civilization. This can be understood as an argument for taking near-term catastrophic risks more seriously.

Second, if the simulation argument fails because advanced civilizations generally decide not to create intelligent simulated beings, this may be because civilizations that avoid catastrophe long enough to be able to produce such simulations are generally very risk averse.[96] There is also reason to think such civilizations would have centralized control and/or coordination schemes with near-universal compliance: absent such control or compliance, we would expect sub-civilizational actors to create many intelligent simulations, given that intelligent simulations would eventually become cheap to produce in advanced civilizations.[97] Similarly, those who are attracted to this response may take it as an indication of the paths that are available to humanity which do not end in near-term extinction.

Third, in informal interactions I've encountered a number of fellow philosophers who are tempted by the instability response. However, for several reasons, this response is unconvincing:

- Any instability induced by the simulation argument afflicts both the simulation hypothesis and its negation. Symmetry suggests that the correct moral of such instability cannot be that we should simply reject the simulation hypothesis.
- The argument can be recast in terms of intelligent beings in *stable simulations*, simulations of scenarios that are similar enough to the level of reality at which

[95] See Bostrom (2003*a*) and Chalmers (2022).

[96] See Chalmers (2022).

[97] See, e.g., Ćirković (2008: 138).

the simulations are run for the simulated beings to be able to reliably reason from their evidence, at least for the reasoning in the simulation argument.[98] While the recast argument would need to be evaluated on its own merits, it is not clear that such recasting introduces any serious flaws.

- We should not in general disregard unstable evidence. This is vividly illustrated through real-life cases involving *hypoxia*,[99] a condition involving oxygen deprivation and impaired reasoning abilities: any reasoning that leads to the conclusion that one is hypoxic is unstable, since it casts doubt on itself. Yet it would be foolish for mountain climbers and pilots to ignore the hypothesis that they are hypoxic on that basis—indeed, despite its instability, such reasoning plausibly sometimes gives people a reason to take that hypothesis seriously.[100] Absent some reason for thinking that any instability associated with the simulation argument is relevantly different from the instability associated with arguments for one's being hypoxic, the simulation argument should not be dismissed on the ground that it is unstable.

Another response that many people find tempting is to dismiss the simulation argument on the ground that it is a skeptical argument. For several reasons, this response is also unconvincing.

- Whereas typical skeptical arguments appeal to mere possibilities to cast doubt on what we take ourselves to know,[101] the simulation argument uses broadly empirical considerations to support positive conclusions, namely simulation dominance and, in turn, the simulation hypothesis. (The susceptibility of the simulation hypothesis to empirical (dis)confirmation also distinguishes it from conspiracy theories that resist (dis)confirmation.)
- The simulation hypothesis is arguably a metaphysical hypothesis about the origins and underlying nature of our universe.[102] Traditional theistic hypotheses hold that our universe was created by an agent and are typically regarded as non-skeptical. But it is hard to see how the sort of creation of our universe

[98] See Dainton (2012); cf. Chalmers (2022: online appendix).

[99] For a summary of literature on neuropsychological responses to hypoxia, see Virués-Ortega (2004).

[100] For relevant discussion, see Elga (2008), Kotzen (2020), and Christensen (2016).

[101] See Bostrom (2005). However, there are exceptions, notably skeptical arguments from dreams, the evolutionary origins of our beliefs, and physical theories that proliferate Boltzmann Brains.

[102] See Chalmers (2003; 2022).

suggested by the simulation argument could engender skepticism when traditional theistic hypotheses lack skeptical import. Similarly, non-skeptical theories in physics posit underlying computational processes to explain the observable universe. But it is hard to see how the sort of underlying nature of our universe suggested by the simulation argument could engender skepticism when these physical theories do not.

- Finally, it should be borne in mind that we are not faced with a choice between embracing the simulation hypothesis and accepting that reality is largely as it appears: even on the assumption that we are not in a simulation, science and philosophy give us good reasons for thinking that the external world is not as our experience presents it, whether or not there is a watered-down sense in which our experiences can be said to be accurate.[103]

9. Shutdown Risk

One corollary of the simulation hypothesis is that our simulation may be shut down.[104] Depending on the axiological trajectory of our universe at the time of shutdown, shutdown could be catastrophic. If a shutdown happened today, it would prematurely end the lives of billions of people and it might destroy immense quantities of expected value that lie in our potential to usher in a grand future. Shutdowns will continue to pose a catastrophic risk, at least as long as we manage to steer clear of other catastrophes.

It may be tempting to think: we shouldn't worry about catastrophic shutdowns because there's no way for us to influence whether they occur. However, this thought is mistaken on two counts. First, even if we cannot influence the risk of catastrophic shutdowns, that risk has implications for the expected value of the long-term future: if we assign a tiny but constant probability to shutdown in any given year (conditional on survival up to that point), that will drive down the expected (dis)value of outcomes that would occur further in the future.[105] Such discounting would drive down the expected (dis)value of nearer term future to a lesser extent. Thus, updates that elevate the risk of catastrophic shutdowns will tend to weaken the case for prioritizing the far future.

[103] See, e.g., Chalmers (2006), Cutter (2021), Hoffman (2019), and Pautz (2014).

[104] See Bostrom (2002*a*), Ćirković (2008), Greene (2020), and Turchin et al. (2019).

[105] See Tomasik (2016); cf. Ord (2020: Appendix B).

Second, there are shutdown triggers that we may be able to influence:

- *Excess computational consumption*:[106] programs are often coded to halt under certain conditions rather than exhaust the computational resources of the systems on which they are run. In the event that we are living in a simulation and our simulators have finite computing resources, there is reason to think they would also program our simulation in an efficient manner—e.g. by not simulating microphysical details that lack observational import for simulation inhabitants—and incorporate such halting mechanisms. In that case, we might trigger shutdown by way of activities that require large quantities of compute and therefore risk engaging halting mechanisms.
- *Moral triggers*: Simulators might shut down simulations for moral reasons. Since shutdown would itself be an existential catastrophe, moral criteria for triggering it would presumably concern some other moral bad such as suffering. This hypothesis suggests a way in which extinction risk might be higher and suffering risks lower than we'd otherwise think.
- *Extinguishing superintelligent threats*: There is an open question in AI safety research about whether superintelligent systems could be safely confined to a virtual environment—a crucial concern is that such systems might use their superintelligence to find clever escapes from their virtual confinement that we are not smart enough to anticipate and thwart. Simulators of our universe might share this concern about superintelligent systems and address it by shutting down simulations when superintelligent systems are created or become likely to emerge within the simulation. For this reason, creating superintelligent systems would raise the shutdown risk. The same goes for making progress toward creating such systems. Of course, creating superintelligent systems might mitigate other catastrophic risks, or even mitigate shutdown risk in a different way.
- *Interest triggers*: If we are in an entertainment or research simulation, we should expect the probability of shutdown to increase if our universe loses its research or entertainment value.[107]
- *Defeat in victory*: If we are in a gaming simulation, meeting a victory condition for the game could result in shutdown.

[106] See, e.g., Ćirković (2008).

[107] See Hanson (2001) and Greene (2020).

- *Simulation awareness*: Finding out that we're in a simulation or in a certain sort of simulation might defeat the purpose of some sorts of simulations—e.g. Fermi research simulations—leading to shutdown.[108]

Developing a more systematic understanding of potential shutdown triggers and the prospects for avoiding them is an underexplored topic.[109]

10. Potential Upsides of Shutdown

The previous subsection noted some obvious reasons for thinking that the shutdown of our simulation might be catastrophic. What may be less obvious is that there are also ways in which shutting down our simulation could mitigate catastrophic risks. These include:

- *Comparative advantage*: In evaluating negative effects of shutdown, we should be mindful of the alternatives, as shutdown's negative effects may be unavoidable and it may turn out that alternatives would be morally worse.
    - It is tempting to think that shutdown is a catastrophe to be avoided because it would result in billions of deaths and the cessation of conscious beings in our universe. However, the available alternatives share these consequences—what is at stake with shutdown is how our universe will end, not whether it will cease realizing value after some time or other.[110]
    - Similarly, one might think that shutdown is a catastrophe to be avoided because it would cut short many lives and projects whose continuation would be valuable. However, it is not clear that there is any realistic alternative available that would avoid this. Indeed, if shutdown is avoided, we would expect there to be a last generation of conscious beings whose lives and projects will be cut short—the same goes for conscious beings belonging to other future generations that will exist only if shutdown is avoided.
    - A more plausible thought is that *we* have reason to prevent shutdown because it would cut short *our* lives and projects. While this point is well

[108] For reasons to think simulation awareness is a risk, see Greene (2020). For reasons to think it is not, see Alexander (2019) and Braddon-Mitchell & Latham (2022).

[109] The most thorough discussions of this that I am aware of are Braddon-Mitchell & Latham (2022), Greene (2020), and Turchin et al. (2019).

[110] See Lenman (2002).

taken as one of prudence, it is doubtful that it has any purchase from a perspective of impartial benevolence: from that moral vantage point, there seems to be no reason for thinking it would be worse for us to be subjected to shutdown than it would be for other people (perhaps much larger in number, with longer and higher quality lives) to have their lives and projects prematurely terminated.[111]

- Suppose that the axiological trajectory of our universe is negative: by default, the expected value of the future is negative—perhaps because we have created digital minds that endure kinds of suffering that cannot be compensated by any sort of good or perhaps because probes are launched to spread simple forms of life across the galaxy, setting the stage for the repetition of the horrors wrought by biological evolution on Earth.[112] In such a case, the default outcome might well be worse than shutdown, even if shutdown itself has negative expected value. If so, then triggering shutdown might serve as a kind of *escape hatch* from a world gone wrong. Indeed, even if one assigned only a small probability to the simulation hypothesis and a still smaller one to triggering attempts inducing shutdown, trying to bring about shutdown might be worth attempting.

- Suppose that in expectation our universe's future will contain more good than bad. Even in this case, there might be decisive moral reasons in favor of shut down as a result of *goods discounting*.
  - For example, maybe a *downside-focused* moral theory is correct.[113] Such a theory would place greater weight on preventing negative outcomes rather than on bringing about positive ones. As a result, the negative outcomes averted by shutdown (astronomical quantities of suffering, say) could justify triggering it, even if doing so would prevent positive outcomes (even larger quantities of happiness, say) that are better than the negative outcomes are bad.

[111] See Lenman (2002) and Prinz (2012). Those who countenance agent-relative moral reasons may see middle ground here—see, e.g., Mogensen (2019*b*) and references therein.

[112] For moral concerns about directed panspermia, see Dello-Iacovo (2017) and O'Brien (forthcoming).

[113] For reasons to doubt that our universe's past or future are net positive, see Anthis (2018; 2022), Benatar (2008), Gloor (2016; 2018), and MacAskill (2022: Ch. 9). For motivation for risk-averse decision theory and arguments that it recommends extinction-hastening interventions over extinction-preventing ones (given longtermist moral assumptions), see Pettigrew (2022).

- Or maybe the correct moral theory assigns *asymmetrically diminishing returns* such that the amount of positive value realized per unit of goods decreases as the number of units of goods increases but the amount of negative value realized per unit of bads does not decrease (by as much) as the number of units of bad increases. In this case, we cannot simply read off the value of bringing about a given future in our universe from the goods and bads it contains: to calculate the value of a given future, we would need to discount the goods it contains in accordance with the relevant aggregation function. Thus, even if a given future of our universe would contain more good than bad, bringing it about might contribute net disvalue once the discounting of the goods is taken into account.
    - Views that embrace asymmetrically diminishing returns lend support to trying to trigger shutdown as follows: either we are in a simulation or we are not.[114] If we are not, then we cannot trigger shutdown and there is little downside to trying to do so. If we are in a simulation, the world probably contains vastly many goods beyond those featured in our universe. In that case, conditional on asymmetrically diminishing returns, we should expect the value of any goods we can bring about to be severely discounted and the disvalue of any bads we can bring about not to be so discounted. Thus, to the extent that, in comparison with our other options, triggering shutdown trades off the realization of goods with the prevention of bads, asymmetrically diminishing returns should push us toward assigning higher expected value to triggering shutdown.

[114] One way to motivate asymmetrically diminishing returns is to note that (1) some bads, such as uncompensated suffering, seem not to diminish at all in disvalue as their quantity increases and (2) countenancing diminishing returns for goods provides an escape from what's known in population ethics as the Repugnant Conclusion, which is the (supposedly) implausible claim that "For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better even though its members have lives that are barely worth living" (Parfit, 1984: 342). For objections to one form of asymmetrical diminishing returns, see *ibid* (§134). For reasons to doubt that avoiding the Repugnant Conclusion is a requirement for an adequate approach to population ethics, see Zuber et al. (2021).

I hasten to add: even in circumstances where evidence indicates that the best option will involve trying to trigger shutdown, it could be a grave error to attempt to trigger shutdown too soon. That's because there's value in delaying—and preserving other options in the meantime—in order to put us in a better epistemic position to determine whether shutdown is the best option.[115] And that value could easily outweigh any time cost associated with delay. Compare: suppose you're on a slowly leaking ship that will probably sink. Even if your evidence indicates that jumping overboard with a life vest is your best option, that doesn't mean you should take it immediately. It might instead be more reasonable to wait to see how the situation evolves in case a better option emerges (e.g. staying on the ship in the event that the leak is fixed or exiting in a lifeboat).[116]

11. The Simulation Argument and Religious Catastrophic Risks

Traditional religions countenance distinctive catastrophic risks ranging from apocalypses inflicted by a creator to afterlives involving eternal torment. (Distinctively) religious catastrophic risks are rarely discussed in the scholarly literature on catastrophic risks. The same goes for public non-academic discussions of catastrophic risks that I am familiar with.[117] Perhaps this is because such risks are typically approached with methods that do not treat these religions—or any religion for that

[115] Cf. Ord (2020: Ch. 2).

[116] The Unilateralist Curse provides a further reason for caution here: when large numbers of at least somewhat error-prone, altruistically motivated agents are each in a position to unilaterally act in ways that affect others, we should expect unilateral interventions to happen more often than is optimal (Bostrom et al., 2016). In the case at hand, the curse we face is: if a sufficiently large number of error-prone, altruistic agents will be in a position to unilaterally take extinction-inducing interventions and their evidence renders extinction non-optimal in expectation, we should expect some such agent to mishandle their evidence, overestimate the value of such interventions, take them, and thereby cause extinction. For general suggestions for avoiding the Curse that can be applied in this case, see *ibid.*

[117] Religious discussions of catastrophic risks that I am familiar with tend not to discuss non-religious catastrophic risks. There are a few notable exceptions. Riedener (2021) argues that existential risk reduction is extremely important from a Thomist Christian perspective. Danaher (2015) brings some ideas from philosophy of religion to bear on catastrophic risk posed by AI: he explores an analogy between the "skeptical theist" view that appearances of evil are not strong evidence against the existence of a benevolent God with Bostrom's (2014: Ch. 8) "treacherous turn" concern that AI systems may behave cooperatively (e.g. while boxed within a simulated environment) and hence appear benign before abruptly changing their behavior (in a potentially catastrophic manner) to pursue their final goals. The (pseudonymous) author of this post argues that longtermist movements have a dismal track-record, that religions have been (for good or ill) influential long-termist movements, and that catastrophic risks that concern longtermists may be more effectively mitigated through near-term focused approach.

matter—as a source of data, much less as authoritative. In any event, given the enormous stakes associated with religious catastrophic risks, they should be considered in analyzing how to reduce catastrophic risk unless we regard religious catastrophes as astronomically unlikely.[118] For those who regard religious catastrophic risks as non-negligible, there is then a question as to how they interact with the simulation argument.[119] There are at least four connections of interest.

Flawed simulators and religious catastrophes. Outside the context of the simulation argument, it is often assumed that our universe either was not created or else it was created by an agent that is all powerful, all knowing, and perfectly good. The simulation argument casts doubt on this assumption: it gives reasons for thinking that our universe is created by intelligent beings, but provides no reason to think such beings would exhibit a maximal degree of power, knowledge, or goodness. Further, taken with the many morally problematic features of our universe, the simulation argument lends support to the hypothesis that our universe was created by a morally flawed or morally indifferent agent. This in turn lends support to the hypothesis that our universe is subject to religious catastrophic risks such as catastrophic interventions by our creator(s) or disvaluable afterlives.[120]

Simulation as a solution to the problem of natural evil. The simulation argument arguably indirectly supports the hypothesis that God exists—this is so even on the assumption that our universe was, if created, created by a non-divine simulator. For the simulation hypothesis offers a candidate solution to the *problem of natural evil*, i.e. the problem of reconciling the apparent existence of natural evils (ones not caused by the free choices of agents) with the existence of God (understood as an agent that is all powerful, all knowing, and perfectly good). The candidate solution on offer is that the appearance of natural evils is an illusion: God created a world devoid of natural evils in which agents make free choices. One of those agents freely chose to create a simulation, a simulation

[118] For an argument that religious catastrophic risks are important, tractable, and neglected relative to other catastrophic risks, see Sampson (2022).

[119] For discussion of the religious implications of the simulation argument or hypothesis, see Bostrom (2003: 254), Chalmers (2022: Ch. 7), Crummett (2020), Dainton (2020), Johnson (2011), and Steinhart (2010).

[120] This support may be somewhat attenuated by the fact that religions that posit catastrophes tend not to countenance a flawed or indifferent simulator of our universe; hence, the supposition that our universe has such a creator tells against those specific catastrophes and so in that respect detracts from overall risk.

that turns out to be our universe. Thus, what appear to be natural evils in our universe are really evils caused by our simulator, not by God.

Extending the solution to help with other problems for theism. The simulation solution to the problem of natural evil can be developed to solve a range of other problems for theism. For suppose we add to the solution that our universe is just one simulation in a much larger reality of which it is, in a certain respect, not representative.[121]

- By supposing that the ratio of good to evil in our universe is much worse than the ratio between good and evil in the larger world to which it belongs, we can extend the solution to help with *the problem of evil*, i.e. the problem of reconciling the evil we find with the existence of God.[122]
- By supposing that our universe is sub-optimal but part of an optimal world, we can extend the solution to help with *the sub-optimality problem*, i.e. the problem of reconciling the sub-optimality of what we observe with the expectation that God would create an optimal world.
- By supposing that our universe is uncharacteristically inefficient in realizing value, we can extend the solution to help with *the problem of scale*, i.e. the problem of reconciling the expectation that God's moral goodness would be reflected in the organization of his creation with the observation that such value seems to be realized at great inefficiency at a cosmically miniscule scale.[123]
- By supposing that beings whose evidence is not manipulated through a simulator's exercise of free will have ample evidence of God's existence, we can extend the solution to help with *the problem of divine hiddenness*, i.e. the problem of explaining why the available evidence for theism is weaker than we would have expected if theism were true.[124]

Simulation solutions to problems for theism have implications for catastrophic afterlives. On the one hand, there is some plausibility to the thought that we are less likely to have the sort of free will required for responsibility if we are simulated than if

[121] For multiverse solutions to theistic problems that do not invoke simulations, see Kraay (2010), Leslie (1989), and Megill (2011). For an overview of surrounding literature, see Kraay (2014: 9-11).

[122] The problem of evil is variously formulated—see, e.g., Benton et al. (2016). The same goes for the other problems considered below.

[123] See Everitt (2004: Ch. 11).

[124] See, e.g., the essays in Green & Stump (2015) along with Schellenberg (1996; 2010).

we are not. This in turn suggests that, on the simulation solution, we are less likely to be subject to catastrophic afterlives imposed by God as just punishment. On the other hand, on simulation solutions, we need to distinguish between afterlives imposed by God and those imposed by the non-divine simulator(s) of our universe. Thus, there remains the possibility that we are in a simulation and will be subjected to a catastrophic afterlife by the non-divine simulator(s). The simulation solution somewhat constrains how bad such an afterlife could be: on pain of reinvigorating the problem of evil, the catastrophic afterlives cannot be so bad (or too inadequately compensated)[125] as to render the existence of God implausible.

Simulations, fine-tuning, and God. Cosmological fine-tuning arguments for the existence of God are among the most popular contemporary arguments for the existence of God. They appeal to the (supposed) fact that fundamental physical parameters take values within narrow ranges required for life.[126] This is claimed to be expected if there is a designer of our universe but not otherwise—the idea is that it wouldn't be surprising if a designer selected those parameters because they are required for life and the designer wanted to create a universe with life (or something for which life is a prerequisite such as biological intelligence or conscious organisms). It's usually assumed that if our universe has a designer, the designer would be God. Insofar as cosmological fine-tuning

[125] See, e.g., Adams (2000), Crummett (2017), and Stump (1985).

[126] For an overview, see Friederich (2018). There are also less explored non-cosmological varieties of fine-tuning such as the match between the universe's boundary conditions with its laws (Cutter & Saad, forthcoming) and the fine-tuning of experiences' causal profiles with their rational profiles (Chalmers, 2020; Goff, 2018, Saad (2019; 2020; forthcoming*a*), James (1890), Mørch (2018), Pautz (2010; 2020). Cutter & Crummett (forthcoming) defend an argument for theism that appeals to psychophysical fine-tuning. While both cosmological and psychophysical fine-tuning can be used to argue for theism, they interact with the simulation hypothesis in different ways. For instance, whereas an ensemble of simulation universes might explain cosmological fine-tuning, it is not clear that they could explain psychophysical fine-tuning—e.g., if the psychophysical laws have a functionalist character at the base level, then they will not be manipulable independently of other factors in the simulations. Similarly, if they involve a non-functional feature that cannot be freely varied in the simulation, the simulators will not be able to vary it. Even if simulators could in principle vary psychophysical laws in simulations, epistemological obstacles associated with consciousness may prevent simulators from figuring out how to do so. And even if simulators could vary the psychophysical laws in simulations and figure out how to do so, they may have no incentive to do so, as they may only be concerned with outputs of the simulation. For simplicity, I mostly hereafter set aside psychophysical fine-tuning and how different varieties of it interact with the simulation hypothesis and associated catastrophic risks.

arguments support the hypothesis that God exists, they presumably also modulate the catastrophic risks associated with the existence of God. The simulation argument bears on cosmological fine-tuning in several respects.[127]

- The simulation argument casts doubt on the assumption that if our universe has a designer, it's God. It does this by pointing to an alternative: the fine-tuner of our universe might be the intelligent but non-divine being running our simulation. To the extent that fine-tuning evidence and the simulation argument together support this non-divine design hypothesis, the simulation argument constrains how much fine-tuning evidence can support the existence of God—the simulation argument in effect redirects support from fine-tuning for theism to a non-theistic design hypothesis.
- A more standard response to fine-tuning arguments is that they neglect the availability of a multiverse explanation of fine-tuning. Multiverse explanations posit a vast ensemble of universes with varying physical parameter values such that it is to be expected that some universe has life-supporting parameter values. One worry about the multiverse explanation is that it evidently requires many universes with different fundamental laws, and thus suffers the theoretical vice of having an immensely complicated set of basic laws.[128] The simulation argument lends to the following response to this worry: an ensemble of simulation-universes explains fine-tuning; however, these simulations are non-fundamental entities within the universe in which our simulation is run; that universe may well have a simple set of fundamental laws; thus, the multiverse explanation does not require the proliferation of fundamental laws.
- It might be thought that cosmological fine-tuning arguments for the existence of God (or a multiverse with different fundamental laws in different universes) could be recast to concern the level of reality in which our simulation is run, if we are in a simulation.[129]

[127] For discussion of the simulation hypothesis and fine-tuning, see Chalmers (2022: Ch. 7) and Steinhart (2010). For an objection to some multiverse hypotheses on the ground that they require a complicated set of basic laws, see Cutter & Saad (forthcoming).

[128] For discussion of this theoretical vice, see, e.g., Chalmers (1996: 213-4), Cutter & Saad (forthcoming), and Sider (2020: 102).

[129] See Chalmers (2022: Ch. 7).

- One problem with this thought is that since we lack access to the physics of any such universe, we lack the relevant empirical data needed to run the argument.
- It might be replied: the most likely scenario in which we are in a simulation is one in which (in accordance with the simulation argument) our simulation is the product of observers like us. But observers like us would tend to exist in situations with physics like our own (navigating a world with radically different physics isn't compatible with being an observer like us). And, being most interested in and capable of designing observers like themselves, such observers would tend to simulate observers in situations with physics like our own. All this suggests that if we are in a simulation, then our physics is similar to the physics of the level at which our simulation is run and hence that cosmological fine-tuning arguments which initially seemed to apply to our universe will at least apply to whatever unsimulated universe we inhabit, even if we happen to inhabit a simulation.
- One limitation of this reply is that if the physics of the universe in which our simulation is run is like the physics of our universe, it is reasonable to expect the simulation argument to apply to our simulators. In that case, just as the simulation argument threatens cosmological fine-tuning arguments that invoke fine-tuning in our universe, so too would the simulation argument threaten cosmological fine-tuning arguments that invoke fine-tuning in our simulators' universe. The reply could be repeated for our simulator's level to try to show that their simulator would also have a physics that is similar to our own. However, even if the reply is plausible at each level, the more times it is iterated, the more likely it is that it will fail at some level.[130] One way to see this is to notice that small differences between simulator and simulatee physics at adjacent levels can add up to big differences in the descent from our level to the basement level. A big difference could well be that whereas our simulation's physics is fine-tuned, basement level physics is not.

We've seen that the simulation argument in different ways supports theism and undermines such support. Likewise, we've seen that the simulation argument in

[130] Though the same goes for the simulation argument.

different ways boosts and diminishes religious catastrophic risks. The net effects of the simulation argument on these issues remain open questions.

12. The Simulation Argument, Self-Location, and Catastrophic Risk

12.1 Background on Self-Location

The simulation argument relies on self-locating information: it assumes you should divide your credence evenly among observers like yourself. There are other arguments that also rely on self-locating information and bear on catastrophic risks. The simulation argument interacts with these arguments in ways that bear on catastrophic risks. This section will describe some of these arguments and interactions. I should flag that there is much controversy within the literature on how to rationally respond to self-locating information and that much of the literature is of a technical nature that I cannot adequately summarize here.[131] I expect that much of the discussion that follows will merit revisiting in a more technical setting and much of what I say in this subsection would need to be qualified or scrapped in light of reassessment. I therefore offer the discussion that follows in a provisional spirit in hopes that it will stimulate further work on the topic, even if what I say turns out to be misguided in important ways.

Before diving into the arguments, I'll use a simple example to put some of the key issues about self-location on the table: suppose an urn contains an unknown number of balls and that you and perhaps some other subjects are each going to draw one ball from. Next, suppose you find out that most subjects who take a ball from the urn draw a red ball. Intuitively, this gives you reason to think that you will probably draw a red ball. This is an instance of 'inward' reasoning that moves from information about a distribution of subjects to a conclusion about yourself—this is the sort of reasoning encoded in the indifference principle invoked by the simulation argument.

In contrast, 'outward reasoning' moves from information about yourself to a conclusion that concerns the distribution of subjects like yourself.[132] To illustrate, suppose that you are initially ignorant about how many subjects will draw from the urn and what it contains. You then draw a red ball. Here, two inferences are prima facie plausible.

[131] Notable work on the topic includes Bostrom (2002*b*), Dorr & Arntzenius (2017), Elga (2000; 2004), Lewis (2001), Titelbaum (2013), Isaacs et al. (forthcoming), and Manley (ms).

[132] For discussion of the inward-outward distinction, see Manley (ms).

First, you might take the fact that you drew a red ball to boost the expected *number* of subjects who will draw red balls. After all, you presumably initially reserved credence for the hypothesis that no subject would draw a red ball. You've now eliminated that hypothesis and presumably redistributed whatever you had in it to hypotheses on which more than one subject draws a red ball. Call this *number boosting*. Second, you might take the fact that you drew a red ball to boost the expected *proportion* of subjects who will draw red balls. After all, the higher proportion of subjects who draw red balls, the more likely it is that you will draw a red ball, in which case those hypotheses get a boost. Call this *proportion boosting*.

A final variation: you are initially ignorant about what the urn contains, except that you know that it contains turquoise balls and that they—unlike balls of any other color the urn contains—are all too deep in the urn for any subject to reach. In this case, draws cannot be treated as representative of the urn's contents: they exhibit a *sampling bias*. For example, drawing a red ball rather than a turquoise ball is not evidence against the urn containing turquoise balls. In this case, the sampling bias is an *observation selection effect*: observations do not qualify as random samples because observations are biased toward certain outcomes.

In practice, factoring in number boosting, proportion boosting, and observation selection effects raises a host of difficult issues, e.g. concerning the individuation of reference classes.[133] A case in point is the bearing of number boosting and proportion boosting on the simulation argument. Insofar as number boosting favors hypotheses with more observers regardless of their observer-type,[134] it will favor hypotheses on

[133] See Bostrom (2002*b*: Ch. 4) and Dorr & Arntzenius (2017).

[134] This is close to what Bostrom (2002*b*: 66) calls the *self-indication assumption*, which claims that "[g]iven the fact that you exist, you should (other things equal) favor hypotheses according to which many observers exist over hypotheses on which few observers exist." Like number boosting, this formulation of the self-indication assumption does not specify how big of a boost hypotheses with more observers should receive. However, Bostrom goes on to offer as a formalization of the self-indication assumption a principle which specifies the size of the boost for pairs of hypotheses and argue that the self-indication assumption should be rejected because of the implausible extent to which it favors certain hypotheses over others (*ibid*: 122-6). The implausible consequences Bostrom derives from what he offers as a formalization of the self-indication assumption do not follow from his initial formulation of the principle, since it does not specify the size of the boost. For the same reason, his criticisms of the assumption do not apply to number boosting, and analogous criticisms would not apply to proportion boosting.

which the world contains large numbers of observers in simulations over non-simulation hypotheses on which the world is relatively sparsely populated. And insofar as number boosting favors hypotheses with more observers like us, it will favor large-scale simulation hypotheses on which many beings like us are simulated over sparsely populated non-simulation hypotheses. Thus, number boosting arguably bolsters the simulation argument's simulation dominance premise that at least a small portion of beings like us will create a large number of beings like us. In contrast, insofar as proportion boosting favors hypotheses on which a higher proportion of beings are observers of some type or other, it will favor panpsychist views over views on which the universe is relatively sparsely populated with macroscopic subjects (whether simulated or not). On the other hand, insofar as proportion boosting favors hypotheses on which a higher proportion of observers are observers like us, it will favor simulation hypotheses on which enough beings like us are simulated for there to be proportionally more beings like us than there are on rival simulation and non-simulation hypotheses. The latter result also fits with the simulation argument's simulation dominance premise, though that premise does not require that there be a higher proportion of simulated beings like us than there are other sorts of observers.

It is not obvious that the individuation of reference class implicit in the foregoing applications of number boosting and proportion boosting are correct. That said, it should be borne in mind that number and proportion boosting may operate across multiple reference classes even within a single case. Indeed, the foregoing applications are all mutually compatible.[135] In what follows, rather than trying to settle the proper

[135] To yield a precise boost from number or proportion boosting in a given case, five parameters need to be set:

(i) your observer type(s),
(ii) your observation type(s),
(iii) the boosted observer type(s),
(iv) the boosted observation type(s), and
(v) the size of the boost.

It is extremely difficult to find a plausible general principle for setting these parameters (cf. *ibid*). That said, reflection on cases show that there are reasonable and unreasonable ways of setting the parameters and applying number and proportion boosting. This mirrors the case of inductive inferences *not* concerning observers: while reasonable and unreasonable inferences are easy enough to find, the prospects for precisely delineating the reasonable inferences in that class seem bleak. This analogy is no accident: number and proportion boosting are self-locating inferences that are special cases of inductive inference from observations of something to a conclusion about a broader reference class that includes

individuation of reference classes and what applications of number and proportion boosting are legitimate, I will instead focus on drawing out implications of different choices on these scores.

12.2 Evolutionary Arguments for Easy Artificial Intelligence

Some authors have appealed to the observation that evolution by natural selection produced human intelligence to promote optimism about our ability to engineer artificial systems with human-level intelligence.[136] Pre-empirically, it's not obvious that a world with our physics and chemistry could give rise to human-level intelligence at all, much less that it could feature beings with such intelligence that would be able to engineer systems with such intelligence. Thus, finding out that evolution produced human-level intelligence eliminates one obstacle—that of incompatibility with the laws of nature—to our engineering such systems and so provides at least a smidgen of support for our capacity to achieve this engineering feat.

However, such optimism does not go beyond whatever optimism is licensed by the observation that physical and chemical processes in our world somehow gave rise to humans. The more interesting sort of evolutionary arguments contend that the fact that evolutionary processes *not aiming for intelligence* gave rise to human intelligence *during the relatively short time* (by cosmological standards) that Earth has existed suggests that engineering such intelligence is not that difficult of a problem. The idea can be spelled out in terms of design space: despite being a very inefficient search procedure relative to those that are available to human engineers, evolution by natural selection managed to hit upon human-level intelligence in design space; so, given our superior search capabilities, we should be able to hit upon it without great difficulty.[137]

If this argument succeeds in reducing the expected difficulty of engineering human-level intelligence, it thereby bolsters the simulation argument by rendering it more plausible that civilizations like ours will be able to create simulated beings like us before going extinct.

---

that thing. Noticing this should help allay any worry that, in the absence of a worked out account, the subject-involving character of number and proportion boosting renders their epistemology suspect.

[136] See Chalmers (2010*b*) and Morevac (1976; 1999); cf. Shulman & Bostrom (2012).

[137] See Cotra (2020) and Shulman & Bostrom (2012: §1.2).

Rendering it more plausible that we will be able to engineer human-level intelligence also renders it more plausible that we will be able to—and, indeed, will—engineer superintelligent AI systems. That outcome would itself bear on catastrophic risks, as superintelligent AI both poses catastrophic risks and harbors the potential to mitigate them. Thus, the evolutionary argument for easy human-level AI bears on catastrophic risks not only by bolstering the simulation argument, but also by raising the probability of superintelligent AI.[138]

Just as the import of the simulation argument is sensitive to how number and proportion boosting are applied, so too is the import of the evolutionary argument for easy AI.[139] To the extent that number boosting favors hypotheses with more observers, a crucial issue is how the difficulty in engineering intelligence co-varies with the number of observers. The relationship is not obvious: while less intelligent observers may require fewer resources per observer (compare, say, fish versus humans), increases in intelligence may yield access (e.g. via space colonization) to large but otherwise inaccessible resource reservoirs. For similar reasons, it is not obvious whether proportion boosting of hypotheses on which a higher proportion of entities are observers favors easy AI hypotheses. On the other hand, to the extent that number boosting favors hypotheses with more observers like us, it favors:

- *easy evolution hypotheses* on which evolution independently produces large numbers of intelligent beings like us across many planets over Rare Earth hypotheses on which Earth is the only planet on which evolution produces human-level intelligence, and
- *easy engineering hypotheses* on which evolution produces intelligent beings like us who go on to create artificial beings with such intelligence over difficult engineering hypotheses on which evolution produces comparably many intelligent beings like us but such beings generally do not go on to create artificial beings with human-level intelligence.

Similarly, if proportion boosting favors hypotheses on which a higher proportion of observers are like us, it presumably favors hypotheses to the extent that it is easier for evolution to create beings like us and easier for us to create AI systems with human-level intelligence. Thus, number and proportion boosting arguably boost easy AI hypotheses and in turn lend support to the simulation argument.

---

[138] For discussion of interactions between superintelligence and the simulation argument, see Prinz (2012).

[139] See Shulman & Bostrom (2012: §§3-4).

The evolutionary argument for easy AI is also sensitive to observation selection effects and choice of reference class. To see this, suppose first that the operative reference class is that of observers with human-level intelligence. That is, suppose that our observations of a given sample space are to be treated not as random samples from that space but as random samples from the observations human-like intelligent observers make of that space. Under this supposition, even if creating human-like intelligence is in fact exceedingly difficult, it might *seem* quite easy before taking into account the observation selection effect.[140] For example, suppose that the expected duration for Earth-like conditions to yield human-like intelligence is many orders of magnitude greater than than have thus far elapsed on Earth. And suppose that the world contains many life-hospitable planets but that they all exist for durations on the same order of magnitude as Earth's history. Then only a tiny fraction of planets would harbor observers with human-level intelligence, but such observers would generally find themselves to be products of evolution that were produced in relatively short order. To resist the temptation to erroneously conclude that engineering human-level intelligence is relatively easy, they would need to take into account this selection effect.

Alternatively, suppose that the operative reference class is a wider one—say, the one that includes observers with human-level intelligence and observers with lower levels of intelligence. In that case, while the observation selection effect would make the evolution of *intelligence* appear easy even if it is hard, it would leave room for observing the absence of *human-level intelligence*. So even once we take this effect into account, the fact that we observe human-level intelligence would rule out some hypotheses on which it goes unobserved because of how difficult it is to produce. Thus, under the noted supposition, the observation selection effect leaves the evolutionary argument intact (albeit attenuated) and permits us to reason from evolution's rapid production of human-level intelligence on Earth to conclusions about how difficult human-level intelligence is to engineer.

In this fashion, the evolutionary argument for easy AI—and in turn its ability to bolster the simulation argument—is sensitive to choice of reference class. I hasten to add that there are additional factors (notably timing of evolutionary transitions, convergent

[140] Cf. Carter (1983).

evolution of prerequisites for intelligence, and Fermi's paradox) that can radically alter the import of observation selection effects on a given choice of reference class.[141]

12.3 Simulation and the Doomsday Argument

The *doomsday argument* holds that taking into account our birth rank should lead us to expect that we will go extinct sooner than we would have otherwise thought.[142] Ćirković (2008: 129-30) helpfully summarizes the argument with the following analogy:

> [Suppose there are] two… urns in front of you, one of which you know contains **ten** balls, the other a **million**, but you do not know which is which. The balls in each urn are numbered 1, 2, 3, 4, … Now take one ball at random from the left urn; it shows the number 7. This clearly is a strong indication that the left urn contains only ten balls…. Now consider the case where instead of two urns you have two possible models of humanity's future, and instead of balls you have human individuals, ranked according to birth order. One model suggests that the human race will soon become extinct (or at least that the number of individuals will be greatly reduced) , and as a consequence the total number of humans that ever will have existed is about **100 billion**…. The other model indicates that humans will colonize other planets, spread through the Galaxy, and continue to exist for many future millennia; we consequently can take the number of humans in this model to be of the order of, say, 10^{18}. As a matter of fact, you happen to find that your rank is about 60 billion. (emphasis mine)

Proponents of the doomsday argument contend that we should reason in the same way about our birth rank as we do about drawing ball #7: our birth rank is more likely on hypotheses on which fewer observers like us will have existed than ones on which observers like us exist in numbers of the sort we would expect if we avoid near-term extinction; since the most plausible way for a smaller number of observers to exist is for

---

[141] See Shulman & Bostrom (2012: §5) and Snyder-Beattie et al. (2021).

[142] The argument traces back at least to Brandon Carter (who did not publish on it). Versions of it were later advanced by Gott (1993) and Leslie (1989: 214; 1996). I work with a formulation that is closer to Leslie's, as his formulation is widely regarded as yielding a more plausible (though still controversial) argument than Gott's. For background and critical discussion of much of the literature on the doomsday argument, see Bostrom (2002*b*: Chs. 6-7). For notable variations of the argument that have received relatively little attention, see Grace (2010), Mogensen (2019*a*), and Turchin (2018).

extinction to happen soon; therefore, our birth rank provides evidence that we will succumb to extinction soon.

The doomsday argument is subject to a few qualifications.

- It does not purport to show that we will go extinct soon, or even that we will probably go extinct soon. Instead, the doomsday argument merely purports to show that our birth rank raises the probability that we will go extinct soon.
- Granting that the argument works, the probability of near-term extinction it pushes one to will depend on how probable one regards near-term extinction before taking the argument into account.
- How probable one regards near-term extinction before taking the argument into account is sensitive to whether one embraces number boosting and/or proportion boosting and, if so, what reference classes one chooses for them and for the doomsday argument.
- Whether the argument works may depend on what other relevant information one has. Even if the argument ordinarily works, it might be screened off by firmer empirical evidence about extinction risk.[143]

I'll now note some interactions between the doomsday argument and the simulation argument.[144]

Like the simulation argument, the doomsday argument relies on an application of the indifference principle: whereas the simulation argument tells you to divide your credence equally among simulated and unsimulated 'locations' in your reference class, the doomsday argument in effect assumes that you should divide your credence equally among birth locations within your reference class. Thus, (dis)confirmation of the indifference principle would tend to (dis)confirm both arguments. However, a challenge to one argument's *application* of the principle need not carry over to the other argument. One reason for this is that it is not obvious that the choices of reference classes presupposed in the arguments' respective applications stand or fall together.

I noted above that number boosting arguably bolsters the simulation argument (via its simulation dominance premise). In contrast, number boosting arguably cancels at least

[143] Cf. Bostrom (2002*b*: 92-3).

[144] See also Lewis (2013) and Richmond (2017).

some of the import of the doomsday argument by boosting our priors in abundant-observer hypotheses.[145]

The doomsday argument can be taken to cast doubt on the simulation argument by raising the probability that we will go extinct before creating many intelligent simulated beings in our reference class. There are several ways this might go.

First, observers like us could generally go extinct before creating any intelligent simulated beings. In this case, given that observers like us would create intelligent simulated beings absent near-term extinction, we will go extinct soon. The plausibility of this outcome is inversely related to the number of civilizations at our level of technological development which contain observers like us: the more such civilizations there are, the less plausible it is that they generally go extinct before creating intelligent simulations.

Second, observers like us could sometimes create intelligent simulated beings, but never in quantities that underwrite the simulation argument. The plausibility of this outcome is also inversely related to the number of civilizations at our level of technological development which contain observers like us. However, this relation is weakened by the fact that the creation of simulated intelligent beings may be strongly technologically coupled with AI risks that tend to result in extinction around the same time as civilizations gain the ability to create intelligent simulated beings.

Third, observers like us could create many intelligent simulated beings albeit ones not in our reference class. In contrast to the previous outcomes, the plausibility of this outcome is not inversely related to the number of civilizations at our level of technological development which contain observers like us. Created simulated beings might fall outside our reference class either because they are unconscious or because they differ from us in some other respect. There is a potentially enormous difference in value between these two possibilities. If we create many simulated beings but not many conscious simulated beings, that would likely be because we mistakenly believe that the simulated beings are unconscious. This scenario would likely involve a catastrophic failure of resource allocation. In the limit, we might set in motion a starfaring civilization, ensure that it will be inhabited by vast populations of intelligent

[145] For discussion and references, see Bostrom (2002*b*: 122-6; Ch. 7, fn11).

simulations, and inadvertently engineer the cessation of value through a design flaw that renders the simulations unconscious. A potential alternative would be to engineer our own extinction and replacement through conscious simulations that fall outside our reference class. For those who place substantial weight on the doomsday argument, there is reason to pursue this outcome, as it seems to be one of the only ones that is reasonably probable and positive by the lights of the doomsday argument. That said, whether the argument deems this outcome reasonably probable turns on the unresolved issue of delineating our reference class.

The simulation argument can also be taken to cast doubt on the doomsday argument by revealing a way in which we might have a typical birth rank in a civilization that is not on the brink of extinction: if the relevant reference class concerns our simulators as well as intelligent simulated beings in other simulations, we might have a typical birth rank in our reference class even if we have a very early birth rank among intelligent beings in our universe-simulation. On the other hand, it is not clear why it would be illicit to run two versions of the doomsday argument: one for a broader reference class that encompasses all beings with minds like ours and another for a narrower reference class that includes only the beings in our universe. The latter version of the doomsday argument would at least not be blocked by the simulation argument, since it would reveal a way in which we might have a typical birth rank within our universe-simulation.

12.4 The Fermi Paradox

Recall that the Fermi paradox is the problem of explaining why we seem to be alone in the universe, given the apparently astronomical number of opportunities for life and advanced civilizations to emerge. In §3.2 we saw that the Fermi paradox points to catastrophic risks. It does this by raising the possibility that we seem to be alone because civilizations generally go extinct before becoming observable by civilizations elsewhere in the universe. We also saw that simulation-based research on Fermi's paradox is one way simulations could be brought to bear on catastrophic risks. In this subsection, I explore other connections between Fermi's paradox, simulations, and catastrophic risks. Specifically, I will describe how some interactions between Fermi's paradox and the simulation argument bear on catastrophic risks.

- *Partial simulation solution*. One candidate solution to Fermi's paradox suggested by the simulation argument is that we're in a simulation-universe that is smaller than it appears.[146] One natural way of spelling out the solution contends that, due to resource constraints, our simulators run partial simulations that fill in the details of the universe as required to keep up appearances of a universe for the simulation inhabitants; as a related measure, they only simulate observers on one planet. On this development of the simulation solution, there were far fewer opportunities for intelligence to emerge in our universe than our astronomical observations suggest, and we are not alone (as our simulators exist as well, along with whatever other intelligent beings exist at their level of reality).
- *Silent simulation solution*: Another candidate solution is that long-lasting civilizations, rather than engaging in observable activities such as space colonization, instead occupy themselves with unobservable activities that unfold within virtual environments of their own making.[147]
    - The plausibility of this solution depends on the vexed issue of how many advanced civilizations we should expect there to be within the observable universe: the higher the number, the less plausible it is that all potentially observable civilizations would be unobservable for this reason.
    - The plausibility of this solution also depends on what reasons civilization would have to engage in silent simulation rather than detectable activities. Here are a couple of notable potential motivations:
        - *Self-interested risk mitigation*. Remaining unobservable is an obvious strategy to avoid threats to civilizations posed by hostile actors that may lurk in our universe.[148] Silent simulation may be an optimal means for realizing value while remaining unobservable. That this motivation would be widespread becomes more plausible on the assumption that civilizations that survive long enough to become technologically mature tend to be risk-averse.
        - *Fairness / altruistic risk mitigation*. For civilizations that face Fermi's paradox, it is an ominous warning sign of existential catastrophes. Any civilization that managed to avoid these catastrophes would be confronted with an issue of cosmic fairness: while becoming a

[146] For discussion, see Ćirković (2018: §4.5).

[147] See Dainton (2020: 220).

[148] This sort of strategy was (famously) promoted by Stephen Hawking—see, e.g., BBC (2010).

space faring civilization might be of enormous benefit to it, doing so would dissolve (or attenuate) Fermi's paradox for other, less technologically advanced civilizations. Removing such a warning sign would therefore put other civilizations at a disadvantage and raise the catastrophic risks faced by such civilizations. That this motivation would be widespread becomes more plausible on the assumption that civilizations that survive long enough to become technologically mature tend to act in accordance with fairness or altruistic values.

    - To the extent that the silent simulation solution is motivated, it provides an additional reason to think that advanced civilizations would run simulations. Hence, the availability of this solution arguably bolsters the simulation argument.

- *Shutdown risk*. If we had awoken to a world evidently teaming with advanced civilizations, that would have been good news about the risk of simulation shutdown: if simulation shutdown were something that civilizations had a remote chance of triggering, we'd expect some other civilization to have triggered it before we came into existence. So the fact that many other advanced civilizations existed would suggest that the risk of simulation shutdown is low. Conversely, instead awakening to the Fermi paradox is bad news, as doing so provides no assurances against the risk of simulation shutdown. For the same reasons, solutions to Fermi's paradox that posit many (unobserved) advanced civilizations will tend to lower the risk of simulation shutdown while solutions that instead posit a small number of advanced civilizations will tend to raise the risk of shutdown.
- *Filter placement and the difficulty of engineering intelligence.* If we had awoken to a world replete with many independently evolved species of human-like intelligence, that would have been evidence that (per the argument in §12.2) that engineering human-level intelligence is not an exceedingly hard problem. (The matter would not have been completely cut and dry: in light of observation selection effects, there'd be room to wonder if evolution almost never produces such species but reliably produces many if it produces any.) Conversely, failing to observe the traces of such species when we gaze at the cosmos through telescopes is evidence that:
    - producing such species is difficult for evolution,

    - (hence) engineering human-level intelligence is difficult for human-level intelligent beings (ourselves included),
    - (hence) we will go extinct before being able to produce simulations with human-level intelligence, and
    - (hence, *pace* the simulation argument) beings like ourselves generally fail to produce many more simulations that are beings like ourselves.[149]
- *Rare Earth solutions.* Rare Earth solutions hold that the universe appears devoid of other civilizations because an Early Filter renders life (or complex life or intelligent life) rare—rare enough for it to be unsurprising that there are no other civilizations for us to observe.[150]
    - Proposed Rare Earth solutions tend to render it very unlikely that there would be even one planet that gives rise to (intelligent) life.
        - For illustration: some estimates for the odds of life arising on a given Earth-like planet include figures like 1 in $10^{40,000}$ and $10^{100,000,000,151}$ while one estimate of the number of planets within the observable universe puts it on the order of 10^{20}, with most of those planets being life-inhospitable.[152]
    - That Rare Earth solutions tend to render it *very* unlikely that there would be even one planet that gives rise to (intelligent) life is not necessarily a bug: given that life is extremely unlikely to arise on any particular Earth-like planet, it would be a striking coincidence if the number of such planets was such as to render the expected number of planets that give rise to intelligent life/civilizations ~1.[153]

---

[149] The import of this evidence is, however, blocked by Early Filter solutions on which the dearth of intelligent life is explained by a filter that renders life rare. Compare Armstrong (2014): "The Great Filter is early, or AI is hard". If correct, this points to a worrying corollary of progress in AI: insofar as it gives us reason to think engineering human-level AI is easier than we thought, it also gives us reason to doubt Early Filter solutions to Fermi's paradox and hence to be more confident that the Filter lies in our future. But see Miller (2019) for argument that evidence for both an Early Filter and AI-based existential risks is less concerning than evidence for either alone would be. See also Hanson et al. (2021) for a solution to Fermi's paradox that lends to an explanation of how an Early Filter could cohere with human-level AI being relatively easy to engineer.

[150] See Ward & Brownlee (2000).

[151] See Monton (2009: 99) for references.

[152] See Zackrisson et al. (2016).

[153] Cf. Carter (1983). Note, however, that it remains a live possibility that the universe has an infinite number of Earth-like planets and that this would in effect render it certain that there are other such

- On the other hand, it is hard to believe that intelligent life—being the striking phenomenon that it is—just happened to arise if the chances or its doing so were miniscule. That would be a form of unexplained fine-tuning akin to physical parameter values all just happening to fall within the narrow ranges required for life.
- Just as such cosmological fine-tuning would arguably be more likely given a designer or multiverse, so too would a Very Rare Earth. Thus, rare earth solutions to Fermi's paradox arguably support multiverse and design hypotheses. For reasons encountered in §11, multiverse and design hypotheses fit with the simulation hypothesis. Moreover, on the face of it, nothing about the simulation versions of the design and multiverse hypotheses renders them worse off at explaining the existence of a Very Rare Earth than their non-simulation counterparts.
- Thus, Rare Earth solutions to Fermi's paradox arguably support the simulation hypothesis.

- *Number boosting.* Recall that number boosting tells us to favor hypotheses on which there are more observers (like us).
  - Given the vastness of our universe, one application of number boosting tells us to favor hypotheses on which there are more observers, and hence to favor solutions to Fermi's paradox on which the apparent dearth of observers is misleading.
    - These include simulation solutions on which many observers exist but we are not in a position to observe them because they inhabit simulations.
  - If we take the apparent dearth of observers at face value, this provides grounds for questioning number boosting. This arguably weakens the simulation argument, since, as seen in §12.1, number boosting arguably bolsters the simulation argument.

---

planets that give rise to intelligent life no matter how unlikely it is that a given planet would—see Monton (2009: 102-4). The existence of an infinite number of observers also raises technical difficulties for the application of indifference principles—see Bostrom (2002*b*), Dorr & Arntzenius (2017), and White (2018).

12.5 Boltzmannian Cosmologies

Physics offers a number of otherwise appealing cosmological models that have a peculiar consequence: they predict that the vast majority of systems in brain states like ours are "Boltzmann brains", short-lived brains that result from thermal or quantum fluctuations.[154] When taken with some innocuous-seeming assumptions, these theories yield the skeptical conclusion that we ourselves are probably Boltzmann Brains. In particular, if we assume that what experience a system has is fixed by its brain state, then these theories predict that almost all observers with our experiences are Boltzmann brains. An application of the indifference principle then dictates that we should divide our credences evenly among observers with our experiences and conclude that, on these theories, we are almost certainly Boltzmann brains. Avoiding this skeptical conclusion is *the Boltzmann brain problem*.

There is much controversy about what to make of the implications of these theories. Here, I will describe some interactions between the Boltzmann brain problem and the simulation argument.[155] As far as I can tell, the main bearing that the Boltzmann brain problem has on catastrophic risks is by way of these interactions.[156]

- The Boltzmann brain problem shows how the simulation argument could fail even granting the indifference principle and that ordinary observers like us are vastly outnumbered by observers like us in simulations: given that simulated observers would themselves be vastly outnumbered by Boltzmannian observers, the indifference principle would require us to think we're probably Boltzmannian observers rather than observers in a simulation.
- Gaining evidence for Boltzmannian cosmologies gives those who wish to resist skepticism reason to revisit the assumption that brain states fix experiences, as well as the indifference principle.

[154] For background on the physical theories, see Carroll (2020). For discussion, see Kotzen (2020).

[155] For discussion of similarities between the simulation argument and the Boltzmann brain problem, see Crawford (2013).

[156] Of course, a world in which almost all observers like us are Boltzmann brains would itself arguably involve catastrophe in the form of premature death on a cosmic scale. However, if there is a real risk of such a catastrophe, the catastrophe is presumably already underway and there is nothing we can do to mitigate it, with the possible exception of non-causal interventions like those discussed in §13. In any event, I will set this sort of catastrophe aside.

- Since the simulation argument relies on closely related assumptions, such evidence would also give us reason to revisit the simulation argument.
  - *Indifference-rejecting solutions to the Boltzmann brain problem constrain indifference-rejecting responses to the simulation argument.* Assuming that brain states fix experiences, Boltzmannian cosmologies predict that almost all observers with experiences like ours are Boltzmann brains on virtually any precisification of 'like ours'. In contrast, as we've seen, the simulation argument's application of the principle of indifference is sensitive to choice of reference class. This yields the following asymmetry: whereas rejecting the simulation argument based on illicit choice of reference class will tend to leave the Boltzmann brain problem unscathed, solving the Boltzmann brain problem based on illicit choice of reference class will tend to generate an objection to the simulation argument.
    - This asymmetry is subject to an important qualification: rejecting the application of the indifference principle in the Boltzmann brain problem on the ground that it leads to skepticism does not jeopardize the simulation argument's application of the indifference principle—at least given that the simulation hypothesis is not a skeptical one.
  - *Zombification solutions to the Boltzmann brain problem undermine the simulation argument.* Zombification solutions to the Boltzmann brain problem hold that Boltzmann brains would be unconscious ('zombies') and therefore not observers like us, meaning that the fact that we're conscious entails that we're not Boltzmann brains. While the Boltzmann brain problem is typically couched in terms of brain states, it is robust under choices of much larger states, including states that involve brain states, bodies, and local environments.[157] Thus, correspondingly robust zombification solutions require experience to be fixed by very large (perhaps global) physical states, rather than any sort of intrinsic or local physical state. If experience has such a large physical basis, it becomes doubtful that duplicating causal structure in the brain in a simulation would suffice to duplicate experiences. Thus,

[157] See Chen (forthcoming).

zombification solutions tell against the simulation argument by giving us a reason to think that simulations would be unconscious.[158]

- *Instability again.* In §11 we saw that some have objected to the simulation argument on the ground that it is cognitively unstable. The same objection has been leveled against Boltzmannian cosmologies.[159] In the context of Boltzmannian cosmologies, the objection is that evidence cannot be stably taken to support such cosmologies since supporting such cosmologies would indicate that we are Boltzmann brains and therefore not subjects who have such evidence. At least some of the reasons given for doubting that the objection works against the simulation argument also suggest that they fail against Boltzmannian cosmologies.
  - For example, take the hypoxia cases in which it is irrational to disregard the cognitively unstable hypothesis that one is hypoxic. These cases show that cognitive instability is generally insufficient as a ground for rejecting a hypothesis.
  - And just as the simulation argument can be recast in terms of stable simulations, so too can the Boltzmann brain problem be recast in terms of *stable Boltzmannian bubbles*, localities that feature observers that are like us in local physical respects and which are large enough and long-lasting enough to avoid rendering the reasoning in the Boltzmann brain problem unstable.

- *Simulation solutions to the Boltzmann brain problem.* We saw in §12.4 that there is reason to think we're in a partial simulation, on which details of our universe-simulation are filled in as we observe, making our universe appear much larger and more detailed than it is. This suggests the following *partial simulation solutions* to the Boltzmann brain problem: it may be that Boltzmann observers do not numerically dominate ordinary observers because we are living in a partial simulation and Boltzmann brains would either (i) exist in the (apparent) parts of our universe that are not simulated, (ii) exist in parts of our universe that are simulated in too little detail to realize consciousness, or (iii) be unconscious because of differences that hold between them and us outside the simulation.

[158] See Saad (forthcoming*c*).

[159] See Carroll (2020).

- *Number boosting.* Number boosting tends to favor Boltzmannian cosmologies over non-Boltzmannian cosmologies and therefore engenders skepticism, absent a solution to the Boltzmann brain problem. This could be taken as a reason to reject number boosting. Since number boosting arguably strengthens the simulation argument, the fact that number boosting tends to favor Boltzmannian cosmologies can also be taken as a reason for reducing confidence in the simulation argument.

13 Simulation as Non-Causal Intervention

This section will introduce *simulations as non-causal interventions*, which I regard as an important and neglected factor with the potential to significantly influence risk levels for a wide range of catastrophes. To a first approximation, the idea is that by raising the probability that certain types of simulations will be run, we can change the probability of our currently inhabiting those types of simulations and (perhaps counterintuitively) thereby meaningfully but non-causally affect the probability of various types of catastrophic risks. The concept will require some unpacking, as it combines insights from the simulation argument and some recent work in decision theory.[160]

To start, recall the simulation argument's premise of simulation dominance: at least a small portion of beings like us will produce a very large number of beings like us that are simulated—a large enough number for most beings like us to turn out to be simulations. The main ways for this premise to be false would be if beings like us went extinct before we created beings like us that are simulations or if we decided never to create such beings. Even if plausible, this premise is at present a speculative empirical claim. So we should not be extremely confident in it. However, this could change: if we began mass producing simulations that realized beings like us, that would give us a powerful reason to accept simulation dominance. Those who have a high credence in the indifference principle (the other premise in the simulation argument) would then be under strong pressure to conclude from the argument that we are probably living in a simulation.[161] Likewise, acquiring evidence that we will mass produce such simulations

[160] To my knowledge, nothing has been published on reducing catastrophic risks by using simulations as non-causal interventions. However, after submitting a research proposal to the Center on Long-Term Risk on the topic, I was informed that the idea had already been considered there. I plan to expand the ideas in this section into a stand-alone paper.

[161] Cf. Bostrom (2003: 253).

should boost our confidence in simulation dominance and in turn the simulation hypothesis via the simulation argument.

Next, consider that there is a family of variations of the simulation argument that concern different types of simulations. For example, there are *salvation simulations* in which minds like ours are immune to catastrophic risks but have much misleading evidence to the contrary. There are also *doom simulations* in which it is inevitable that minds like ours will collectively succumb to certain catastrophic risks, regardless of their evidence concerning risk levels and regardless of their risk mitigation efforts. And there are myriad other sorts of simulations in which minds like ours face risks that are bizarrely related to their evidence and actions. Just as evidence that we will one day mass produce simulations with minds like ours should boost our confidence that we are in a simulation, so too should evidence that we will one day mass produce simulations of a given type that contain minds like ours boost our confidence that we are in a simulation of that type. Thus, evidence that we will run such salvation simulations would be good news while evidence that we will run such doom simulations would be bad news. And we are in a position to choose what type of news we receive—for example, setting up a fund to sponsor salvation simulations when they become technologically feasible would be a way of bringing about good news about the catastrophic risks we face.

Of course, we are either in a given type of simulation or we are not. And we control neither whether we are in a simulation nor which type of simulation we are in if we are in one. Thus, there is no way to causally exploit the described connection between our advancing the creation of certain sorts of simulations and our inhabiting such simulations. So it may seem that this connection is not decision-relevant in the context of catastrophic risks. Endorsing this appearance would be to take a stand on an ongoing debate in decision theory. The debate concerns whether *causal decision theory* is correct, rather than one of its rivals such as *evidential decision theory*. Roughly, whereas "[evidential decision theory] tells you to perform the action that would be the best indication of a good outcome, … [causal decision theory] tells you to perform the action that will tend to causally bring about good outcomes" (Levinstein & Soares, 2020).

To briefly illustrate, suppose you are deciding whether to vote. Holding fixed how everyone else votes, you know (let's suppose) that your vote won't make a difference as to who wins and hence that you could cause more good by doing something else instead of voting. You also know that the people like you will vote for the better candidate just in case you vote for that candidate, and that that candidate will probably win just in case

the people like you vote for her. In this case evidential decision theory recommends voting, while causal decision theory recommends not voting.[162] Similarly, evidential decision theory will tend to recommend running salvation simulations, preventing doom simulations, and bringing about evidence that the former will be run and the latter prevented. On evidential decision theory, such actions affect catastrophic risk levels in a decision relevant sense without causally affecting them—in this sense, evidential decision theory licenses non-causal interventions. In contrast, causal decision theory will tend to judge these actions unworthy of promotion and will recommend allocating our resources elsewhere should these actions come with even the slightest opportunity cost.

Causal decision theory seems more widely favored than evidential decision theory.[163] And it might seem that non-causal interventions can influence the expected value of outcomes in decision relevant ways only to those who accept evidential decision theory. Putting these two points together, one might be tempted to conclude that simulations as non-causal interventions bear on catastrophic risks in decision-relevant ways only given a certain minority view (evidential decision theory). Contrary to this response, simulations as non-causal interventions have broader significance. This is for several reasons:

- While evidential decision theory is perhaps the most often discussed rival to causal decision theory, it is not the only rival. There are various other non-causal decision theories that agree with evidential decision theory in recommending non-causal interventions via simulations.[164]
- Perhaps pluralism is true of rationality and different decision theories accurately describe different forms of rationality corresponding to different foci of evaluation.[165]

---

[162] See Leslie (1991).

[163] In a recent survey of professional philosophers, participants were asked about their view of Newcomb's problem, a case that is standardly taken to elicit different recommendations from causal decision theory and evidential decision theory. 31.2% favored the response standardly associated with evidential decision theory while 39.0% favored the response standardly associated with causal decision theory (Bourget & Chalmers, 2021). It is thus natural to interpret these results as indicating that causal decision theory is more widely favored. However, there is also debate about whether Newcomb's problem bears on causal decision theory and evidential decision theory in the way that is standardly supposed—see Knab (2019: Ch. 3) for references and reasons to think not.

[164] See Easwaran (2021) for an illuminating taxonomy and references.

[165] See Easwaran (2021: §3.2), Kagan (2000), and Nozick (1993).

- We should be uncertain about whether causal decision theory is correct. Arguably, factoring this uncertainty into our decision making should make us act in accordance with evidential decision theory's recommendations concerning non-causal interventions even if we think causal decision theory is probably true instead.
    - We should be uncertain about which decision theory is correct because decision theory is a difficult and highly technical subject and there is disagreement among relevant experts about whether causal decision theory is correct.
    - Given uncertainty about whether causal decision theory is correct, there is a question of how to factor in our decision-theoretic uncertainty when deciding what to do.
    - One natural option is to use *meta decision theory*. It recommends taking the option with the highest meta expected value, where an action's meta expected value = the sum of expected values assigned to it by rival first-order decision theories weighted by their probability of being correct.[166] By the lights of meta decision theory, there can be strong reasons to act in accordance with the recommendations of evidential decision theory (or other non-causal decision theories) rather than causal decision theory even if we are much more confident in causal decision theory.
        - Meta decision theory leads to *the evidentialist's wager*, which contends that we should in this fashion act against the odds in a wide range of cases in which causal decision theory and evidential decision theory issue conflicting recommendations.[167] The basic reason for this is that evidential decision theory channels expected value to acts via agents in similar decision contexts whose acts are non-causally correlated with that of the actor (for instance, evidential decision theory amplifies the expected value of your voting, given that your voting gives you strong evidence of how similarly minded agents whose decisions are non-causally correlated with your own will act); since causal decision theory does no such thing, the meta expected value of evidential decision

[166] See MacAskill (2016).

[167] See MacAskill et al. (2021).

theory's recommendations tend to swamp causal decision theory's in such cases. Running salvation simulations is a case in point: taking actions to raise the probability of running salvation simulations that contain many minds like ours raises the probability that the same sort of action will be taken by the many other agents in advanced civilizations whose decisions are non-causally correlated with our own; by the above argument, that action thereby diminishes the probability of catastrophic risks and so accrues much expected value by the lights of evidential but not causal decision theory. Thus, the evidentialist's wager applies in this context: even if we are confident but not certain that causal decision theory is correct, we still have strong reason to act on evidential decision theory's recommendation to raise the probability that we will run salvation simulations, since this would in turn raise the probability that we are in a salvation simulation and so reduce catastrophic risks. Similar reasoning applies to other non-causal interventions through simulations of observers like us.

- Meta decision theory is not the only option for dealing with decision theoretic uncertainty.
  - Some alternatives recommend in effect ignoring decision theoretic uncertainty and acting in accordance with the decision theory you believe, have the highest credence in, is most supported by your evidence, or which is correct. Given corresponding auxiliary assumptions about causal decision theory, these proposals will deliver the verdict that you should act in accordance with it rather than any theory that promotes non-causal interventions. However, these proposals are subject to severe limitations and powerful objections. For instance, the proposal that you should act in accordance with the correct theory provides no guidance if you are uncertain which theory is correct. Acting in accordance with the theory you have the highest credence allows small differences in credence to outweigh arbitrarily large differences in stakes. Acting in accordance with the theory you believe precludes dominance reasoning in cases where that theory is indifferent between two

options and the other theories you have credence in agree about which option you should take.[168]

- Other alternatives to meta decision theory agree with meta decision theory that we should factor in decision theoretic uncertainty but propose a different way of taking it into account. These include variations of meta decision theory that incorporate an extra parameter for risk aversion or to correct for counterintuitive effects of certain theories swamping other theories by assigning much higher stakes.[169] They also include bargaining and parliamentary proposals that treat different theories as if they were engaged in moral trade or voting members in a parliament.[170] On the face of it, we'd expect these and other alternatives to meta decision theory that take decision theoretic uncertainty into account and which are at least somewhat stake sensitive to promote non-causal interventions via (evidence for) simulations. The space of such theories is underexplored, as is the application of such theories to non-causal interventions via simulations.

Some might take the promotion and prevention of certain simulations as non-causal interventions as an implausible result that indicates that the argument has gone wrong somewhere. My own view is that this reaction is understandable but unwarranted. Admittedly, that simulations as non-causal interventions should be promoted is a weird and wacky idea. However, this comes with the territory: it's no surprise that interactions between decision theory, decision theoretic uncertainty, self-locating belief, and simulation hypotheses have bizarre consequences. If there is something especially objectionable about this one, it remains to be specified. I expect others to disagree here. So there is a project of exploring the prospects for logically weakening the argument while preserving its upshot along with the prospects for a robust escape from the argument.

If simulation as non-causal interventions is admitted as relevant to analyzing and seeking to reduce catastrophic risks, how should we devise a risk-reduction portfolio

[168] See Lockhart (2000), Bykvyst (2017), and MacAskill et al. (2020).

[169] For a decision theory that treats risk aversion as a basic parameter, see Buchak (2013).

[170] Respectively, see Greaves & Cotton-Barratt (2019) and Newberry & Ord (2021).

that is appropriately sensitive to them? This is a large question that I cannot hope to settle here. This is partly because of the residual issues concerning how to deal with decision theoretic uncertainty that we encountered above, and partly because addressing it would require a systematic investigation of what sorts of non-causal interventions via simulation are possible, their expected (causal and non-causal, descriptive and axiological) consequences. While I will not attempt such an investigation here, I will lay out some considerations that would need to be addressed in such an investigation.

First, we need to consider various types of non-causal interventions that could affect catastrophic risk levels via simulation hypotheses about different types of simulations. Types to consider include not only salvation and doom simulations, but also:

- Research simulations
- Entertainment simulations
- Catastrophic simulations
- Different types of salvation simulations
    - Compensatory simulations in which the good and badness of the simulation is set in advance and minds are compensated in a simulation afterlife to achieve the preset value.
    - Miraculous simulations in which certain sorts of would-be catastrophes are miraculously prevented at the last moment.
    - Merely apparent suffering simulations in which the vast majority of apparently suffering minds are either not minds or not suffering.
    - Simulations that terminate once the overall expected value of their continuation becomes negative
- Debunking simulations
    - For a given belief, we could undermine it by running simulations of minds like ours in which that belief has a debunking causal origin, thereby raising the probability that our belief has a debunking causal origin.
- Simulations that vindicate/falsify different philosophical views.
    - Personal identity simulations: by running simulations in which a certain view of personal identity is true, we could raise the probability that it is true. Some views of personal identity may have the potential to influence individual's levels of altruism—for example, the distinction between self and others is diminished on some eliminativist and reductionist views of

personal identity.[171] By bringing it about that we are probably in a simulation in which one of those views is true of us and broadcasting that result thus constitutes a strategy for persuading people to be less self-concerned.

    - Free will and anti-free will simulations: by running simulations in which minds like ours have (lack) free will, we would raise (lower) the probability that we have free will. E.g. if free will is less likely given determinism, we might lower the probability that we have free will by running deterministic simulations containing minds like ours.
    - By running simulations that conform to a certain model of time, we would raise (lower) the probability that we are in a simulation that conforms to that model. Arguably, different models of time have different axiological consequences.[172] So modifying the probability of different models can be expected to affect the distribution of value in our universe.
    - One of the leading interpretations of quantum mechanics (the many worlds interpretation) has the unsettling consequence (given auxiliary assumptions about personal identity) that each of us should expect to survive with certainty arbitrarily long into the future, probably via means that involve great suffering.[173] By running simulations of universes inhabited by minds like ours and underpinned by an interpretation of quantum mechanics that does not have the unsettling consequence, we could lower the probability that we are subject to it.
    - Running simulations in which minds like ours are skeptically situated would arguably raise the probability that we are skeptically situated in such a simulation (subject to complications involving cognitive instability discussed in §8)

- Anti-simulation hypothesis simulations are simulations such that running them (and evidence that we will run them) lower the probability that we are in a simulation. This would tend to lower simulation termination risks for us.
    - Totem simulations: in such simulations, minds like ours would generally have a "totem", a test that they could easily run to tell that they are in a simulation.

[171] See Parfit (1984: §95).

[172] See Saad (forthcoming*b*).

[173] See Lewis (2004).

    - Unconscious simulations: if we ensured that any simulations of minds like ours are unconscious, that would cast doubt on the simulation dominance premise and so give us reason to reduce our confidence in the simulation hypothesis.
    - Anti-nesting simulations: if we run simulations in which minds like ours cannot create minds like ours via simulation, that would cast doubt on the simulation dominance premise and so give us reason to reduce our confidence in the simulation hypothesis.
    - Skeptical simulations: running simulations in which minds like ours are skeptically situated would arguably raise the probability that we are in a skeptical situation if we are in a simulation. If we have independent warrant for believing we are not in a skeptical situation,[174] this would then tell against our being in a simulation.
- Level coordination simulations: by running simulations that are coordinated with our simulation with respect to certain features (e.g. physics), we could raise the probability that we are in a simulation that shares those features with the level of its simulators.[175] By requiring these features to be strictly preserved in nested simulations, we could raise the probability that they are preserved in any simulations in which ours is nested.

If simulation as non-causal interventions is admitted as relevant to analyzing and seeking to reduce catastrophic risks, there are some potential mistakes that we should take care to avoid or else ensure that they are not in fact mistakes. These include:

- Researching catastrophic risks through simulations of minds like ours facing such risks. This would raise the probability that we're in such a simulation and hence that we will succumb to catastrophe. (To avoid this mistake, we could run only unconscious simulations or ensure that catastrophic simulations are offset by sufficiently many non-catastrophic simulations, though the latter option would be morally objectionable.)
- Running simulations in which minds like ours are skeptically situated could constitute an epistemic catastrophe for us. The extent to which such epistemic catastrophe is a real risk is unclear: maybe we would be warranted in concluding that simulations are unconscious if we found out that most minds like ours would

[174] Cf. White (2005) and Wright (2004).

[175] Cf. Dainton (2012: 68).

be skeptically situated if simulations were conscious. Or maybe epistemic catastrophes would be psychologically ignored and as a result not be catastrophes from a moral or all-things-considered perspective.

- Running simulations that induce cognitive instability (such that we cannot form a coherent view about our evidence and whether we are in a certain type of simulation for which we apparently have evidence) could undermine our ability to evaluate and reduce catastrophic risks, perhaps while leaving our epistemic standing in other domains intact.
- Running simulations in which minds like ours are short-lived would reduce our life expectancy. The same goes for our civilization's life expectancy if we run simulations with civilizations that are short-lived and which contain minds like ours.
- Some catastrophes would—if they occurred at all—occur before the creation of salvation simulations. The risk of such catastrophes is subject to screening off effects that severely limit the potential for reducing them by raising the probability of salvation simulations. For such catastrophes, interventions that raise the probability of salvation simulations in which they are avoided will do so only conditional on the catastrophe not happening—but, conditional on the catastrophe not happening, its probability of happening is insensitive to the probability of salvation simulations. Failing to take this into account would encourage misallocation of resources to salvation simulations.
- Neglecting world-scale when analyzing simulations as non-causal interventions.
    - The most obvious way this could happen would be if we initially analyzed catastrophic risks in our universe and then considered the impact of simulation hypotheses and simulations as non-causal interventions on *those* risks without considering their impact on the world that contains our simulation along with our simulator's level of reality and whatever else exists. Taking the latter into account could be crucial on some views of value, notably those on which a world's value doesn't scale with good- and bad- making features. For example, suppose there is an upper bound on how much value good-making features in a world can collectively generate but no bound on how much disvalue bad-making features in a world can generate. On this view, finding out that our universe is a simulation in a world that likely contains many other simulations would suggest that there is almost certainly no way for us to make the world better by

increasing the number of good making features but much we can do to improve the world by reducing the number of bad-making features that our universe will contain. This view would then give us reason to, for example, prioritize preventing a future with catastrophic quantities of suffering over bringing about a grand future. Similarly, views on which scaling up good- and bad-making features equally tends to diminish the value of a world would give us reason to run anti-simulation hypotheses so as to reduce the expected size of our world and lessen this unwanted scaling effect.

14. Open Questions and Avenues for Future Research

By way of conclusion, I'll highlight what I regard as some key open questions we've encountered along with some promising avenues for future research. I'll classify these in a rough and ready way by research area. A disproportionate number of questions and research avenues will belong to areas of philosophy, which is my own field of expertise. This merely reflects my being in a better position to identify promising topics in philosophy. This collection is by no means comprehensive, and I would be delighted if others improved upon it.

- Epistemology
    - How should unstable evidence for the simulation hypothesis be handled? Under what conditions if any does the fact that a hypothesis renders evidence unstable mean that the hypothesis should be ignored?
    - The reference problem for observation selection effects has largely been approached through a priori arguments. There is a project of teasing out empirical consequences of different reference class hypotheses, comparing them with our evidence, and tracing the impact on associated risks via Fermi's paradox, the simulation argument, the doomsday argument, and evolutionary arguments for easy AI.
    - The literature on self-locating belief in the last two decades is sprawling, technical, and lacking in uniform terminology. This suggests a few projects that could facilitate analysis of catastrophic risks that interact with self-locating beliefs:
        - It would be helpful to have an accessible and up-to-date synthesis of the self-locating belief literature.

        - A critical mass of researchers could agree to terminological conventions concerning self-locating belief and announce this to relevant research communities.
        - A sensitivity analysis of how different interactions between self-locating belief and catastrophic risks are fragile/robust under different views about self-locating belief could provide strategic guidance even without resolving debates surrounding self-locating belief.
    - A sensitivity analysis of how different views of general accounts about justification and evidence (ones not specifically concerned with self-locating information) bear on the simulation argument.
    - Formulating and evaluating the simulation argument in terms of evidentially stable simulations
- Philosophy of mind / cognitive science / AI
    - What tests should we use to assess whether simulations are conscious?
    - What criteria should we use when evaluating whether simulations are conscious and, if so, what sorts of experience they have?
    - What can we do to reduce the probability that simulated systems we create are conscious?
    - What can we do to reduce the risk that simulated systems we create suffer?
    - What can we do to increase the probability that simulated systems we create have positively valenced experiences?
    - How does psychophysical fine-tuning evidence bear on the simulation hypothesis and associated catastrophic risks?
- Philosophy of physics
    - Interactions between the simulation argument and the Boltzmann brain problem
    - Interactions between cosmological and planetary fine-tuning arguments and the simulation hypothesis
- Metaethics
    - Which evolutionary debunking arguments are amenable to simulation testing? How can we test them? What is the role of such tests in solving the alignment problem?
- Ethics

    - Aside from hedonic constraints, what other ethical constraints need to be respected in creating simulations containing minds? What would be best practices for satisfying these constraints?
    - Sensitivity analysis of how bad shutdown would be on different ethical theories
    - Taxonomy of theories in population ethics that lead to downside focus or asymmetrically diminishing returns in large-worlds, along with those theories' motivations and problems
- Philosophy of Religion
    - How do different simulation solutions to problems for design hypotheses affect the probability of religious catastrophes?
    - How does the simulation hypothesis interact with arguments for/against different design hypotheses and associated catastrophic risks? For instance, which arguments (if successful) support the existence of God but not the existence of a non-divine simulator?
- Decision theory
    - Research on the following could put us in a better position to evaluate the expected value of non-causal simulation interventions.
        - Adjudicating between causal vs. non-causal decision theories
        - Systematic exploration of the space of decision theories (in contrast to narrow focus on causal vs. evidential vs. functional decision theory)
        - What is the best version of the evidentialist wager argument for non-causal simulation interventions?
        - How robust is that argument to different decision-theoretic assumptions concerning both first-order and higher-order theories?
        - What should we make of arguments for causal and non-causal theories sharing predictions in a wider range of cases than is standardly supposed? If any of these arguments work, under what conditions do these theories in fact yield the same predictions?
- Forecasting
    - Taxonomy of technologies that are relevant to both simulation and catastrophic risk
    - Forecasting development timelines for these technologies
    - Forecasting use of these technologies conditional on arrival

    - Forecasting catastrophe conditional on the use of these technologies
- Risk modeling
    - How can simulations be used to yield better estimates of catastrophic risks and the efficacy of candidate interventions?
    - What are the main risk factors for misestimation? How can these be mitigated?
    - How can existing risk models be adapted to take into account information about changes in information about whether we're in a simulation
- Strategic analysis
    - Shutdown risk. Research on the following could put us in a better position to evaluate and mitigate shutdown risks:
        - How can we fruitfully taxonomize shutdown risks?
        - How credible are different shutdown risks?
        - How can different shutdown risks be mitigated?
        - How do shutdown risks interact with other risks?
    - How do different simulation scenarios impact religious catastrophic risks?
        - Systematic exploration religious catastrophic risks
        - Systematic exploration of interactions between simulation and religious catastrophic risks
    - Developing non-causal simulation intervention as a strategy for risk reduction
        - What type of interventions to use?
        - Are there any viable near-term non-causal interventions?
        - What's the relative importance of different non-causal interventions vs. other such interventions and vs. causal interventions?
    - What are the risk-reducing prospects for simulation refuges and fallbacks? How can these be improved?
    - What sort of differential technological progress would reduce the catastrophic risks associated with simulation? How can we induce such progress?
        - To what extent does the potential for simulation refuges and fallbacks to reduce risks tell in favor of speeding up simulation technology?

    - What sorts of policies would reduce simulation-related catastrophic risks? How would these policies bear on other catastrophic risks?
    - What roles for simulations are promising in grand future scenarios?
- Social science
    - There are various projects of simulating different catastrophic risks and factors that are directly or indirectly relevant to such risks in order to reduce our uncertainty about them and to test risk reduction strategies.
    - There are also various projects of simulating catastrophic risks in order to illustrate realistic catastrophic scenarios in a way that engages public and policymaker attention more than abstract analyses and models.
    - Do games, interactive simulations, and/or immersive simulations that simulate risks to promote risk responsiveness? If so, how can they be designed to better promote risk responsiveness?
    - Simulations that test hypotheses about value dynamics.
    - Simulations of risk scenarios that test hypotheses about the impact of different cognitive capacities and/or biases on navigating those scenarios.
    - Research on using simulations to enhance cognitive capacities and diminish biases that affect catastrophic risk.
- AI Governance
    - What policies should be put in place to reduce the catastrophic risks associated with simulations?
        - What sorts of policies would prevent catastrophic simulations (ones realizing immense quantities of moral disvalue) in the context of games, entertainment, research, digital economies, wars, and power-seeking, morally indifferent, or malevolent actors?
        - What policies should be enacted to reduce catastrophic risks associated with dual uses of simulations?
    - When is the best time to promote such policies?
    - What policies does it make sense to push for now? How should this be done?
- AI Safety
    - Developing nested simulations for safety testing
    - Developing nested simulations for boxing purposes or for the purpose of incentivizing aligned behavior
    - Using simulations to find overlooked threat models and interventions

- Using simulations to vividly demonstrate dangers posed by AI to relevant parties that underestimate these dangers

References

Adams, M. (2000). *Horrendous evils and the goodness of God*. Cornell University Press.

Alexander, S. (2019) Don't Fear The Simulators. Slate Star Codex.

Althaus, D., & Baumann, T. (2020). Reducing long-term risks from malevolent actors. Publications of Center of Long Term Risk.

Anthis, J.R. (2018) "Why I Prioritize Moral Circle Expansion Over Artificial Intelligence Alignment." Effective Altruism Forum. https://forum.effectivealtruism.org/posts/BY8gXSpGijypbGitT/why-i-prioritize-moral-circle-expansion-over-artificial.

Anthis, J.R. (2022) The Future Might Not Be So Great. URL: https://www.sentienceinstitute.org/blog/the-future-might-not-be-so-great

Armstrong, S.; Sandberg, A. & Bostrom, N. (2012). Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines* 22 (4):299-324.

Babcock, J., Kramár, J., & Yampolskiy, R. V. (2019). Guidelines for artificial intelligence containment. *Next-Generation Ethics: Engineering a Better Society*, 90-112.

Banks, I.M. (2010) *Surface Detail*, London: Orbit.

Barak B. & Edelman B. (2022) AI will change the world, but won't take it over by playing "3-dimensional chess". URL: https://www.alignmentforum.org/posts/zB3ukZJqt3pQDw9jz/ai-will-change-the-world-but-won-t-take-it-over-by-playing-3#Our_argument__an_executive_summary_

Barnett, Z. & Li, H. (2016). Conciliationism and merely possible disagreement. *Synthese* 193 (9):1-13.

Baum, S. D., Denkenberger, D. C., & Haqq-Misra, J. (2015). Isolated refuges for surviving global catastrophes. *Futures*, 72, 45-56.

Baumann, T. (2022) *Avoiding the Worst: How to Prevent a Moral Catastrophe* URL: https://centerforreducingsuffering.org/wp-content/uploads/2022/10/Avoiding_The_Worst_final.pdf

BBC News. (2010) "Stephen Hawking warns over making contact with aliens". URL: http://news.bbc.co.uk/2/hi/uk_news/8642558.stm

Beckstead, N. (2015). How much could refuges help us recover from a global catastrophe?. *Futures*, 72, 36-44.

Bedke, (2009). Intuitive non-naturalism meets cosmic coincidence. *Pacific Philosophical Quarterly* 90 (2):188-209.

Benatar, D. (2008). *Better never to have been: The harm of coming into existence*. OUP.

Bennett, C.H., Hanson, R., & Riedel, C.J. (2019). Comment on 'The Aestivation Hypothesis for Resolving Fermi's Paradox'. *Foundations of Physics*, 49(8), 820-829.

Bentham, J. (1791). Panopticon: or, The inspection-house. Containing the idea of a new principle of construction applicable to any sort of establishment, in which persons of any description are to be kept under inspection, etc. Thomas Byrne.

Benton, M. A., Hawthorne, J., & Isaacs, Y. (2016). Evil and evidence. *Oxford Studies in Philosophy of Religion*, Vol. 7, 1-31.

Birch, J. (2013). On the 'simulation argument' and selective scepticism. *Erkenntnis*, 78(1), 95-107.
Bhogal, H. (forthcoming). What's the Coincidence in Debunking? *Philosophy and Phenomenological Research*.
Bogardus, T. (2016). Only All Naturalists Should Worry About Only One Evolutionary Debunking Argument. *Ethics* 126 (3):636-661.
Bostrom, N. (2002*a*). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and technology*, 9.
Bostrom, N. (2002*b*) *Anthropic bias: observation selection effects in science and philosophy*. New York, NY: Routledge.
Bostrom, N. (2003*a*). Are we living in a computer simulation?. *The philosophical quarterly*, 53(211), 243-255.
Bostrom, N. (2003*b*). Ethical issues in advanced artificial intelligence. In *Science fiction and philosophy: from time travel to superintelligence*.
Bostrom, N. (2005). The simulation argument: Reply to Weatherson. *The Philosophical Quarterly*, 55(218), 90-97.
Bostrom N. (2008) Where are they? Why I hope the search for extraterrestrial life finds nothing. *MIT Technology Review* May/June: 72-77.
Bostrom, N. (2011) The Simulation Argument FAQ. URL: http://www.simulation-argument.com/faq.html
Bostrom, N. (2014). *Superintelligence: paths, dangers, strategies*. OUP.
Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, *10*(4), 455-476.
Bostrom, N., & Ćirković, M. M. (Eds.). (2008). *Global catastrophic risks*. OUP.
Bourget, D. & Chalmers, D.., (2021) Philosophers on Philosophy: The 2020 PhilPapers Survey. URL: https://philpapers.org/go.pl?id=BOUPOP-3&u=https%3A%2F%2Fphilpapers.org%2Farchive%2FBOUPOP-3.pdf
Braddon-Mitchell, D., & Latham, A. J. (2022). Ancestor simulations and the Dangers of Simulation Probes. *Erkenntnis*, 1-11.
Buchak, L. M. (2013). *Risk and rationality*. OUP.
Bykvist, K. (2017). Moral uncertainty. *Philosophy Compass*, 12(3), e12408.
Carroll, S.M. (2020). Why Boltzmann brains are bad. In *Current Controversies in Philosophy of Science* (pp. 7-20). Routledge.
Carter B. (1983) The anthropic principle and its implications for biological evolution. *Philos Trans R Soc Lond A Math Phys Sci* 310:347–363.
Chalmers, D. (2003). The Matrix as metaphysics. *Science Fiction and Philosophy: From Time Travel to Superintelligence*.
Chalmers, D. (2006). Perception and the Fall from Eden. *Perceptual experience*, 49-125.
Chalmers, D. (2010). The Singularity. *Journal of Consciousness Studies*, 17(9-10), 7-65.
Chalmers, D. (2018*a*). The meta-problem of consciousness. *Journal of Consciousness Studies*, 25(9-10).
Chalmers, D. (2018*b*). Structuralism as a response to skepticism. *The Journal of Philosophy*, 115, 625–660.
Chalmers, D. (2020). Debunking Arguments for Illusionism about Consciousness. *Journal of Consciousness Studies*, 27(5-6), 258-281.
Chalmers, D. J. (2022). *Reality+: Virtual worlds and the problems of philosophy*. Penguin UK.
Chen, E.K. (forthcoming). Time's Arrow and Self-Locating Probability. *Philosophy and Phenomenological Research*.

Christensen, D. (2016). Conciliation, Uniqueness and Rational Toxicity. *Noûs*, 50(3), 584-603.
Ćirković, M.M. (2008). Observation selection effects and global catastrophic risks. *Global Catastrophic Risks*. OUP. pp. 120-145.
Ćirković, M.M. (2015). Linking simulation argument to the AI risk. *Futures*, 72, 27-31.
Ćirković, M.M. (2018). *The great silence: Science and philosophy of Fermi's paradox*. OUP.
Clarke-Doane, J. (2020). *Morality and Mathematics*. OUP.
Christiano, P. (2019) "What failure looks like" *Alignment forum*.
Cotra, A. (2020) Forecasting Transformative AI with Biological Anchors. Open Philanthropy.
Cotra, A. (2022) Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. *Alignment forum*.
Crawford, L. (2013). Freak observers and the simulation argument. *Ratio*, 26(3), 250-264.
Critch, A. (2021)What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes. *Alignment forum.*
Crummett, D. (2017). Sufferer-centered requirements on theodicy and all-things-considered harms. *Oxford Studies in Philosophy of Religion*, 8.
Crummett, D. (2020). The real advantages of the simulation solution to the problem of natural evil. *Religious Studies*, 1-16.
Cuneo, Terence (2007). *The Normative Web: An Argument for Moral Realism*. OUP.
Cutter, B., & Crummett, D. (forthcoming) Psychophysical Harmony: A New Argument for Theism. *Oxford Studies in Philosophy of Religion*.
Cutter, B. (2021). Perceptual illusionism. *Analytic Philosophy*.
Cutter & Saad (forthcoming) The Problem of Nomological Harmony. Manuscript. *Nous*.
Dainton, B. (2012). On singularities and simulations. *Journal of Consciousness Studies*, 19(1-2), 42-85.
Dainton, B. (2020). Natural evil: the simulation solution. *Religious Studies*, 56(2), 209-230.
Danaher, J. (2015). Why AI doomsayers are like sceptical theists and why it matters. *Minds and Machines*, 25(3), 231-246.
Daniel, M. (2017). S-risks: Why they are the worst existential risks, and how to prevent them (EAG Boston 2017). Foundational research institute.
Davidson, T. (2023) Continuous doesn't mean slow. URL: https://www.planned-obsolescence.org/continuous-doesnt-mean-slow/
Dello-Iacovo, M.A. (2017) On terraforming, wild-animal suffering and the far future. *Sentience Politics*.
Dreier, J. (2012). Quasi-realism and the problem of unexplained coincidence. *Analytic Philosophy*, 53(3), 269-287.
Dogramaci, S. (2017). Explaining our moral reliability. Pacific Philosophical Quarterly, 98, 71-86.
Dogramaci, S. (2020). Does my total evidence support that I'm a Boltzmann Brain? *Philosophical Studies* 177 (12):3717-3723.
Doody, R. (2022). Don't Go Chasing Waterfalls: Against Hayward's "Utility Cascades". *Utilitas* 34 (2):225-232.
Dorr, C., & Arntzenius, F. (2017). Self-locating priors and cosmological measures. *The philosophy of cosmology*, 396-428.
Easwaran, K. (2021). A classification of Newcomb problems and decision theories. *Synthese*, 198(27), 6415-6434.

Elamrani, A., & Yampolskiy, R. V. (2019). Reviewing tests for machine consciousness. *Journal of Consciousness Studies*, 26(5-6), 35-64.
Elga, A. (2000). Self-locating belief and the sleeping beauty problem. *Analysis* 60 (2):143–147.
Elga, A. (2004). Defeating dr. evil with self-locating belief. *Philosophy and Phenomenological Research* 69 (2):383–396.
Elga, A. (2008). "Lucky to be rational." Unpublished manuscript. URL: https://www.princeton.edu/~adame/papers/bellingham-lucky.pdf.
Enoch, D. (2011). *Taking Morality Seriously: A Defense of Robust Realism*. OUP.
Erez, F. (2023). Ought we align the values of artificial moral agents?. *AI and Ethics*, 1-10.
Everitt, N. (2004). *The non-existence of God*. Routledge.
Fitzpatrick, L. (2009). "A Brief History of China's One-Child Policy". *Time*. ISSN 0040-781X
Friederich, S., "Fine-Tuning", *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), E.N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2018/entries/fine-tuning/>.
Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*, 30(3)411-437.
Gloor, L. (2016) "The Case for Suffering-Focused Ethics" URL: https://longtermrisk.org/the-case-for-suffering-focused-ethics/
Gloor, L. (2018) "Cause prioritization for downside-focused value systems" URL: https://longtermrisk.org/cause-prioritization-downside-focused-value-systems/
Goff, P. (2018). Conscious thought and the cognitive fine-tuning problem. *Philosophical Quarterly*, 68(270), 98–122.
Grace, C. (2010). Anthropic reasoning in the great filter. Unpublished manuscript.
Greaves, H. (2016, October). Cluelessness. In *Proceedings of the Aristotelian Society* (Vol. 116, No. 3, pp. 311-339). OUP.
Greaves, H., Cotton-Barratt (2019) "A bargaining-theoretic approach to moral uncertainty" URL: https://globalprioritiesinstitute.org/wp-content/uploads/2020/Cotton-Barratt_%20Greaves_bargaining_theoretic.pdf
Green, A., & Stump, E. (Eds.). (2015). *Hidden Divinity and Religious Belief*. Cambridge.
Greene, J. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (ed.), Moral Psychology, Vol. 3. MIT Press.
Greene, P. (2020). The Termination Risks of Simulation Science. *Erkenntnis* 85(2):489-509.
Gustafsson, J.E., & Peterson, M. (2012). A computer simulation of the argument from disagreement. *Synthese*, 184(3), 387-405.
Häggström, O. (2021) AI, orthogonality and the Müller-Cannon instrumental vs general intelligence distinction. URL: https://arxiv.org/ftp/arxiv/papers/2109/2109.07911.pdf
Hanson, R., (1998). The Great Filter - Are We Almost Past It? URL: http://hanson.gmu.edu/greatfilter.html
Hanson, R. (2001). How to live in a simulation. *Journal of Evolution and Technology*, 7(1), 3-13.
Hanson, R. (2016). *The age of Em: Work, love, and life when robots rule the Earth*. OUP.
Hanson, R., Martin, D., McCarter, C., & Paulson, J. (2021). If Loud Aliens Explain Human Earliness, Quiet Aliens Are Also Rare. The Astrophysical Journal, 922(2),182.
Hayward, M.K. (2020). Utility cascades. *Analysis* 80 (3):433-442.
Hill, R. R., & Tolk, A. (2017). A history of military computer simulation. In *Advances in Modeling and Simulation* (pp. 277-299). Springer.

Hoffman, D. (2019). *The case against reality: Why evolution hid the truth from our eyes.* WW Norton & Company.
Huang, S., ... & Wei, F. (2023). Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045.
Isaacs, Y.; Hawthorne, J. & Russell, J.S. (forthcoming). Multiple Universes and Self-Locating Evidence. *Philosophical Review*.
Ichikawa, J. (2011). Quantifiers, Knowledge, and Counterfactuals. *Philosophy and Phenomenological Research* 82 (2):287 - 313.
James, W. (1890). *The principles of psychology* (Vol. 2). New York: Henry Holt and Company
Jaquet, F. (2022). Speciesism and tribalism: embarrassing origins. *Philosophical Studies*, 179(3), 933-954.
Johnson, D.K. (2011). Natural evil and the simulation hypothesis. *Philo*, 14(2), 161-175.
Joyce, R. (2007). *The evolution of morality*. MIT press.,
Karnofsky, H. (2021). "All Possible Views About Humanity's Future Are Wild" URL: https://www.cold-takes.com/all-possible-views-about-humanitys-future-are-wild/
Kagan, S. (2000). Evaluative focal points. Morality, rules, and consequences: A critical reader, 134-55.
Knab, B. (2019) *Three Problems in Formal Epistemology*. Dissertation.
Arden Koehler, Benjamin Todd, Robert Wiblin and Keiran Harris (2020) "BenjaminTodd on varieties of longtermism and things 80,000 Hours might be getting wrong" Podcast. URL: https://80000hours.org/podcast/episodes/ben-todd-on-varieties-of-longtermism/#articles-books-and-other-media-discussed-in-the-show
Korman, D.Z. (2019). Debunking arguments. *Philosophy Compass* 14 (12).
Kotzen, M. (2020). What Follows from the Possibility of Boltzmann Brains?. In *Current Controversies in Philosophy of Science* (pp. 21-34). Routledge.
Kraay, K. (2010) "Theism, Possible Worlds, and the Multiverse." *Philosophical Studies* 147:355-368.
Kraay, K. (ed.) (2014). *God and the Multiverse: Scientific, Philosophical, and Theological Perspectives*. Routledge.
Ladak, A.(2022) Is Artificial Consciousness Possible? A Summary of Selected Books. URL: https://www.sentienceinstitute.org/blog/is-artificial-consciousness-possible.
de Lazari-Radek, K., & Singer, P. (2012). The objectivity of ethics and the unity of practical reason. *Ethics*. 123(1)9-31.
Lawsen, A. (2023) AI x-risk, approximately ordered by embarrassment. *Alignment forum*.
Lee, G. (2019). 13 Alien Subjectivity and the Importance of Consciousness. *Blockheads!: Essays on Ned Block's Philosophy of Mind and Consciousness*, 215.
Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., ... & Legg, S. (2017). AI safety gridworlds. URL: https://arxiv.org/pdf/1711.09883.pdf
Lenman, J. (2002). On becoming extinct. *Pacific Philosophical Quarterly* 83 (3):253–269.
Leslie, J. (1989) *Universes*. Routledge. [reprinted in 1996]
Leslie, J. (1991). Ensuring two Bird deaths with one throw. *Mind* 100 (1):73-86.
Leslie, J. (1996), *The End of the World: The Science and Ethics of Human Extinction*. Routledge.
Lewis, D. (2001). Sleeping beauty: reply to Elga. *Analysis*, 61(3), 171-176.
Lewis, D. (2004). How many lives has Schrödinger's cat?. *Australasian Journal of Philosophy*, 82(1), 3-22.
Lewis, P.J. (2013). The doomsday argument and the simulation argument. *Synthese*, 190(18), 4009-4022.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2022). Emergent world representations: Exploring a sequence model trained on a synthetic task. arXiv preprint arXiv:2210.13382.
Lockhart, T. (2000). Moral uncertainty and its consequences. OUP.
Long, R. Key questions about artificial sentience: an opinionated guide. Effective Altruism Forum. URL: https://forum.effectivealtruism.org/posts/gFoWdiGYtXrhmBusH/key-questions-about-artificial-sentience-an-opinionated
Joyce, R. (2001). *The Myth of Morality.* Cambridge University Press.
MacAskill, W. (2016). Smokers, psychos, and decision-theoretic uncertainty. *The Journal of Philosophy*, 113(9), 425-445.
MacAskill, W. (2022) *What We Owe the Future*. Basic Books.
MacAskill, W., Bykvist, K., & Ord, T. (2020). *Moral Uncertainty.* OUP.
MacAskill, W.; Vallinder, A.; Oesterheld, C.; Shulman, C. & Treutlein, J. (2021). The Evidentialist's Wager. *Journal of Philosophy* 118 (6):320-342.
Mackie, J.L. (1977). *Ethics: Inventing Right and Wrong*. Penguin Books.
Madden JC, Enoch SJ, Paini A, Cronin MTD. A Review of In Silico Tools as Alternatives to Animal Testing: Principles, Resources and Applications. *Alternatives to Laboratory Animals*. 2020;48(4):146-172.
Manley, D. Manuscript. "On Being a Random Sample"
Mason, K. (2010). Debunking arguments and the genealogy of religion and morality. *Philosophy Compass*, 5(9), 770-778.
Megill, J. (2011). Evil and the many universes response. *International Journal for Philosophy of Religion* 70:127–138.
Mogensen A. (2019). Doomsday rings twice. Available at https://globalprioritiesinstitute.org/wp-content/uploads/2019/Mogensen_doomsday_rings_twice.pdf
Mogensen, A. (2019). 'The only ethical argument for positive *δ*'?. Working paper.
Miller, J.D. (2019). When two existential risks are better than one. Foresight, 21(1), 130–137. https://doi.org/10.1108/FS04-2018-0038
Miller, J.D., & Felton, D. (2017). The Fermi paradox, Bayes' rule, and existential risk management. *Futures*, 86, 44-57.
Monton, B. (2009). *Seeking God in science: an atheist defends intelligent design*. Broadview Press.
Moravec, H. (1976) "The Role of Raw Rower in Intelligence." Unpublished manuscript. URL: http://www.frc.ri.cmu.edu/users/hpm/project.archive/general.articles/1975/Raw.Power.html
Moravec, H. (1999) *Robot: Mere Machine to Transcendent Mind*. OUP.
Mowshowitz, Z. (2023) A Hypothetical Takeover Scenario Twitter Poll.
Mørch, H. (2018). The evolutionary argument for phenomenal powers. *Philosophical Perspectives*, 31, 293–316.
Muehlhauser, L. (2021) "Treacherous turns in the wild" http://lukemuehlhauser.com/treacherous-turns-in-the-wild/
Muehlhauser, L. (2017) A software agent illustrating some features of an illusionist account of consciousness, OpenPhilanthropy, [Online], https://www.openphilanthropy.org/software-agent-illustrating-some-features-illusionistaccount-consciousness

Müller, V.C. & Cannon, M. (2021). Existential risk from AI and orthogonality: Can we have it both ways? *Ratio* 35(1):25-36.
Murphy, P., & Black, T. (2012). Sensitivity meets explanation: an improved counterfactual condition on knowledge. In K. Becker & T. Black (eds.), *The Sensitivity Principle in Epistemology*. Cambridge.
Newberry, T., & Ord, T. (2021). The Parliamentary Approach to Moral Uncertainty. Technical Report 2021-2, Future of Humanity Institute, University of Oxford. URL: https://www. fhi. ox. ac. uk/wpcontent/uploads/2021/06/Parliamentary-Approach-to-Moral-Uncertainty. pdf
Nowak, A., Gelfand, M. J., Borkowski, W., Cohen, D., & Hernandez, I. (2016). The evolutionary basis of honor cultures. *Psychological science*, 27(1), 12-24.
Nozick, R. (1981). *Philosophical Explanations*. Harvard University Press.
Nozick, R. (1993). The nature of rationality. Princeton University Press.
Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative & Physiological Psychology*, 47, 419–427.
Owen Cotton-Barratt (2021) Web of Virtue thesis. URL: https://forum.effectivealtruism.org/posts/yejoA6bmQSjRqEGK3/web-of-virtue-thesis-research-note
Parfit, D. (1984). *Reasons and persons*. OUP.
Parfit, D. (2011) *On what matters: Vol. 2*. OUP.
Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. arXiv preprint arXiv:2304.03442.
Passini, E., Britton, O. J., Lu, H. R., Rohrbacher, J., Hermans, A. N., Gallacher, D. J., ... & Rodriguez, B. (2017). Human in silico drug trials demonstrate higher accuracy than animal models in predicting clinical pro-arrhythmic cardiotoxicity. *Frontiers in physiology*, 668.
Pautz, A. (2014). The real trouble with phenomenal externalism: New empirical evidence for a brain-based theory of consciousness. In *Consciousness inside and out: Phenomenology, neuroscience, and the nature of experience* (pp. 237-298). Springer, Dordrecht.
Pautz, A. (2020). Consciousness and coincidence: Comments on Chalmers. *Journal of Consciousness Studies*, (5-6).
Pautz, A. (2021). *Perception*. Routledge.
Peterson, M. (2019). The value alignment problem: a geometric approach. *Ethics and Information Technology*, 21, 19-28.
Perez, E. (2022) A Test for Language Model Consciousness. URL: https://www.greaterwrong.com/posts/9hxH2pxffxeeXk8YT/a-test-for-language-model-consciousness
Pettigrew, R. (2022) Effective altruism, risk, and human extinction. URL: https://globalprioritiesinstitute.org/wp-content/uploads/Pettigrew-Effective-altruism-risk-and-human-extinction-2.pdf
Prinz, J. (2012). Singularity and inevitable doom. *Journal of Consciousness Studies*, *19*(7-8), 77-86.
Ord, T. (2020). *The precipice: existential risk and the future of humanity*. Hachette Books.
Olson, J. (2014). *Moral Error Theory: History, Critique, Defence*. OUP.
Rowland, R. (2019). Local evolutionary debunking arguments. Philosophical Perspectives, 33(1), 170-199.
Richmond, A.M. (2017). Why doomsday arguments are better than simulation arguments. *Ratio*, 30(3), 221-238.

Riedener, S. (2021) Existential risks from a Thomist Christian perspective. URL: https://globalprioritiesinstitute.org/wp-content/uploads/Stefan-Riedener_Existential-risks-from-a-Thomist-Christian-perspective.pdf
Saad, B. (2019). A teleological strategy for solving the meta-problem of consciousness. *Journal of Consciousness Studies*, 26(9-10), 205-216.
Saad, B. (2020). Two solutions to the neural discernment problem. *Philosophical Studies*, 177(10), 2837-2850.
Saad, B. (forthcoming*a*) Harmony in a Panpsychist World. *Synthese*.
Saad, B. (forthcoming*b*) "The sooner the better: an argument for bias toward the earlier". *Journal of the American Philosophical Association*
Saad, B. (forthcoming*c*) "Lessons from the void: what Boltzmann brains teach" *Analytic Philosophy.*
Saad, B. & Bradley, A. (2022). The Problem of Digital Suffering. *Inquiry*.
Sampson, E. (2022) Disaster Prevention and the Possibility of Hell: A Dilemma for Longtermist Effective Altruists. Manuscript.
Sandbrink, J., Hobbs, H., Swett, J., Dafoe, A., & Sandberg, A. (2022). Differential technology development: A responsible innovation principle for navigating technology risks. Available at SSRN.
Sandberg, A., Armstrong, S., & Cirkovic, M.M. (2017). That is not dead which can eternal lie: the aestivation hypothesis for resolving Fermi's paradox. arXiv preprint arXiv:1705.03394.
Sandberg, A., Drexler, E., & Ord, T. (2018). Dissolving the Fermi paradox. arXiv preprint arXiv:1806.02404.
Schellenberg, J. L. (1996). Divine Hiddenness and Human Reason. *International Journal for Philosophy of Religion* 40(2):121-124.
Schellenberg, J. L. (2010). The Hiddenness Problem and the Problem of Evil. *Faith and Philosophy* 27(1):45-60.
Schelling, T.C., (1971). "Dynamic Models of Segregation," *Journal of Mathematical Sociology*, 1: 143–186.
Schwartz, S. H., & Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. *Journal of research in personality*, 38(3), 230-255.
Schwitzgebel, E. (2017). 1% Skepticism. *Noûs*, 51(2), 271-290.
Schwartz, S. H., & Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. Journal of research in personality, 38(3), 230-255.
Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, 1-14.
Shafer-Landau, R. (2012). Evolutionary debunking, moral realism and moral knowledge. *J. Ethics & Soc. Phil.*, 7, i.
Shiller, D. (2017). In Defense of Artificial Replacement. *Bioethics*, 31(5)393-399.
Shulman, C. (2010). Whole brain emulation and the evolution of superorganisms. Mach. Intell. Res. Inst. Work. Pap. Httpintelligence OrgfilesWBESuperorgs Pdf.
Shulman, C., & Bostrom, N. (2012). How hard is artificial intelligence? Evolutionary arguments and selection effects. *Journal of Consciousness Studies*, 19(7-8), 103-130.
Shulman, C., & Bostrom, N. (2021). Sharing the World with Digital Minds. In *Rethinking Moral Status.*
Sider, T. (2020). *The tools of metaphysics and the metaphysics of science.* OUP.
Silva Jr, P. (forthcoming). Debunking Objective Consequentialism: The Challenge of Knowledge-Centric Anti-Luck Epistemology. In M. Klenk (ed.), *Higher Order Evidence and Moral Epistemology*. Routledge.
Singer, P. (2005). 'Ethics and Intuitions'. *The Journal of Ethics* 9: 331–52.

Snyder-Beattie, A. E., Sandberg, A., Drexler, K. E., & Bonsall, M. B. (2021). The timing of evolutionary transitions suggests intelligent life is rare. *Astrobiology*, 21(3), 265-278.
Sotala, K., & Gloor, L. (2017). Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica*, 41(4).
Steinhart, E. (2010). Theological implications of the simulation argument. *Ars Disputandi*, 10(1), 23-37.
Street, S. (2006). A Darwinian dilemma for realist theories of value. *Philosophical studies*, 127(1), 109-166.
Street, S. (2010). What is constructivism in ethics and metaethics? *Philosophy Compass* 5 (5):363-384.
Street, S. (2011). Mind-Independence Without the Mystery: Why Quasi-Realists Can't Have it Both Ways. In Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics*, Volume 6. OUP.
Streumer, B. (2017). *Unbelievable Errors: An Error Theory About All Normative Judgments*. OUP.
Stump, E. (1985). The problem of evil. *Faith and philosophy*, 2(4), 392-423.
Summers, M., & Arvan, M. (2021). Two New Doubts about Simulation Arguments. *Australasian Journal of Philosophy*, 1-13.
Tersman, Folke (2017). Debunking and Disagreement. *Noûs* 51 (4):754-774.
Thomas, T. (2022) Simulation expectation. URL: https://globalprioritiesinstitute.org/wp-content/uploads/Teruji-Thomas-Simulation-Expectation-2.pdf
Titelbaum, M.G. (2013). Ten reasons to care about the Sleeping Beauty problem. *Philosophy Compass*, 8(11), 1003-1017.
Todd, B. (2020). The emerging school of patient longtermism. 80,000 Hours. https://80000hours.org/2020/08/the-emerging-school-of-patient-longtermism/
Tomasik, B. (2016). How the Simulation Argument Dampens Future Fanaticism. URL: https://longtermrisk.org/files/how-the-simulation-argument-dampens-future-fanaticism.pdf
Tomasik, B. (2017). What Are Suffering Subroutines. URL: https://reducing-suffering.org/what-are-suffering-subroutines/
Tufekci, Z. (2022) Open Thread. URL: https://www.theinsight.org/p/open-thread-heres-hoping-we-dont
Turchin, A. (2016) "The Map of Shelters and Refuges from Global Risks". URL: https://www.academia.edu/50829599/The_Map_of_Shelters_and_Refuges_from_Global_Risks_Plan_B_of_X_risks_Prevention_
Turchin, A. (2018). A Meta-Doomsday Argument: Uncertainty About the Validity of the Probabilistic Prediction of the End of the World. URL: https://philarchive.org/rec/TURAMA-4
Turchin, A., Batin, M., Denkenberger, D., & Yampolskiy, R. (2019). Simulation Typology and Termination Risks. arXiv preprint arXiv:1905.05792.
Udell, D. B. & Schwitzgebel, E. (2021). Susan Schneider's Proposed Tests for AI Consciousness: Promising but Flawed. *Journal of Consciousness Studies* 28 (5-6):121-144.
Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3), 189-191.
Van Fraassen, Bas C. (1989). *Laws and Symmetry*. OUP.
Vavova, K. (2014). Debunking evolutionary debunking. *Oxford studies in metaethics*, 9(7), 6-101.
Vavova, K. (2015). Evolutionary debunking of moral realism. *Philosophy Compass*, *10*(2), 104-116.
Vemprala, S., Bonatti, R., Bucker, A., & Kapoor, A. (2023). Chatgpt for robotics: Design principles and model abilities. Microsoft.

Virués-Ortega, J., Buela-Casal, G., Garrido, E., & Alcázar, B. (2004). Neuropsychological functioning associated with high-altitude exposure. *Neuropsychology review*, 14(4), 197-224.
Ward P, & Brownlee D. (1999) *Rare earth*. In Copernicus, Springer-Verlag, New York, NY.
Webb, S. (2015). *If the universe is teeming with aliens... where is everybody?: fifty solutions to the Fermi paradox and the problem of extraterrestrial life*. 2nd ed. New York, NY: Copernicus Books.
White, R. (2006). Problems for dogmatism. *Philosophical Studies*, 131, 525–557.
White, R. (2010). You just believe that because.... *Philosophical Perspectives* 24 (1):573-615.
White, R. (2018). Reasoning with Plenitude. In *Knowledge, Belief, and God: New Insights in Religious Epistemology*
Wright, C. (2004). Warrant for nothing. *Proceedings of the Aristotelian Society*, Supplementary Volume, 78, 167–212.
Xia, L., Robock, A., Scherrer, K., Harrison, C. S., Bodirsky, B. L., Weindl, I., ... & Heneghan, R. (2022). Global food insecurity and famine from reduced crop, marine fishery and livestock production due to climate disruption from nuclear war soot injection. Nature Food, 3(8), 586-596.
Yampolskiy, R. V. (2014). Utility function security in artificially intelligent agents. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3), 373-389.
Yampolskiy, R.V. (2018). Why we do not evolve software? Analysis of evolutionary algorithms. *Evolutionary Bioinformatics*, 14, 1176934318815906.
Yudkowsky, E. (2004). Coherent extrapolated volition. Singularity Institute for Artificial Intelligence.
Zackrisson, E., Calissendorff, P., González, J., Benson, A., Johansen, A., & Janson, M. (2016). Terrestrial planets across space and time. *The Astrophysical Journal*, 833(2), 214.
Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.
Zhang, L. What Rethink Priorities General Longtermism Team Did in 2022, and Updates in Light of the Current Situation. URL: https://forum.effectivealtruism.org/posts/C26RHHYXzT6P6A4ht/what-rethink-priorities-general-longtermism-team-did-in-2022#Shelters_and_Other_Civilizational_Resilience_work
Zuber, S., Venkatesh, N., Tännsjö, T., Tarsney, C., Stefánsson, H. O., Steele, K., ... & Asheim, G. B. (2021). What should we agree on about the repugnant conclusion?. *Utilitas*, 33(4), 379-383.

Funding: This project has been supported by the Long-Term Future Fund, the Sentience Institute, and Utrecht University. The views expressed in this report do not necessarily reflect those of funders.