The Termination Risks of Simulation Science

Preston Greene

Forthcoming in *Erkenntnis* —
https://link.springer.com/article/10.1007%2Fs10670-018-0037-1

Abstract

Historically, the hypothesis that our world is a computer simulation has struck many as just another improbable-but-possible "skeptical hypothesis" about the nature of reality. Recently, however, the simulation hypothesis has received significant attention from philosophers, physicists, and the popular press. This is due to the discovery of an epistemic dependency: If we believe that our civilization will one day run many simulations concerning its ancestry, then we should believe that we are probably in an ancestor simulation right now. This essay examines a troubling but underexplored feature of the ancestor-simulation hypothesis: the termination risk posed by both ancestor-simulation technology and experimental probes into whether our world is an ancestor simulation. This essay evaluates the termination risk by using extrapolations from current computing practices and simulation technology. The conclusions, while provisional, have great implications for debates concerning the fundamental nature of reality and the safety of contemporary physics.

1 Introduction

In 2012, physicists from the University of Washington, led by Martin Savage, posted a preprint of a paper titled "Constraints on the Universe as a Numerical Simulation."[1] In

[1] Beane et al., 2014.

the paper, they discuss theoretical limits on computer simulations, and most interestingly, they describe possible experiments that could be used to test whether our universe is a simulation. This paper created a media storm: by the end of the year, most of the major news organizations in the United States and internationally had covered their work. In general, people seem to be fascinated with the hypothesis that we might live in a computer simulation, and specifically with the idea that this hypothesis can be scientifically tested.

An understandable reaction is to regard this interest as just more popular hoopla. Why, after all, should we think that the hypothesis that we are living in a computer simulation is probable enough to test? The answer lies in an interesting dependency amongst our beliefs originally pointed out by the philosopher Nick Bostrom [2003b]. The physicists, as well as the media reports, credit Bostrom as the originator of "the simulation hypothesis," which proposes that we may be living in a computer simulation set up at an advanced stage of our civilization. For example, a University of Washington press release states, "A decade ago, a British philosopher put forth the notion that the universe that we live in might in fact be a computer simulation run by our descendants. While that seems far-fetched, perhaps even incomprehensible, a team of physicists at the Universe of Washington has come up with a potential test to see if the idea holds water" [Stricherz, 2012]. Even if our work is called "incomprehensible" in comparison to that of physicists, philosophers can still rejoice in the attention, since it is not often that articles in philosophy journals receive international news coverage.

What is the argument for the simulation hypothesis, and why has it had this uncommon influence? These questions are answered in Section 2. The principal part of the paper is an analysis of a troubling but relatively underexplored feature of the simulation hypothesis: the possibility that either simulation technologies or experimental probes would cause the simulators to terminate our universe.[2] In Section 3, I show that the

[2] Simulation shutdown is identified as a significant existential risk and discussed briefly by Bostrom [2002, Sections 4.3 and 8.4], Ćirković [2008, Section 6.5], and Torres [2017, Chapter 5.1]. Bostrom [2002, Section 4] identifies it as the third most probable cause of the extinction of human life. Torres [2017, Chapter 5] writes that it is "a threat that, untrained intuitions aside, could be more worrisome than it initially appears."

termination risk of simulation technologies must be taken into account when deciding whether to create them, and that doing so reveals a reason to doubt that their expected value is positive. This conclusion is relevant to current theorizing about the simulation hypothesis because it showcases a previously unnoticed reason to believe that our world is unsimulated. Section 4 analyses a corresponding termination risk created by experimental probes, such as those proposed by Beane et al. [2014], and develops an approach to calculating the expected value of such probes in light of the risk. This calculation has direct implications for the safety of near-future physics.

2 The Simulation Argument

One way to approach the simulation argument is through thinking about the sorts of future technologies that our civilization is likely to develop. If we currently had the technology to create complete simulations of the history of humanity and its countless potential variations (what Bostrom calls "ancestor" simulations) would we be interested in doing so?

In the empirical sciences, experiments and computer simulations have a shared functional description, and differ mostly in the types of questions they are used to answer [El Skaf and Imbert, 2013]. Simulations that respect the laws of nature and allow for initial conditions to be varied would represent the holy grail for many forms of research. Such simulations would allow for better answers to counterfactual questions about the past (i.e., "What would have happened if things were different?") and the assignment of probabilities to counterfactual conditionals generally. Relatedly, simulations can be used to efficiently distinguish causes from mere correlations in forming social-scientific theories. One day, researchers may use advanced simulations to pursue knowledge in many domains of inquiry. Already, inchoate research simulations are being used for these purposes.[3]

[3] For an overview of the scientific status of agent-based simulations, see Grüne-Yanoff and Weirich, 2010 and Grüne-Yanoff, 2011. See Grüne-Yanoff, 2009 for a discussion of the research potential of simulated societies.

If we accept—along with the majority of those working in the philosophy of mind—that mental states are *substrate-independent*, then we should think that agents in advanced simulations are conscious. The substrate-independence thesis holds that mental states can be realized by different types of physical states than those present in the human brain, provided that a system implements similar computational structures and processes. If the behavior of simulated agents is to match what one would expect of non-simulated agents, then the computational processes governing simulated agents would need to nearly match those carried out by the human mind. Given that agents in advanced simulations are supposed to display the expected behavior of non-simulated agents, we should expect such agents to be conscious.

Furthermore, if posthumans decide to create ancestor simulations, then they would be able to create a very large number of them. Assuming only computational mechanisms that we currently understand, a conservative estimate is that a computer could simulate the entire history of humankind a million times over every second.[4]

With these assumptions, we can state the basic idea of the simulation argument in one sentence:

> If we believe that there is a substantial chance that our civilization will one day run many ancestor simulations, then we should believe that we are probably in an ancestor simulation right now.

Bostrom introduces two potential reasons to deny the antecedent. The first of which involves the thought that "posthumans" (i.e., the inhabitants of civilizations utilizing advanced technologies like ancestor simulations) might tend to have desires radically different from our own. One possibility is that posthumans would not derive much research or entertainment value from ancestor simulations. Perhaps they would no longer be interested in answering the kinds of questions that ancestor simulations target, or the sort

[4] For technical questions outside the scope of this review, concerning, for example, what sort of computers ancestor simulations would require and what parts of the universe would need to be simulated, see Bostrom, 2003b, Section 3.

of entertainment that they can provide. It is also possible that they would find ancestor simulations immoral.

Bostrom also considers the possibility that posthuman civilizations tend to never be in a position to create ancestor simulations, because they go extinct before developing the required technology. This might be the case if highly destructive weapons tend to become ubiquitous before civilizations are in a position to simulate.

That concludes the full argument. There are three options: (1) almost all posthuman civilizations are not interested in creating ancestor simulations, (2) almost all civilizations go extinct before reaching a posthuman stage, or (3) we are probably living in an ancestor simulation. And there are two seemingly humdrum assumptions: the substrate-independence thesis and a "bland indifference principle" licensing the following inference: "if there are many more simulated realities than unsimulated ones (and the two are indistinguishable), then we are probably in a simulated reality."[5] Bostrom recommends that we resist further speculation and split our confidence between (1), (2), and (3), and thus he suggests that the odds are about 1:2 that we are living in an ancestor simulation. Presumably, before engaging with the argument, most people take the probability that we are living in a computer simulation to be much less.

It is worth noting that the ancestor-simulation hypothesis is not technically a skeptical hypothesis, but rather a metaphysical hypothesis about the nature of reality. A skeptical hypothesis is a hypothesis such that if it were true, then nearly all our beliefs about the external world would be false. As Chalmers [2010] argues, common beliefs about the external world, such as "I have hands," would not be falsified by the discovery that the universe was intelligently created and that bits represent the fundamental unit of reality. Instead, if we discover that the world is made of bits and was created by beings living outside of space and time, we should regard this as revealing interesting metaphysical truths about our world, rather than revealing that our world is not "real."[6]

[5] For criticism of Bostrom's use of an indifference principle see Weatherson, 2003. For Bostrom's reply: Bostrom, 2005.

[6] Note that an argument to a similar conclusion can be found in Putnam, 1981. Further note that failure to grasp this point can lead to errors in judging the epistemic force of the ancestor-simulation argu-

It should now be clear that the ancestor-simulation hypothesis is more compelling than run-of-the-mill “skeptical” hypotheses, including the hypothesis that we live in a computer simulation that is not an ancestor simulation. Arguments concerning traditional skeptical hypotheses (or the hypothesis that we live in a non-ancestor computer simulation) first note that the hypothesis might be true, and they then challenge us to explain how we could know it is false. Such arguments do not give us a reason to believe that the hypothesis is true. The ancestor-simulation argument, by contrast, utilizes everyday beliefs about how our civilization is likely to develop to reveal empirical evidence for a metaphysical hypothesis. Since the argument appeals to empirical evidence in favor of the hypothesis (rather than reasons to think we cannot know the hypothesis is false), the appropriate way to engage with it is as we would with a scientific claim about the nature of reality, and not as we would with an epistemological claim about the limits of our knowledge.[7]

3 The Termination Risk of Simulation Technologies

When first entertaining the ancestor-simulation hypothesis, many wonder whether there might be multiple levels of simulation. If it is possible for those on the basement-level to simulate their ancestors, would it not be possible for those in an ancestor simulation to do the same? And if so, might this lead to the unbounded development of simulations within simulations?

ment. For example, Birch [2013, 98] claims that for the argument to succeed, we must have just as much evidence for its assumptions as we do for the proposition “I possess two real human hands.” He rejects the ancestor-simulation argument because he thinks this “limb scepticism” is unjustified [101]. However, pace Birch, everyday objects exist, including “real, human hands,” even if the ancestor-simulation hypothesis is true.

[7] As Torres [2017, Section 5.1] writes, “While this scenario may still sound fantastical, no philosopher has discovered, to the satisfaction of most other philosophers, a broken gear in the argument’s logical machinery.” He adds, “Most of contemporary science consists of patently ’fantastical’ claims about the nature and the workings of reality.... What matters isn’t how crazy an idea sounds to epistemically naive ears, but the extent to which that idea is positively supported by the totality of available and intersubjectively verifiable evidence” [Section 5.1, fn. 13].

The problem with this reasoning is that any computation performed in a simulation must ultimately be supported by computation on the basement level. The computing constraints and the objectives of the simulators on the basement level, therefore, put a limit on the amount of nesting that can occur.[8] It is partly for this reason that ancestor simulations entail a *termination risk* to those that create them. A similar risk is mentioned by Bostrom [2003b, 253], who writes: "One consideration that counts against the multi-level hypothesis is that the computational cost for the basement-level simulators would be very great. Simulating even a single posthuman civilization might be prohibitively expensive. If so, then we should expect our simulation to be terminated when we are about to become posthuman." John Tierney, writing in *The New York Times*, gives the idea a flashier description:

> It's also possible that there would be logistical problems in creating layer upon layer of simulations. There might not be enough computing power to continue the simulation if billions of inhabitants of a virtual world started creating their own virtual worlds with billions of inhabitants apiece. If that's true, it's bad news for the futurists who think we'll have a computer this century with the power to simulate all the inhabitants on earth. We'd start our simulation, expecting to observe a new virtual world, but instead our own world might end — not with a bang, not with a whimper, but with a message on the Prime Designer's computer. It might be something clunky like "Insufficient Memory to Continue Simulation." But I like to think it would be simple and familiar: "Game Over."[9]

At first glance, this kind of worry might seem like wild speculation. However, on full reflection, I believe the idea is worthy of philosophical analysis. If we seriously entertain the possibility that we are living in an ancestor simulation, then we should seriously entertain the things we know are entailed by it. In what follows, I argue that the termination risk is one of those things. Given the connection between the simulation hypothesis and

[8] Cf. Bostrom, 2009, 459.

[9] Tierney, 2007.

the termination risk, and given the catastrophic consequences of termination, we should think hard about how the risk affects both the practical and epistemological upshots of simulation technologies.

3.1 Counterfactual Historical Simulations

Imagine attempting to use your computer to simulate yourself working on your computer. In this way, the problem with nested levels of simulation can be appreciated intuitively. The first problem is that the simulation would create an infinite loop. Call this *the hanging problem*. Even if shortcuts were used to eliminate insignificant aspects of your actual computer's computation in the simulation, if your simulated computer itself included a simulation, and so on, then the program would never halt. Similarly, a mega computer using its resources to run billions of ancestor simulations would face the hanging problem if it allowed those simulations to run their own ancestor-simulation mega computers. Even if an individual ancestor simulation running a similar computer to that on the basement level did not create an infinite loop, it would still put a large demand on the resources of the basement-level computer, which would make it impossible to run many such programs concurrently, which is a probable objective of the basement-level simulators.

The most straightforward solution to the hanging problem, which is ubiquitous in modern computing, is to terminate a program that creates an infinite loop or triggers resource exhaustion. Ironically, an attempt to avoid the problem by eliminating the possibility of nesting only serves to heighten the termination risk. If we were to program our ancestor simulations to terminate before the simulated inhabitants switch on their own ancestor simulations, then that would provide us with evidence that our own simulation would be similarly programmed if we inhabit one. If we were to program our ancestor simulations to terminate after a specific number of nestings, then we would have to consider the possibility that our universe is a final allowable nesting. Such strategies would, therefore, serve to heighten the perceived termination risk, rather than diminish it.

Furthermore, the probability that one inhabits an ancestor simulation is greatest when one is in a position to create one's own. As [Bostrom, 2006, 39] writes, "[An] event that would let us conclude with a high degree of confidence that we are in a simulation is if we ever reach a point when we are about to switch on our own ancestor simulations. That would be very strong evidence against the first two propositions, leaving us only with the third." For this reason, the creators of ancestor simulations should conclude that they already inhabit an ancestor simulation, and therefore that they are actually creating *nested* ancestor simulations. Thus, if nested ancestor simulations create a termination risk, then this risk must always be accounted for in deciding whether to develop ancestor simulations.

Of course, solutions to the hanging problem more sophisticated than termination are possible. A program could be designed to self-modify rather than terminate. This solution is rarely pursued in modern computing because programmers consider it far simpler and safer for a system to terminate a program and restart it than to attempt modification. However, there may be little reason to assume that advanced programmers would retain this practice.

The issue with modification, instead, is a *relevance problem* for ancestor simulations used for research purposes: modifications to an ancestor simulation risk making it irrelevant to the question being studied. As discussed in Section 1, simulations that respect the laws of nature and allow for initial conditions to be varied would represent the holy grail for forms of research that deal with counterfactuals. While many such simulations might study counterfactuals about the future, we can be certain that our simulation, if we inhabit one, concerns the past (since advanced simulation technology has not been invented yet). Call the type of simulation we could inhabit a *counterfactual historical simulation.*

A system that is studying counterfactual historical simulations cannot produce useful data if its simulations involve large divergences from physical laws. While it is hard to know what counts as a large divergence, we can safely assume that likely candidates include changing the simulated laws of physics so simulation computers do not function, or adjusting the psychology of simulated human agents such that they no longer desire to build simulation computers. This assumption is particularly strong in the case of counterfactual historical simulations concerning the beliefs and behaviors of human beings.

Furthermore, the study of counterfactual conditionals requires a great number of simulations; simulators cannot study such questions by running a single simulation (as opposed to the use of simulations for entertainment purposes, which do not require more than a single simulation). For counterfactual questions, simulators must run a vast number of simulations that vary the initial conditions at the time of the targeted divergence from actuality. For example, if one wants to study what caused the outbreak of World War I, then one must run a very large number of simulations and look for trends; whereas, if one wants merely to be entertained by experiencing a simulation of World War I, then a single simulation, or a small number, would suffice. Given that current interest in simulations seems to be split between research and entertainment purposes, our starting assumption should be that, at the least, it is not more probable that we inhabit an entertainment simulation than that we inhabit a research simulation. If this is true, then we should seriously consider the relevance problem when estimating the probability that our simulators would pursue some strategy other than termination.

In sum, while it is difficult to predict the motivations or programming practices of advanced simulators, our current motivations and programming practices highlight the possibility that they might use simulations for research into counterfactual questions, and that they might solve the hanging and relevance problems through termination. While this is just one possibility amongst many, it is one that merits our attention since it represents an outcome of great disutility that is supported by current trends.

3.2 Simulation Technologies and Decision-Theoretic Instability

The termination risk of ancestor simulations is complicated by a heretofore unnoticed *instability* in the decision to create them. On the one hand, as we have just seen, the probability that one inhabits an ancestor simulation would increase if one decides to create one's own (thus raising the termination risk). On the other hand, this probability would *decrease* if one decides to refrain from creating one's own (thus lowering the termination risk). In this section, I introduce a decision-theoretic model to help study this aspect of the termination risk. Once this is accomplished we will be ready, in Section 3.3, to ask what parts of the model might apply to the actual world.

Consider the following case:

> Margaret is a computer scientist developing a novel simulation program that will give her the opportunity to simulate her past life. The program will run millions of simulations of her past, in which initial conditions and chance events are varied. She expects to succeed in creating the program a few years from now. The simulations take just a few seconds to run but are indistinguishable from the basement level for the agent inside. To prevent unbounded nesting, any simulation in which the simulated Margaret chooses to create her own simulations is terminated; otherwise, a natural lifespan is simulated. Margaret wonders whether she is currently in a simulation. She reasons that she almost certainly is unless basement-level Margaret would fail to complete the program or decide not to use it.
>
> Margaret has now completed the program, and she must decide whether to use it. Seeing the results of the simulations would be immensely rewarding, but Margaret knows that if she is already in a simulation, then choosing to run the program will cause her termination.

Let us consider the following questions, in this order: What should Margaret do once she has created the program? What should Margaret believe about her situation before she has created the program? And should Margaret have created the program?

To better specify Margaret's decision problem once she has created the program, let us assign the following values to the outcomes of the decision (letting "u($B \wedge S$)" represent the value of creating past simulations while existing in the basement level):

u($B \wedge S$) = 1

u($\neg B \wedge S$) = -1

u($B \wedge \neg S$) = 0

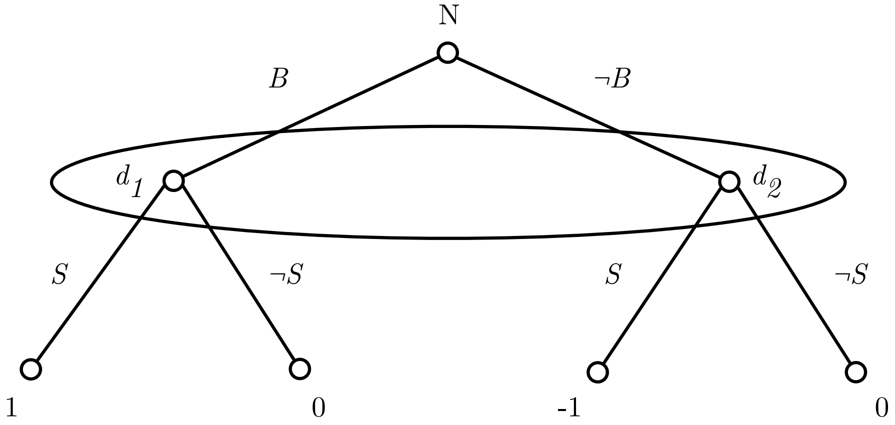

Figure 1: The Simulation Decision

$$u(\neg B \wedge \neg S) = 0$$

Thus, for the best outcome, she creates simulations and exists in the basement level; for the worst, she creates simulations and does not exist in the basement level. The outcomes for which she does not create simulations are between the two. I have purposefully made the relative value of running sims unrealistically high in order to show how generally applicable the model is. We will see later that making these values more realistic will only serve to strengthen the applicability of the model to the actual world.

We can represent the case by using a decision tree, as in Figure 1. Nature decides whether Margaret exists in the basement level or a simulation, and then she must decide whether to create simulations. Since she cannot know what nature has decided, she does not know whether she is located at d_1 or d_2. (The ellipse encircling the two decision nodes represents the fact that both d_1 and d_2 are compatible with her information set).

A striking feature of Margaret's decision situation, and the one that creates the instability, is that the result of her deliberation provides evidence bearing on the hypothesis that she is in the basement. This feature is present whenever there is an expected correlation between the decisions of basement-level Margaret and her simulations. Margaret should expect a correlation whenever her simulations base their decision on reasons also available in the basement level. Given their similar epistemic perspectives, basement-level

Margaret and her simulations would have identical reasons against running the program stemming from the possibility of being located at d_2.

Therefore, with the expected correlation of reasons, Margaret should think that d_1 is probable given $\neg S$ and that d_2 is probable given S. Since this feature of Margaret's decision situation is relevant to the disagreement between evidential and causal decision theorists, we need to consider Margaret's decision from the perspective of both evidential and causal decision theory.

According to evidential decision theory, to evaluate the expected utility of an act we take the product of the probabilities of the possibilities conditional on the act being performed and the utility of those possibilities. Using the "w_i"s to refer to elements of the space of possibilities and "A" to the act in question, the equation for expected utility is:

$$\mathrm{EU}(A) = \textstyle\sum_i \mathrm{p}(w_i \mid A)\mathrm{u}(w_i)$$

In Margaret's case, the evidential expected utility of $\neg S$ is 0, and the evidential expected utility of S is negative, since given an expected correlation, $\mathrm{p}(B \wedge S \mid S) < \mathrm{p}(\neg B \wedge S \mid S)$. Evidential decision theory would thus recommend that she not run the program.

The recommendation of causal decision theory is more complicated than that of evidential decision theory. Causal decision theory holds that the calculation of expected utilities needs to make explicit use of the agent's beliefs about causal processes.[10] Following Lewis [1981], an agent's beliefs about causal processes are represented by credences over *dependency hypotheses*. Dependency hypotheses are long conjunctions of subjunctive conditionals that specify how things depend causally on what the agent does.[11] Causal decision theory's recommendation is determined by Margaret's confidence in dependency hypotheses that include the counterfactual conditional $S \mathrel{\Box\!\!\rightarrow} T$, read as "If I were to cre-

[10] See, e.g., Egan, 2007, 95 and Peterson, 2009, 190.

[11] Causal decision theorists share a common idea, and differ mostly on matters of emphasis [Lewis, 1981, 5]. Some causal decision theorists, for example, emphasize chances in their formulations rather than counterfactual conditionals. These concepts are interrelated and, on some views, interdefinable, and, in any event, do not affect the analysis of Margaret's decision situation.

ate simulations, then I would be terminated." Given the values we specified, the causal expected utility of creating simulations is greater than that of not creating them when $\mathrm{p}(S \mathrel{\Box\!\!\rightarrow} T) < .5$. If $\mathrm{p}(S \mathrel{\Box\!\!\rightarrow} T) > .5$, then the causal expected utility of not creating simulations is less than that of creating them.

So what probability does Margaret assign to $S \mathrel{\Box\!\!\rightarrow} T$? Recall that Margaret believes: "I am almost certainly in a simulation unless basement-level Margaret fails to complete the program or decides not to use it." And after completing the program, Margaret can *rule out* the possibility that basement-level Margaret does not complete the program. She can eliminate this possibility by using a type of constructive dilemma: $((\phi \vee \psi) \wedge (\phi \rightarrow \chi) \wedge (\psi \rightarrow \chi)) \rightarrow \chi$. Viz.,

1. Either I am in the basement level or a simulation.
2. If I am in the basement level, then basement-level Margaret has completed the program.
3. If I am in a simulation, then basement-level Margaret has completed the program.
4. Therefore, basement-level Margaret has completed the program.

She should, therefore, believe that she inhabits a simulation if she expects to run the program.[12] Accordingly, the more certain Margaret becomes that she will run the program, the more certain she should become that doing so will cause her to be terminated.

[12] Decision-theory enthusiasts might be interested to note that it is possible for $\mathrm{p}(S \mathrel{\Box\!\!\rightarrow} T) < .5$, even while $\mathrm{p}(S \mathrel{\Box\!\!\rightarrow} T \mid S) > .5$ (i.e., Margaret might be confident that creating simulations would probably not cause termination, while simultaneously thinking that creating simulations would probably cause termination conditional on it being true that she will create them). Since causal expected utilities are determined by $\mathrm{p}(S \mathrel{\Box\!\!\rightarrow} T)$ and not $\mathrm{p}(S \mathrel{\Box\!\!\rightarrow} T \mid S)$ (i.e., they are determined using the unconditional probabilities of dependency hypotheses, and not the probabilities of dependency hypotheses conditional on the agent's choice), it is possible for the causal expected utility of creating simulations to exceed that of not creating them even when $\mathrm{p}(S \mathrel{\Box\!\!\rightarrow} T \mid S) > .5$. Andy Egan [2007] has claimed that this result is highly counterintuitive, and that it amounts to a counterexample to causal decision theory. One of Egan's ostensible counterexamples, *The Psychopath Button*, mirrors the structure of Margaret's decision situation on the assumption that $\mathrm{p}(S \mathrel{\Box\!\!\rightarrow} T) < .5$ and $\mathrm{p}(S \mathrel{\Box\!\!\rightarrow} T \mid S) > .5$. See Egan, 2007, 97. Wedgwood [2013] goes so far as to offer a replacement for causal decision theory in light of *The Psychopath Button*,

Given this, it seems that deciding to run the program would be irrational from the causal perspective.

However, what makes the recommendation of causal decision theory complicated is that the opposite is also true: the more certain Margaret becomes that she will *not* run the program, the more certain she should become that doing so will *not* cause her to be terminated. This kind of decision situation, according to causal decision theorists like Arntzenius [2008] and Joyce [2012], results in an intractable *instability* in the agent's epistemic perspective. Thus, according to Arntzenius [2008] and Joyce [2012], while causal decision theory does not recommend running the program, it also does not recommend refraining from running it either. I will leave further discussion of this point to a footnote, but it is a fascinating way in which decision-theoretic discussions of instability are relevant to simulation theorizing, and likely an important area of future research.[13]

which has been called "benchmark theory." Briggs [2010] concurs with Egan's and Wedgwood's intuitions about the rational decision in *The Psychopath Button*, but argues that Wedgwood's theory suffers from serious problems.

[13] Given their analyses of Egan's [2007] *The Psychopath Button*, Arntzenius [2008] and Joyce [2012] might argue that causal decision theory does not produce a straightforward recommendation in Margaret's decision situation. Rather, they might claim that the application of causal decision theory results in an intractable *instability* because both options are *causally unratifiable*. Consider the following conditional probabilities of termination:

$$\mathrm{p}(S \mathrel{\Box\!\!\rightarrow} T \mid S) > .5$$
$$\mathrm{p}(S \mathrel{\Box\!\!\rightarrow} T \mid \neg S) < .5$$
$$\mathrm{p}(\neg(S \mathrel{\Box\!\!\rightarrow} T\ ) \mid S) < .5$$
$$\mathrm{p}(\neg(S \mathrel{\Box\!\!\rightarrow} T\ ) \mid \neg S) > .5$$

Given these conditional probabilities, Arntzenius and Joyce would argue that Margaret should become more confident in $S \mathrel{\Box\!\!\rightarrow} T$ as she becomes more confident that she will choose S, and more confident in $\neg(S \mathrel{\Box\!\!\rightarrow} T)$ as she becomes more confident that she will choose $\neg S$. As such, Arntzenius and Joyce would argue that rational deliberation on Margaret's part results in shifting calculations of the causal expected utilities of S and $\neg S$ with no obvious resolution. A major problem with their analysis is that it gives counterintuitive recommendations in other cases involving causal unratifiability. For example, as Ahmed [2014] shows, it recommends not using a randomization device in the *Death in Damascus* case even when this saves the life of half who do so. See Greene [2018] for discussion of Ahmed's example

Let us now turn to what Margaret should believe about her situation before she has completed the program. If she believes that she will decide not to run the program, and she believes that she will so decide for reasons that would be applicable on the basement level (such as her desire not to risk termination, which given the epistemic situation would be equally applicable on the basement level), then she should believe that she does not inhabit a simulation. This inference would, therefore, undermine the simulation hypothesis for any rational agent in Margaret's situation. If Margaret believes, instead, that she will decide to run the program, and that she will so decide for reasons that would be applicable on the basement level (such as her desire to answer counterfactual questions about her life), then she can infer that she *does* inhabit a simulation.

Finally, let us consider whether Margaret should have created the simulation program in the first place. If she believes that she will not use the program, then, presumably, she should not create it. Therefore, unless there is some independent reason to develop the technology but not run it, Margaret has strong reason to not pursue past-self simulations.

3.3 Implications for the Actual World

Let us now consider the implications for the actual world. As we have seen, Margaret has strong reasons not to create past-self simulations and to believe that she does not inhabit one. While this is relatively clear in Margaret's case, the important object of reflection, for our purposes, is how our situation differs. Do we have the same reasons that Margaret does? Answering this question involves some speculation, due to the absence of simplifying assumptions, but when something as important as the termination of our world is at stake, it is essential to take these possibilities seriously even if they require some speculation.

To start, consider that in Margaret's case we assumed that she is certain that creating simulations from within a simulation causes termination, but we lack the corresponding certainty when it comes to creating ancestor simulations from within an ancestor sim-

and a diagnosis of the disagreement. In any event, as Egan [2007], Briggs [2010], and Wedgwood [2013] would claim, running the computer is intuitively irrational.

ulation. However, in Margaret's case, we also assumed that the creation of past-self simulations is as beneficial as death would be costly. The reasons to not create ancestor simulations become stronger as we make this ratio more in line with the intuitive idea that the termination of our level of reality is many times worse than the creation of ancestor simulations is good. The disutility of complete destruction motivates the following premise:

> *Scientific-Conduct Premise*: Unless it is exceedingly improbable that an experiment would result in our destruction, it is not rational to run the experiment.

This premise is deeply ingrained in modern scientific practice.

A more familiar example of this sort of reasoning may help strengthen its plausibility. Research into particle physics, such as that conducted at CERN's Large Hadron Collider, has the potential to produce both new theoretical insights and advancements in technology. However, prior to the completion of the LHC, there was much discussion regarding the safety of the high-energy particle-collision experiments that would be conducted. For the most part, the worries over safety—including dubious doomsday scenarios involving microscopic black holes—were confined to the general public, and serious concerns over these issues among physicists did not survive peer review. The LHC Safety Group [2003] concluded that the collisions posed no threat. If, however, they had found that there is a small chance—say, 1%, or perhaps even .001%—that research conducted at the LHC would result in a giant implosion, then the rewards of performing these experiments would have to be extreme for the research to continue. Furthermore, it would have to be proven that similar benefits could not be attained through less dangerous research. Rejecting this perspective would require a significant change in basic human values. Extrapolating from this perspective would show that an uncertain termination risk leaves the relevant considerations relatively unchanged from those of facing a certain one.[14]

[14] An anonymous referee points out that some may view the termination of our simulation as much less bad than basement-level destruction. For example, one may appeal to the fact that sentient life would

The same point applies to a situation in which Margaret's program allows for finite nesting of simulations before termination. If Margaret knows how many layers are permitted, then she can assign explicit probabilities less than one to termination. However, the risk of termination would remain non-negligible unless a very large number of nested simulations are allowed. Similarly, the possibility that posthumans would allow for some finite nesting of simulations serves to lessen the termination risk, but seemingly not enough to overcome the scientific-conduct premise. It would be impossible to know how many nested ancestor simulations are permitted; the simulators may even prefer to use all their computational resources on non-nested simulations.

Another potential asymmetry is that one might expect a reward upon the termination of our simulation, rather than annihilation, as Margaret does. Perhaps we will not be terminated but rather uploaded to paradise upon termination of our simulation. Perhaps, even, our simulators will feel morally compelled to do so.

The main worry about this response is that current moral beliefs differ widely, and this makes it especially difficult to predict the moral beliefs of posthumans.[15] Whether it would even be morally permissible to *create* our world is an extraordinarily difficult question that concerns such familiar debates as those over the problem of evil and the non-identity problem. Similarly, we can speculate about moral beliefs regarding termination — perhaps by studying trends in the ethics of medical research, as Jenkins [2006, 26–31]

continue after our simulation is terminated, or that beings very similar to us would continue to exist (if there are very many similar simulations running in parallel). However, many cosmologists believe that the situation is the same even if we do live in the basement: there is probably much sentient life in the observable universe that would continue after our destruction, and the entire universe is probably infinite and therefore contains many beings just like us. (See Bostrom, 2011 for a discussion of these views and their potential impact on moral theory.) Alternatively, one might prefer simulation termination because one believes we should defer judgment to our simulators about whether our universe should exist, or because basement-level destruction implies a massive amount of "astronomical waste" (see Bostrom, 2003a for a discussion of astronomical waste).

[15] Though we can note that some uses of simulation technology are more problematic than others. Consider, for example, the proposed development of simulations for studying criminal sentencing [Auerhahn, 2008], and imagine a corresponding advanced simulation of jails.

does in an interesting paper — but this is unlikely to be a method for predicting the moral beliefs of posthumans reliable enough to rule out the termination risk.

Instead, we should again note the force of the scientific-conduct premise and see that it leaves the relevant considerations unchanged given our current state of uncertainty about the moral beliefs of posthumans. If we gain strong reasons to believe that uploading-to-paradise would follow termination, then that would indeed be an important and exciting asymmetry between our situation and Margaret's. For now, such a hypothesis is similar to speculative theology.

A different kind of potential asymmetry lies in the assumptions that Margaret's simulator would be like her in being rational and assigning great disutility to termination. In fact, it seems Margaret needs to take into account the possibility that her simulator is different in at least one of these ways (either because basement-level Margaret is different or Margaret is actually part of the past-self simulations of some other agent). Similarly, it is possible that our posthuman simulators either do not assign great disutility to termination or that they did not handle the decision rationally. Perhaps this is more difficult to believe when we imagine large-scale multi-agent scientific endeavors like the Large Hadron Collider as our model for the creation of ancestor simulations, but eventually, the technology enabling the creation of ancestor simulations could become so trivial that individuals could decide for themselves to create them.

A worry for this argument is that similar considerations apply to *any* dangerous technology. For example, unless effective restrictions are put in place, the technology enabling the creation of nuclear weapons, or vicious nanobots, will also become so trivial that individuals could decide for themselves whether to use it. Should we, therefore, conclude that technological progress ensures these routes to annihilation? If so, then the "doomsday" horn of Bostrom's trilemma is true. However, if this horn is false, then it will be because advanced civilizations do indeed tend to assign great disutility to annihilation, and, furthermore, they tend to figure out how to prevent the use of dangerous technologies before they destroy them. If the arguments of this paper are sound, then ancestor-simulation technology is even more dangerous than nanobots.

We should also note that this asymmetry cannot give posthumans *more reason* to create ancestor simulations. It *could* serve to break the decision instability: they would no longer be confident that they are not in a simulation conditional on deciding not to create their own. In that case, the reasons for refraining from creating ancestor simulations would be strengthened.

Finally, one might lodge a “bad company” objection against the application of Margaret’s conclusion to the actual world. The objection is that the same considerations that tell against creating ancestor-simulation technology also tell against creating any sort of advanced technology, since all technologies add to the computation that would be necessary to simulate humanity. In fact, computational demands increase as the population increases, when we explore other planets, and so on. Should we attempt to stop these things as well?

While we can grant that non-simulation technology may imply some sort of termination risk, it is of a different character from that implied by ancestor simulations. Recall that the creation of simulation technology poses a problem for basement-level computation because it creates an unbounded explosion of computational demands if left unchecked. In contrast, even the most sophisticated non-simulation technology — such as a mega computer proving theorems or one simulating billions of agents experiencing euphoria — requires a bounded and (in-principle) predictable amount of computation to run. But a simulation computer (including a mega one running billions of ancestor simulations) requires a check on stacking to be a useful tool. Termination of simulations is the most straightforward check. Thus, at the least, the creation of ancestor simulations seems to pose a greater risk of termination than other increases in computational complexity, and this would be enough to establish the applicability of Margaret’s inferences to the actual world.

In sum, while there are many possible asymmetries between Margaret’s situation and our own, it is not clear that any succeed in showing that the same inferences are not applicable to the actual world. Since the ostensible asymmetries are so speculative, I am hesitant to conclude that our situation and Margaret’s are relevantly different. Therefore, given the magnitude of the potential outcomes, we currently do not possess enough certainty about any asymmetry premise to rationally rely on it in practical reasoning about simulation technologies or theoretical reasoning about the simulation hypothesis. If so, then we should take Margaret's inferences to be a starting point for reasoning in both domains.

4 The Termination Risk of Simulation Investigations

Experimental probes represent a significant development in the study of the ancestor-simulation hypothesis. Previously, research into the hypothesis has only involved philosophical theorizing. In contrast to experimental probes, philosophical theorizing does not attempt to create observational divergences from the basement level. In other words, theorizing allows us to increase our confidence that we live in an ancestor simulation without the need for observations that would only occur in a simulated reality. (On the contrary, the same theorizing would occur on the basement level). For this reason, the benefits and costs of observational probes differ from those of philosophical theorizing. Since we are currently on the brink of creating such probes, and any associated termination risk is of relevance to near-future physics, now is the time to reflect on their expected value.

Beane et al. [2014] claim that if our universe is an ancestor simulation, then there might be observable consequences in the spectrum of high-energy cosmic rays. They speculate that the required measurement is achievable with current technology if we live in an "early" (i.e., low-grade) simulation. Even if we do not live in an early simulation, we still should expect some in-principle observability of a simulated universe if the computational resources of the basement level are finite. Beane et al. note: "Assuming that the universe is finite and therefore the resources of the potential simulators are finite, then a volume containing a simulation will be finite and space-time will be discretized, and the discretization will imprint itself upon our universe and may be observable" [8].

Undoubtedly, some will regard these sorts of experiments as having negative value because they are certain that we do not live in a simulation. However, for those that assign some probability to the ancestor-simulation hypothesis, such as the experimental researchers themselves, what is the expected value of these experiments?

Imagine that you are uncertain whether the gun you are holding is loaded with a bullet. One way to test is to play a single round of Russian roulette. If the gun does not discharge, then this creates weak evidence that the gun is unloaded. Call this the "boring" result. Call the result in which the gun discharges the "shocking" result. The shocking result proves that the gun was loaded, but, in some important sense, it does not matter, because you are dead. Since the shocking result has negative value, it is irrational to use Russian roulette as a means of investigating unless you assign an overriding positive value to the boring result.

How do the roulette example and the situation with simulation probes differ? The equivalent boring result is one in which experimenters find only very weak evidence that we do not live in a simulation. The shocking result is one in which they discover evidence that shows that we do. On the one hand, experimental observation that fails to detect abnormalities cannot get beyond the boring result. As Beane et al. [2014] note, there are many plausible ways in which we could inhabit a simulated reality and not be able to detect it through observing cosmic rays. More fundamentally, no matter how sophisticated our experiments become, they could *never* allow us to conclude that we do not live in a simulation, since if we live in a simulation, then all our observations are part of the simulation programming.[16] To show that we do not live in a simulation an argument must conclude: *this pattern of observation could not be programmed into a simulation.* Physicists have so far not made claims like this, and for good reason. Demonstrating that we do not live in a simulation does not seem to be an attainable goal of experimental observation.[17]

[16] Cf. Nozick, 1981, 591: "I don't say there is no ground floor..., just that we wouldn't know it if we reached it."

[17] The distinction between simulating the world and simulating observations of the world is important to keep in mind when evaluating evidence relating to the simulation hypothesis produced by physics. Recently, for example, Ringel and Kovrizhin [2017] have shown that not all quantum systems can be quickly simulated using classical computing. Many news articles reported this as proof our world is

On the other hand, experimental observation of abnormalities can demonstrate that we *do* live in a simulation. Abnormalities would demonstrate that we live in a simulation by, again, a type of constructive dilemma. Viz.,

1. If the observations are actual observations of a non-basement reality, then we live in a simulation.
2. If the observations are illusions created by the simulation's programming, then we live in a simulation.
3. Either way, we live in a simulation.

Given the negligible value of the boring result, if experimental probes have non-negligible positive expected value it must be because of the value of the shocking result. Therefore, we should focus our attention on determining the expected value of the shocking result. To do this, we need to weigh the potential benefits and costs of a probe demonstrating that we inhabit a simulation.

Much of the calculation depends on our beliefs about what is likely to happen to an ancestor simulation in which the shocking result occurs. Since such a result creates a significant divergence between the behavior of a basement-level civilization and a simulated one, it is likely to destroy the value of a counterfactual historical simulation in just the way that large modifications do (discussed in Section 3.1). If the inhabitants of an ancestor simulation learn that they inhabit a simulation, and this has a significant effect on the course of human history, then the value of the simulation for answering counterfactual social-scientific questions is destroyed. Perhaps an ancestor simulation used for entertainment purposes would be more amenable to such a divergence, but further speculation is awkward given our current lack of evidence as to what type of ancestor simulation we occupy.

unsimulated, with headlines like "Physicists Confirm That We're Not Living In a Computer Simulation" [Eck, 2017]. However, Ringel and Kovrizhin's result does not show that our universe is unsimulated (even if we accept their assumptions about computing constraints). To show that, they would need to make a claim about the computability of our *observational patterns*, and not of hypothesized quantum systems.

As argued in Section 1, ancestor simulations hold tremendous potential for research purposes as a tool for studying counterfactual conditionals. Since the study of counterfactual conditionals requires a great number of simulations, our starting assumption should be that if we inhabit an ancestor simulation, then, at the least, it is not more probable that we inhabit an entertainment simulation than that we inhabit a research simulation (as argued in Section 3.1). If this is true, then we should seriously consider the possibility that, conditional on our universe being an ancestor simulation, it is a counterfactual historical simulation. If simulated inhabitants in a counterfactual historical simulation observe that they inhabit a simulation, and this has a significant impact on the course of human events, then the data provided by such a simulation would no longer suit its purposes, with one exception. The exception would be counterfactual historical simulations that are explicitly designed to study what would happen if people found out they inhabit a simulation. For other types of counterfactual historical simulations, such as one studying the effects of world wars on the development of political systems, the discovery that the world is simulated (which represents an additional large divergence from actuality) would make the data unreliable.

Given, therefore, that extrapolation from current trends in computing and simulation technologies point to termination as the possible result of a successful simulation probe, how should we think of their expected value? The most obvious benefit is that of adding to our knowledge of reality. Would that benefit outweigh the potential cost?

Most would believe that the termination of our level of reality is many times worse than adding to our knowledge of reality is good. This premise is deeply embedded in human scientific practice. Given this, an experiment with a similar utility and probability profile to those proposed by simulation investigators might not be tolerated if proposed in a more familiar context. Consider a proposal for an experiment at the Large Hadron Collider with the following properties: *This experiment is very unlikely to succeed in producing an interesting result, but if it does succeed, it will reveal a shocking truth about our universe and may cause its annihilation.* The fact that people do not view experimental simulation investigations from a similar perspective may be the result of bias — due to the abstract or “science-fictional” seeming nature of the ancestor-simulation hypothesis. However, those who care about the simulation hypothesis enough to test it presumably do not view the hypothesis as abstract or science-fictional. Ironically, from their perspective the risk should seem the greatest.

For these reasons, the benefit of knowledge alone may not be a sufficient justification for experimental simulation investigations. As discussed in Section 3.3, it is possible that our simulators would give us some reward for discovering that we live in a simulation. This possibility perhaps has the power, in principle, to provide a sufficient justification for experimental probes when combined with the possibility of knowledge acquisition. As far as I am aware, however, the possibility of reward has never been seriously put forward in defense of such investigations. More reflection on these issues is therefore required before we can reasonably regard experimental simulation investigations to have positive expected value.

4.1 Pascal’s Wager?

The most prominent response to this argument, in my experience, is that it mirrors the structure of Pascal’s Wager, and is therefore susceptible to similar objections. Actually, the argument is very different from Pascal’s Wager. It is worth noting these differences in detail, and explaining why objections to the Wager, therefore, do not apply.

What the argument has in common with the Wager is that in each instance we are appealing to possibilities about how our actions may affect the behavior of a creator. This, however, is where the commonalities stop. Translated into the language of modern decision theory, Pascal’s argument is that we should wager for God because the expected value of doing so is infinite while that of the alternative is merely finite. As detailed by Hájek [2003], the most important objections to Pascal can be broken into three categories. They are: i) one might assign probability 0 to God’s existence [Rescher, 1985], ii) the notion of infinite utility is incomprehensible [Jeffrey, 1983, McClennen, 1994], or the use of such a notion renders the argument invalid [Rescher, 1985, Hájek, 2003], and iii) perhaps there is more than one God [Diderot, 1746], or perhaps God does not reward those who wager on the basis of an expected utility calculation [James, 1896].

These categories serve to highlight the superiority of our discussion over Pascal's Wager. First, an argument against simulation probes can accept the possibility that one assigns probability 0 to the simulation hypothesis. If so, one will not be interested in simulation probes because one is already certain of their result. Second, and perhaps most importantly, the argument need not appeal to infinite utilities, and so does not risk incomprehensibility or invalidity. Relatedly, third, note that the objections of (iii) gain their force precisely because of Pascal's appeal to infinite utilities: if a possibility other than wagering for God in Pascal's way is assigned infinite utility, then it must be assigned probability 0 or else Pascal's argument is not sound. Thus the objections of Diderot and James gain their force.

Pascal's Wager can, of course, be reformulated to avoid these objections by removing the appeal to infinite utilities. If Pascal had merely pointed to some hypothesis about God that he took to be probable and assigned specific outcomes finite utilities, then the argument would have no in-principle objection—this would be standard philosophical debate. The important issues, in that case, would be the evidence we have for the possibilities and the assignments of utility. This is exactly the format of the potential arguments against simulation technologies and probes. We have a finely specified possibility—we inhabit an ancestor simulation programmed to terminate given certain events—and we have the suggestion that such a possibility has great, but finite, disutility. How this should affect our practical decision making depends on one's agreement with the assigned probabilities and utilities.[18]

[18] For an example of a paper that focuses on the sorts of behaviors that posthumans would find pleasing—and which seems to have more in common with Pascal's Wager—see Hanson, 2001. Even this argument, however, is significantly different from Pascal's Wager because it does not appeal to infinite utilities.

5 Conclusion

We have explored the expected value of two aspects of simulation science in light of the termination risk implied by the ancestor-simulation hypothesis. First, it was shown that the creation of ancestor simulations entails a risk of termination that must be accounted for in the decision to create them. This decision was examined further, through the lens of decision theory, with the help of a simplified model. Applied to the actual world, the model would show that we should believe that we exist in the basement level but should refrain from creating ancestor simulations if given the chance. To resist this conclusion, one needs to either reject the conception of rationality assumed by the model, or posit a relevant asymmetry. Several potential asymmetries were considered, but none were found to be fully compelling. Further research should capitalize on the ongoing debate over causal and evidential conceptions of decision making, as this issue impacts simulation theorizing directly.

We have also discovered implications for the safety of the sort of experimental simulation investigations that scientists could conduct in the near future. These experiments may have negative expected value given the possibility that we inhabit a counterfactual historical simulation, and given the assumption that the destruction of our universe is many times worse than adding to our knowledge of reality is good. In the popular press, it is often remarked in a hopeful way that one day experimental probes will demonstrate that we live in a simulation. On the contrary, if we live in a counterfactual historical simulation, then perhaps it would be best if our simulation is advanced enough to ensure that this day never comes. In any case, scientists hoping to conduct such probes need to do more than merely appeal to the ancestor-simulation hypothesis in justifying their research. They must also countenance the termination risk.

References

Arif Ahmed. Dicing with death. *Analysis*, 74(4):587–92, October 2014.

Frank Arntzenius. No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis*, 68(2): 277–297, 2008.

Kathleen Auerhahn. Using simulation modeling to evaluate sentencing reform in California: Choosing the future. *Journal of Experimental Criminology*, 4(3):241–266, 2008.

Silas R. Beane, Zohreh Davoudi, and Martin J. Savage. Constraints on the universe as a numerical simulation. *The European Physical Journal A*, 50(148), 2014. arXiv:1210.1847.

Jonathan Birch. On the 'simulation argument' and selective scepticism. *Erkenntnis*, 78: 95–107, 2013.

Nick Bostrom. Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9(1), March 2002.

Nick Bostrom. Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, 15(3):308–14, 2003a.

Nick Bostrom. Are we living in a computer simulation? *The Philosophical Quarterly*, 53 (211):243–55, 2003b.

Nick Bostrom. The simulation argument: Reply to Weatherson. *The Philosophical Quarterly*, 55(218):90–7, 2005.

Nick Bostrom. Do we live in a computer simulation? *New Scientist*, 192:38–39, 2006.

Nick Bostrom. The simulation argument: Some explanations. *Analysis*, 69(3):459–61, July 2009.

Nick Bostrom. Infinite ethics. *Analysis and Metaphysics*, 10:9–59, 2011.

Rachel Briggs. Decision-theoretic paradoxes as voting paradoxes. *The Philosophical Review*, 119(1):1–30, 2010.

David J. Chalmers. The Matrix as metaphysics. In *The Character of Consciousness*, pages 455–79. Oxford University Press, 2010.

Milan M. Ćirković. Observation selection effects and global catastrophic risks. In Nick Bostrom and Milan M. ÂŽCirkoviÂŽc, editors, *Global Catastrophic Risks*, pages 120–145. Oxford University Press, 2008.

Denis Diderot. *Pensées Philosophiques*. Reprinted in Kessinger Publishing, 2009, 1746.

Allison Eck. Physicists confirms that we're not living in a computer simulation. *PBS*, Retrieved from http://www.pbs.org/wgbh/nova/next/physics/physicists-confirm-that-were-not-living-in-a-computer-simulation/, 2017.

Andy Egan. Some counterexamples to causal decision theory. *The Philosophical Review*, 116(1):93–114, 2007.

Rawad El Skaf and Cyrille Imbert. Unfolding in the empirical sciences: Experiments, thought experiments and computer simulations. *Synthese*, 190(16):3451–74, November 2013.

Preston Greene. Success-first decision theories. In Arif Ahmed, editor, *Newcomb's Problem*. Cambridge University Press, 2018.

LHC Safety Study Group. Study of potentially dangerous events during heavy-ion collisions at the LHC. *CERN*, 2003.

Till Grüne-Yanoff. The explanatory potential of artificial societies. *Synthese*, 169(3): 539–55, August 2009.

Till Grüne-Yanoff. Artificial worlds and agent-based simulation. In Ian C. Jarvie and Jesus Zamora-Bonilla, editors, *The Sage Handbook of the Philosophy of the Social Sciences*, pages 613–31. Sage, 2011.

Till Grüne-Yanoff and Paul Weirich. The philosophy and epistemology of simulation: A review. *Simulation and Gaming*, 41(1):20–50, 2010.

Alan Hájek. Waging war on Pascal's wager. *The Philosophical Review*, 112(1):27–56, January 2003.

Robin Hanson. How to live in a simulation. *Journal of Evolution and Technology*, 7(1), 2001.

William James. The will to believe. In J. J. McDermott, editor, *The Writings of William James*. Random House, 1896.

Richard Jeffrey. *The Logic of Decision*. University of Chicago Press, 1983.

Peter S. Jenkins. Historical simulations — motivational, ethical and legal issues. *Journal of Futures Studies*, 11(1):23–42, August 2006.

James M. Joyce. Regret and instability in causal decision theory. *Synthese*, 187(1): 123–45, 2012.

David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59(1):5–30, March 1981.

Edward McClennen. Pascal's wager and finite decision theory. In Jeff Jordan, editor, *Gambling on God: Essays on Pascal's Wager*, pages 115–37. Rowan I& Littlefield, 1994.

Robert Nozick. *Philosophical Explanations*. Harvard University Press, 1981.

Martin Peterson. *An Introduction to Decision Theory*. Cambridge University Press, 2009.

Hilary Putnam. *Reason, Truth, and History*. Cambridge University Press, 1981.

Nicholas Rescher. *Pascal's Wager*. Notre Dame University Press, 1985.

Zohar Ringel and Dmitry L. Kovrizhin. Quantized gravitational responses, the sign problem, and quantum complexity. *Science Advances*, 3(9), September 2017.

Vince Stricherz. Do we live in a computer simulation? UW researchers say idea can be tested. *The University of Washington [Press Release]*, Retrieved from http://www.washington.edu/news/2012/12/10/do-we-live-in-a-computer-simulation-uw-researchers-say-idea-can-be-tested/, 2012.

John Tierney. Our lives, controlled from some guy's couch. *The New York Times*, Retrieved from http://www.nytimes.com/2007/08/14/science/14tier.html, 2007.

Phil Torres. *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks*. Pitchstone Publishing, 2017.

Brian Weatherson. Are you a sim? *The Philosophical Quarterly*, 53:425–31, 2003.

Ralph Wedgwood. Gandalf's solution to the Newcomb problem. *Synthese*, 190(14):2643–75, 2013.