Effective Altruism, Disaster Prevention, and the Possibility of Hell:
A Dilemma for Secular Longtermists

Eric Sampson

Abstract: Longtermist Effective Altruists (EAs) aim to reduce the risk of existential catastrophes. This paper identifies a catastrophic risk that EAs have almost entirely ignored. I call it *religious catastrophe*: the threat that, as Christians and Muslims have long warned, billions may suffer eternal damnation. I argue that, even by secular EA standards, religious catastrophe is at least as bad, at least as probable, and therefore at least as important as many familiar EA catastrophic risks, such as catastrophic climate change, stellar explosions, or global nuclear winter. Secular EAs therefore face a dilemma: either adopt religious catastrophe as an EA cause, or ignore it, but also ignore all EA causes with a comparable or lower expected value (i.e., most or all of them). Either way, business as usual—ignoring religious catastrophe while championing the standard EA causes—is incompatible with secular longtermist principles.

Keywords: Effective Altruism; Longtermism; Existential Risk; Expected Utility; Consequentialism; Pascal's Wager; Pascal's Mugging

Word Count: 6,736

1. Introduction

Effective Altruists (EAs) aim to do as much good as they can with the limited resources they have. The movement began with a focus on urgent present-day problems such as global poverty and factory farming.[1] Those concerns remain important, but the movement's attention has increasingly shifted toward the distant future. EAs now argue that, although our influence over the distant future is less certain, the value at stake across such a vast stretch of time is astronomically greater than the value generated by solving even the most pressing problems here and now. Thus, the *expected* value of positively shaping the long-term future vastly exceeds that of securing certain benefits now. EAs call this new outlook *longtermism*.

[1] Singer (1972) and Singer (1975) have been especially influential in EA circles.

Many longtermists believe that the best way to maximize expected good over the long run is to reduce *existential risk*: catastrophic events that would cause human extinction. Toby Ord, one of EA's founders, makes this case in *The Precipice: Existential Risk and the Future of Humanity* (2020), a nearly 500-page catalog and assessment of every existential threat (he thinks is) worth taking seriously. These risks fall into two broad categories: anthropogenic catastrophes, such as nuclear war, engineered pandemics, misaligned AI,[2] and catastrophic climate change; and natural catastrophes such as supervolcanoes, natural pandemics, stellar explosions, and asteroid impacts. Ord argues that existential risk is the defining moral issue of our time because so much value hangs in the balance. If we destroy ourselves, all future value is lost. If we survive, unimaginable value lies ahead.

This paper identifies a catastrophic risk that EAs have almost entirely ignored. I call it *religious catastrophe*: the threat that, as Christians and Muslims have warned for centuries, billions may suffer eternal damnation. I argue that, even by secular EA standards, religious catastrophe is at least as bad, at least as probable, and therefore at least as important, as the catastrophic risks that have dominated EA discussion. Secular EAs therefore face a dilemma: either adopt religious catastrophe as an EA cause, or ignore it, but also ignore all EA causes with a comparable or lower expected value (i.e., most or all of them). Either way, business as usual—ignoring religious catastrophe while championing the standard EA causes—is incompatible with secular longtermist principles.

2. What is Religious Catastrophe and How Bad Would it Be?

Orthodox forms of Christianity and Islam hold that eternal bliss awaits those who embrace the one true religion, while eternal punishment awaits those who don't. To be sure, both traditions also contain universalist strands according to which everyone is ultimately saved. But I set those views aside here, since my argument requires only that, on orthodox versions of these religions, some people

[2] Bostrom (2014) is the canonical text on the dangers advanced artificial intelligence may pose to humanity.

are damned. Consider, for example, this famous passage from *Matthew* 25:31-46 (ESV; emphases added):

> When the Son of Man comes into his glory, and all the angels with him, then he will sit on his glorious throne. Before him will be gathered all the nations and he will separate people one from another as a shepherd separates the sheep from the goats. And he will place the sheep on his right, but the goats on the left. Then the King will say to those on his right, "Come you who are blessed by my Father, inherit the kingdom prepared for you from the foundation of the world. For I was hungry and you gave me food…
>
> Then he will say to those on his left [the goats], '*Depart from me you cursed, into the eternal fire prepared for the devil and his angels*. For I was hungry and you gave me no food…Truly, I say to you, as you did not do it to one of the least of these you did not do it to me.' *And these will go away into eternal punishment, but the righteous into eternal life*."

The Qur'an contains many similar passages (emphasis added):

> "And if you are in doubt about what We have revealed to Our servant, then produce a chapter like these, and call your witnesses apart from Allah, if you are truthful. But if you do not—and you will not—then *beware the Fire whose fuel is people and stones, prepared for the disbelievers*." (2:23-24).
>
> "Those who reject Our revelations—*We will scorch them in a Fire. Every time their skins are cooked, We will replace them with other skins, so they will experience the suffering*. Allah is Most Powerful, Most Wise. As for those who believe and do good deeds, We will admit them into Gardens beneath which rivers flow, abiding therein forever…" (4:56-57).
>
> "As for those who disbelieve, *garments of fire will be tailored for them, and scalding water will be poured over their heads. Melting their insides and their skins*. And they will have maces of iron. Whenever they try to escape the gloom, they will be driven back to it: 'Taste the suffering of burning.' But Allah will admit those who believe and do good deeds into Gardens beneath which rivers flow" (22:19-23).[3]

You get the point. In both sacred texts, hell is depicted as a state of *unending* punishment, anguish, misery, and loss. Not good. No doubt, many believers interpret these descriptions metaphorically rather than literally—as images of alienation, despair, regret, meaninglessness, self-loathing, or separation from God rather than physical torment. But nothing in my argument depends on a literal

[3] https://m.clearquran.com/downloads/quran-english-translation-clearquran-edition-allah.pdf

reading. No matter how we interpret the imagery, hell is unimaginably bad. If such a fate awaits those who fail to believe or act correctly, then few things could be more morally significant.

In both traditions, a minority of adherents argue that a perfectly good God would never subject anyone—even the very wicked—to eternal punishment of this sort. They therefore propose revisionary interpretations of these passages to avoid attributing such actions to God. I am sympathetic to this line of thought. I, too, am inclined to think that a perfectly loving God would not subject anyone to eternal conscious torment. But I'm not *certain* about that. "God would never punish someone eternally" does not strike me as epistemically comparable to something like "2+2=4." I retain some (reasonable, in my view) doubts. Perhaps I misunderstand the gravity of sin or the demands of perfect justice and goodness. Perhaps, as some theologians argue, eternal punishment is not inflicted merely for finite sins, but rather because the damned *continue* sinning eternally under the conditions of hell—persisting forever in pride, resentment, deceit, rebellion, or hatred. We need not settle these theological questions here. What matters for present purposes is simply this: we cannot reasonably be *certain* that no such punishment awaits those who reject the correct religion (assuming there is one).

So, how bad would it be if large swathes of humanity were damned for eternity? Unimaginably bad, obviously. Indeed, it seems easy to show that eternal damnation for billions is at least as bad (and almost certainly far worse) than *any* of the catastrophic risks that dominate secular EA discussions. There are two ways to make this case. The first is a "one-shot" argument. The second is piecemeal.

The one-shot argument emphasizes the contrast between the infinite (or finite but ever-increasing) badness of religious catastrophe and the finite badness of standard EA catastrophes. Consider unaligned AI. Suppose it exterminates humanity, cruelly and painfully, wiping out all present and future value. However terrible this outcome would be, its badness remains finite. Only a finite number of people would die or fail to exist, and only a finite amount of value would be lost, since the

universe itself can persist only for a finite period before eventual heat death. By contrast, when even one person is condemned to hell for eternity, they experience infinite (or finite but ever-increasing) suffering and infinite (or finite but ever-increasing) loss since they forfeit eternal bliss. Thus, *even one person*'s eternal damnation would be worse than humanity's extinction by unaligned AI. Eternal damnation for billions would be incalculably worse. And note that there is nothing special about AI here. The same finite/infinite asymmetry arises when religious catastrophe is compared with any standard EA catastrophe. That's the one-shot argument that religious catastrophe is at least as bad as—and likely unimaginably worse than—the catastrophes that dominate longtermist EA discourse.

The piecemeal argument proceeds by comparing religious catastrophe case by case with each EA existential risk. Which is worse: eternal damnation for billions of people, or catastrophic climate change culminating in human extinction? Which is worse: eternal damnation for billions, or nuclear war and extinction? Which is worse: eternal damnation for billions, or an asteroid impact that wipes out humanity? And so on, across the familiar catalog of existential threats. In each comparison, eternal hell for billions—especially when contrasted with the eternal bliss they otherwise might have enjoyed—seems *far worse* than the corresponding secular catastrophe. But my argument does not require that stronger claim. It's sufficient for present purposes that the badness of religious catastrophe is comparable to that of the standard EA existential risks.

3. How Probable is Religious Catastrophe Compared to Other Catastrophic Risks?

If damnation for billions is at least as bad as every EA catastrophic risk, then EAs can justify ignoring religious catastrophe only by assigning it a zero or infinitesimal probability. In this section, I argue that assigning a zero or infinitesimal probability to religious catastrophe is irrational. Even if such a catastrophe is unlikely, the evidence does not justify treating it as effectively impossible.

To see this, we need to distinguish several kinds of probability. There are *subjective* probabilities: the credences people in fact assign to propositions. There are *evidential* probabilities: the credences people are epistemically justified in assigning, given the evidence. There are *frequentist* probabilities: probabilities grounded in the observed frequencies of outcomes in past cases. And there are *objective* probabilities: mind-independent propensities for events to occur. Our concern here is evidential probability—the degree of confidence a rational agent ought to assign to a proposition, represented on a scale from 0 (certainty the proposition is false) to 1 (certainty it's true). No doubt many EAs assign a vanishingly small (or even zero) *subjective* probability to religious catastrophe. But I will argue that this is irrational, since the available evidence does not warrant such extreme confidence. Indeed, I will argue that the evidential probability of religious catastrophe lies at least in the same general neighborhood as the probabilities routinely assigned to many standard existential risks.

Let's begin by considering the probabilities assigned to the standard EA catastrophic risks. Doing so will clarify how probable religious catastrophe would need to be to warrant serious EA attention. Toby Ord provides a helpful summary chart containing the following probability estimates.[4]

Existential Catastrophe via	*Chance within next 100 years*
Asteroid or comet impact	~ 1 in 1,000,000
Supervolcanic eruption	~ 1 in 10,000
Stellar explosion	~ 1 in 1,000,000,000
Total natural Risk	**~ 1 in 10,000**
Nuclear war	~ 1 in 1,000
Climate Change	~ 1 in 1,000
Other environmental damage	~ 1 in 1,000
"Naturally" arising pandemics	~ 1 in 10,000
Engineered pandemics	~ 1 in 30
Unaligned artificial intelligence	~ 1 in 10
Unforeseen anthropogenic risks	~ 1 in 30
Other anthropogenic risks	~ 1 in 50
Total anthropogenic risks	**~ 1 in 6**
Total existential risk	**~ 1 in 6**

[4] This chart appears in *The Precipice* (p. 167).

While these are Ord's estimates, he developed them in consultation with many of the EA movement's leading figures. Notice that some existential risks widely regarded as serious—even outside EA circles—receive quite low probabilities. Ord assigns both nuclear war and catastrophic climate change a probability of roughly 1 in 1,000 (0.001) in the next 100 years. This provides a useful benchmark. If religious catastrophe has a probability somewhere in that general range, then by EA standards it would seem to deserve serious attention and resources. That sharpens the central question of this section: is religious catastrophe's probability plausibly in the neighborhood of 1 in 1,000? I will argue that it is.

We could proceed by painstakingly surveying the first-order evidence for and against Christianity, Islam, and other religious hypotheses, and then use that evidence to assign probabilities to orthodox Christianity and Islam. There's no space for that here. Fortunately, we don't need it. Since the present argument requires only rough probability estimates, we can instead appeal to higher-order evidence: evidence about what the first-order evidence supports.

Here is a quick argument that, for most people in the modern West—especially the analytic philosophers reading this paper—Christianity or Islam should receive a non-negligible credence, plausibly somewhere in the vicinity of one in a thousand. It appeals to the testimony of both ordinary people and professional philosophers. Start with ordinary believers. Roughly 57% of the world's population adheres to a religion with heaven-and-hell stakes—about 33% to Christianity[5] and 24% to Islam.[6] Millions also report religious experiences associated with these faiths. That, by itself, suggests that belief in a heaven-and-hell religion has *something* going for it. Of course, this is a very low evidential bar. But a very low bar is all we need to clear to establish that religious catastrophe warrants a non-zero, non-infinitesimal probability assignment—that treating religious catastrophe as effectively *impossible* is unwarranted.

[5] https://worldpopulationreview.com/country-rankings/most-christian-countries
[6] https://www.pewresearch.org/fact-tank/2017/08/09/muslims-and-islam-key-findings-in-the-u-s-and-around-the-world/

Still, you might think (even if you'd be reluctant to say it aloud), "Who cares what ordinary people think? They're idiots!" Consider, then, professional philosophers: among the most educated, reflective, and skeptical people on the planet. According to the 2020 PhilPapers survey, 18.83% of philosophers accept or lean toward theism (14.25% accept theism; 4.58% lean toward it), while another 7.21% identify as agnostic.[7] Thus, over a fourth (26.05%) of philosophers surveyed either accept theism or suspend judgment about it. In other words, 26.05% had a credence in the vicinity of 0.5 or greater—*not at all what you'd expect if the evidential probability of Christianity or Islam were zero or infinitesimal*. (After all, these are people who have been professionally trained to evaluate arguments carefully, scrutinize evidence, and avoid obvious fallacies.) If we play it safe and suppose that only a third of those philosophers who believe in theism also believe in heaven and hell, that's about 6%. (The true figure is almost certainly higher, but again, the evidential bar my argument needs to clear is so low I can afford to be absurdly concessive.) Thus, on a very conservative estimate, about 6% of the most educated, reflective, and skeptical people on the planet believe in heaven and hell. Combined with the fact that more than half the global population adheres to Christianity or Islam, this strongly suggests that the evidential probability of religious catastrophe lies at least somewhere in the vicinity of 0.001—not the sort of probability a rational expected-value maximizer can ignore.

But we're not done. The case grows even stronger when we narrow our focus to philosophers of religion. According to the 2020 PhilPapers survey, 77.77% of respondents specializing in Philosophy of Religion accept theism.[8] In other words, among the highly educated, reflective, and generally skeptical people most familiar with the arguments for and against God's existence, more than three-fourths believe in God. Of course, there is an obvious selection effect here. People drawn to philosophy of religion are more likely than average to regard religious questions as intellectually

[7] https://survey2020.philpeople.org/survey/results/4842
[8] https://survey2020.philpeople.org/survey/results/correlations

serious and genuinely open rather than obviously settled in favor of atheism. We should therefore expect theists and agnostics to gravitate toward philosophy of religion at higher rates than atheists—just as we'd expect Kantians and utilitarians to gravitate toward ethics more than moral nihilists. Still, the point remains significant. The philosophical case for atheism is plainly not so decisive that every professionally trained philosopher who studies the issue carefully comes away convinced that atheism is true. Quite the opposite: among those most acquainted with the arguments against God's existence, a large majority believe in God. That fact alone makes it difficult to justify assigning theism—and, by extension, religious catastrophe—a zero or infinitesimal evidential probability.

So far, I've done little more than cite the opinions of ordinary people and philosophers. But neither evidential probabilities nor philosophical truths are determined (in any direct way) by polling data. The significance of such data emerges only once we consider the epistemology of disagreement. Views in this literature fall along a spectrum between so-called *conciliationist* and *steadfast* positions.[9] Conciliationists argue that when you encounter disagreement from an epistemic peer—someone roughly as informed, intelligent, and intellectually conscientious as you—you rationally ought to suspend judgment or substantially reduce your confidence in the disputed proposition. Steadfasters, by contrast, argue that it is at least sometimes rationally permissible for you to retain your original belief despite the disagreement. Crucially, however, nearly everyone in the debate agrees on one central point: disagreement from epistemic peers normally provides *some* reason to reduce your confidence in the disputed proposition. (A notable exception was Tom Kelly's 2005 "right reasons" view, which he later abandoned.)

Although the literature focuses primarily on *peer* disagreement, most philosophers hold that disagreement with epistemic superiors and inferiors has epistemic significance, too. In general, you

[9] Canonical statements of conciliationist views appear in Christensen (2007) and Elga (2007). Canonical statements of steadfast views appear in Kelly (2005) and Kelly (2010).

should place more weight on the judgments of epistemic superiors than on your own and less weight on the judgments of epistemic inferiors. Suppose, for example, that you're a first-year biology student interning at a pharmaceutical company. Your supervisor asks you to review data from a randomized controlled trial and determine whether a drug meets a relevant safety threshold. You examine the data and conclude that it does. Your supervisor tells you that it doesn't. In this context, he is plainly your epistemic superior, so you should defer almost entirely to his judgment. But now consider a different case in which you lead a team of one hundred scientists investigating a drug's safety. You have five more years of experience than any of them and an IQ ten points higher. After reviewing the data, you conclude that the drug is safe. The other ninety-nine scientists—all your epistemic inferiors—disagree. Many philosophers think it's obvious that, even if you're rationally permitted to retain your *belief* that the drug is safe, you're still rationally required to reduce your *confidence* in it. If so, then even disagreement from epistemic inferiors carries evidential weight.

Here, then, is the application to religious catastrophe. Most secular EAs disagree not only with ordinary religious believers but also with many epistemic peers and superiors about whether religious catastrophe is a genuine threat. Their epistemic superiors include, for example, theistic philosophers of religion who are often better trained in philosophy and far better informed about the arguments for and against Christianity and Islam. Their peers include many intelligent, informed, and philosophically sophisticated religious thinkers. And, as noted above, even disagreement from epistemic inferiors—such as ordinary believers around the world—still carries at least some evidential weight. If the epistemology of disagreement literature is even approximately correct, all of this should substantially increase secular EAs' credence in religious catastrophe relative to whatever credence they would assign based solely on their own evaluation of the arguments. Importantly, this conclusion is perfectly compatible with assigning a fairly low overall probability to religious catastrophe. My claim is not that EAs should regard religious catastrophe as likely. It's that they are not rationally entitled to

treat it as effectively *impossible*. Even if religious catastrophe warrants only a probability comparable to that of nuclear war (1 in 1,000) or a supervolcano eruption (1 in 10,000) over the next century, that would still not be low enough—by standard EA reasoning—to justify ignoring the threat altogether.

But perhaps you think I have overlooked some overwhelmingly powerful reason to rule out religious catastrophe altogether. If so, what could it be? Presumably, the only candidates are philosophical arguments. The arguments from evil, divine hiddenness, religious diversity, and related atheological considerations are undeniably formidable. But they are opposed by an equally familiar body of theistic arguments: cosmological, fine-tuning, ontological, and moral arguments, among others. Are the atheological arguments so decisive that they justify assigning a zero—or near-zero—credence to every religion with heaven-and-hell stakes? Not even close. To think otherwise, one would have to believe that one sees the matter more clearly than the many theistic and agnostic philosophers of religion—specialists!—who assign non-zero credence to theism, as well as the overwhelming majority of ordinary religious believers. One would have to regard claims like "orthodox Christianity is false" or "orthodox Islam is false" as epistemically comparable to simple arithmetic truths such as "5 + 3 = 8." Finally, one would have to think that atheological arguments occupy a uniquely privileged position within philosophy. In almost no other area do philosophers take philosophical arguments alone as sufficient to justify near-certainty that a substantive view—especially one widely held by both philosophers and ordinary people—is false. Thus, assigning a zero or infinitesimal probability to religious catastrophe cannot be justified—it's positively irrational.

Again, none of this is meant to show that religious catastrophe is imminent or likely. For all I have argued, the probability is quite low—perhaps somewhere around one in one thousand. But that is all my argument requires.

4. Can Anything be Done about Religious Catastrophe?

I have argued that religious catastrophe is at least as bad, at least as probable, and therefore at least as important as the standard catastrophic risks EAs emphasize. But nothing yet follows about what we should do about it. If there is nothing we can do to reduce the risk of religious catastrophe, then we almost certainly have no obligation to try. Ought implies can. The crucial question, then, is this: can anything be done to reduce the risk of religious catastrophe?

This one's easy: yes. On both Christianity and Islam, people can avoid religious catastrophe by embracing the correct religion. And this is plainly possible. Conversions to these religions occur all the time. Most of us likely know someone who has undergone one. Of course, few people are especially skilled at bringing about such conversions in others. Doing so requires a rare combination of religious knowledge, rhetorical ability, and persuasive power. But some people possess those abilities, and the rest of us can support them financially. Christianity has missionaries and evangelists. Islam has practitioners of *dawah*. These individuals often devote years to learning how to communicate and defend their religious traditions in ways designed to persuade others. So, one obvious strategy for reducing the risk of religious catastrophe would be to fund effective missionaries or to research the most effective methods for producing religious conversion. Importantly, one need not personally believe in Christianity or Islam to support such efforts—just as one need not believe that nuclear war or catastrophic climate change is imminent to support efforts aimed at reducing those risks. Even someone who does not believe, and cannot bring themselves to believe, in Christianity can fund people capable of persuading others to adopt it. The same is true for Islam. The crucial point, then, is this: we are not powerless in the face of religious catastrophe. There are concrete actions we can take to reduce the risk—at least for vast numbers of people. Secular EAs therefore cannot justify ignoring religious catastrophe by claiming that we're powerless to stop it.

5. A Dilemma for Secular Longtermist Effective Altruists

If religious catastrophe is at least as bad, at least as probable, and therefore at least as important as the catastrophic risks EAs emphasize, and we can do something about it, then longtermist EAs face the following dilemma:

> **The Dilemma:** Either adopt religious catastrophe as an EA cause, or ignore it, but also ignore all EA causes with a comparable or lower expected value (i.e., most or all of them). Either way, business as usual—ignoring religious catastrophe while championing the usual EA causes—is incompatible with secular longtermist principles.

I suspect most EAs will find the first horn of the dilemma unattractive because it would require taking religion and religious proselytization seriously. Anyone familiar with the EA movement and its culture will know that, while it is not explicitly anti-religious, it has not given religious values and concerns a prominent place. Its intellectual stars are secular academics—figures such as Peter Singer, Will MacAskill, and Toby Ord—and religiosity has been unfashionable within academia for at least a century. Among many academics, religion is viewed as irrational, harmful, silly, embarrassing, or some combination thereof. I suspect this partly explains the relative silence surrounding religious concerns within the EA movement. Placing religious catastrophe among the standard EA cause areas thus threatens to embarrass both the movement and many of its members.

The second horn of the dilemma—ignoring religious catastrophe while also abandoning catastrophic risks with comparable or lower expected value—is unattractive for obvious reasons. EAs have devoted enormous attention and resources to causes such as unaligned AI, nuclear war, pandemics, supervolcanic eruptions, and catastrophic climate change. To abandon those causes merely to avoid taking religious catastrophe seriously would itself be embarrassing, since it would be obvious that the retreat was motivated solely by a desire to avoid engaging with religion. And that would seem deeply unprincipled—precisely the kind of inconsistency the EA movement takes pride in rejecting.

Of course, there is always a third option: embrace inconsistency. EAs could concede that the expected value of addressing religious catastrophe is at least as high as that of addressing other catastrophic risks, yet simply refuse to apply their principles consistently. But this option is especially unattractive because it abandons a central EA value: following the argument wherever it leads. Effective Altruists have long prided themselves on supporting the most effective causes regardless of how strange, unpopular, or unglamorous they initially appear. Concern about unaligned AI, for example, was once widely dismissed as science-fiction nonsense. Many still view it that way. Yet EAs never backed down in the face of ridicule. Ignoring religious catastrophe merely because it is culturally unfashionable would therefore betray one of the movement's central commitments.

My own view is that the least costly option for secular EAs is to embrace the first horn of the dilemma and adopt religious catastrophe as another catastrophic risk worthy of research, funding, and public advocacy. But I leave that judgment to them. My aim here is only to show that a choice must be made.

6. Pascal's Wager and Religious Catastrophe

Many will notice the resemblance between my argument and Pascal's Wager. No surprise: EAs wholeheartedly embrace Pascalian—that is, expected-value—reasoning, and my argument is simply a consistent application of that reasoning. Pascal argued that, although the evidence for Christianity is inconclusive, the expected benefits of believing and practicing Christianity are infinitely positive, while the expected consequences of failing to do so are infinitely negative (or at least not infinitely positive). On that basis, he claimed that it is practically rational even for nonbelievers to take steps to bring themselves to believe.

Critics of Pascal's Wager have raised countless objections over the centuries, and you might think many of them apply to my argument as well. But they don't, for two reasons. First, many

objections to the Wager target its recommendation that one believe—or get oneself to believe—religious claims for prudential reasons. Such belief is often said to be immoral, psychologically impossible, or doomed to fail because God would reject this kind of self-interested, "for-profit" faith. But my argument does not claim that anyone ought to believe, or get oneself to believe, any religious proposition. I am simply applying standard EA expected-value reasoning about catastrophic risk to the case of religious catastrophe. It is certainly true that anyone attempting to mitigate religious catastrophe will need to support activities that lead others to adopt certain religious beliefs. But presumably the people secular EAs would support in such efforts—missionaries, evangelists, practitioners of *dawah*—already believe the relevant religious claims themselves. So, objections to Pascal's Wager that target belief formation do not apply here.

Second, because EA reasoning is fundamentally Pascalian—that is, grounded in expected-value calculations—many objections raised against Pascal's Wager apply just as readily to standard EA arguments as they do to mine. So, if those objections defeat the argument I have advanced here (and I will argue that they don't), then they would equally undermine familiar EA arguments for prioritizing catastrophic-risk mitigation.

There are, however, two objections to Pascal's Wager worth addressing. I turn to those now.

6.1 Catastrophic Risk and the Many-gods Objection

I've argued that, when consistently applied, EA principles commit EAs to supporting efforts to convert people to religion to reduce their risk of religious catastrophe. This raises a crucial question: *which* religion should EAs promote? At first glance, there seems to be no rational basis for choosing. If each religion offers an infinitely positive payoff, then converting to any of them appears to have infinitely positive expected value. Thus, converting someone to Christianity has the same expected value as converting them to Islam, which has the same expected value as converting them to some

variant of Old Norse religion, according to which Odin grants eternal bliss to those who worship him and eternal torment to those who don't. Now the objection gets its bite. If my argument implies that EAs should adopt religious catastrophe as an official cause and that promoting Odin worship would be just as reasonable as promoting Christianity or Islam, then something has clearly gone wrong with the argument. This is the catastrophic-risk analogue of the familiar "many-gods objection" to Pascal's Wager.

While it's true that the expected value of causing someone to convert to each religion with heaven-and-hell stakes is the same, it does not follow that we have no rational basis for choosing which religion(s) to promote. Suppose I offer you one of two lottery tickets and you're certain the offer is genuine:

Ticket 1: a one-in-ten-thousand chance of infinite bliss

Ticket 2: a one-in-five chance of infinite bliss

Because both outcomes are infinitely positive, the expected value of choosing either ticket is infinitely positive. But ask yourself: Are you indifferent between them? Obviously not. Ticket 2 is plainly the only rational choice. The lesson: when multiple options offer infinitely positive payoffs, you should choose the option with the highest probability.[10] If a decision theory tells you to be indifferent between Tickets 1 and 2 merely because they have the same expected value, you should ditch that decision theory. Nothing could be more obvious than that you ought to choose Ticket 2, and any plausible decision theory will vindicate that thought.

We can answer the many-gods objection to religious catastrophe mitigation in the same way. EAs should promote whichever heaven-and-hell stakes religion they justifiably regard as most probable. For some EAs, that's Christianity. For others, it's Islam. For still others, maybe something else. Each religion's evidential probability will differ from person to person because different people

[10] This response is inspired by Jackson and Rogers (2019) and Rota (2017).

possess different evidence, and evidential probability is determined by evidence. Nevertheless, there is a straightforward method for choosing among multiple options with infinitely positive payoffs: choose the most probable option. So, if the question is, "Which religion should EAs promote?" the answer is: the most probable one.

Notice that this is how EAs *already* think about catastrophic risk. How do EAs decide which risks to prioritize when many of them could wipe out all conscious life and eliminate all future value? Answer: They prioritize by evidential probability (with some consideration for tractability). This is why unaligned AI has loomed so large in EA circles in recent years: EAs widely regard it as the most likely catastrophic risk. The same applies to religious catastrophe: prioritize which religions to promote by evidential probability.

Objection: If we convert someone to Christianity when Islam is true, we'll do infinite harm! Likewise, if we convert someone to Islam when Christianity is true. So, it's best to avoid attempting to convert anyone at all.

Reply: Prioritize converting atheists, Scientologists, and adherents of other worldviews whose probability of escaping religious catastrophe is far lower than that associated with the religion to which they may be converted. Doing so substantially reduces the risk of converting someone away from the correct religion, thereby making the expected value of attempted conversion positive.

Note that the reply to the many-gods objection applies here, too. If Islam has the highest evidential probability, then converting someone from Christianity to Islam is, on balance, more likely to help than to harm them. More generally, converting someone from a lower-probability religion to a higher-probability religion still makes sense. Given your evidence, leaving someone in a lower-probability religion when you could move them to a higher-probability one is riskier than the possibility of converting them incorrectly. That follows directly from the fact that the religion to which you are converting them is, by hypothesis, more likely to be true than the religion they currently accept.

6.2 Catastrophic Risk and Pascal's Mugging

One might worry that my argument illicitly holds secular EAs hostage to causes involving infinite or unimaginably large (dis)utilities. This concern is captured by Bostrom's (2009) "Pascal's Mugging" thought experiment. A stranger approaches you: "Give me your wallet today, and tomorrow I'll return it with one hundred quadrillion utils." You're virtually certain they're lying, but there's a tiny chance they're telling the truth. If the promised payoff is sufficiently large, expected-value reasoning says you're rationally required to hand over your wallet, despite your near-certainty you'll never see it again. Something has gone wrong. Perhaps the lesson is that cases involving vanishingly small probabilities and astronomically large (dis)utilities just don't behave like ordinary expected-utility cases. So, perhaps we can rationally ignore them, including the prospect of religious catastrophe.

Perhaps. But no matter how one resolves Pascal's Mugging, it does not rescue secular EAs from the force of The Dilemma. If Pascal's Mugging licenses ignoring religious catastrophe because of its tiny probability and extreme stakes, it equally licenses ignoring EA's own catastrophic risks, which share the same structure. If, on the other hand, Pascal's Mugging shows we really are hostage to such possibilities, then we're hostage to all of them: EA risks and religious catastrophe alike. Either way, Pascal's Mugging cannot justify EAs in championing their preferred causes while ignoring religious catastrophe.

To see this, consider the two natural responses to Pascal's Mugging. The first is to pay the mugger. After all, that is the straightforward implication of expected-value reasoning. Applied to religious catastrophe, this response implies that EAs ought to adopt religious catastrophe as an official cause area and devote resources to mitigating it, just as they do with other catastrophic risks.

The second response is to ignore the mugger, perhaps on the grounds that vanishingly small probabilities can be rationally disregarded. On this view, once a probability falls below some threshold, it may be treated as effectively zero. In Pascal's Mugging, this means treating the probability of the

mugger's return with the promised reward as zero. Applied to catastrophic risk, it means treating the probability of religious catastrophe as zero.

This response fails for two reasons. First, it assumes that there exists some non-arbitrary probability threshold below which outcomes may be rationally ignored. But even many EAs have questioned whether such a threshold can be justified, since introducing one would generate serious paradoxes for one's decision theory (Isaacs 2016; Beckstead and Thomas 2021). Second, even if such a threshold could be justified, it would need to be lower than the probabilities assigned to standard EA catastrophic risks. Otherwise, one would justify ignoring religious catastrophe only at the cost of justifying indifference toward most (or all) other EA catastrophes as well. Consider, for example, Toby Ord's estimate that the probability of a stellar explosion causing human extinction is roughly one in a billion. He nevertheless treats the risk as worthy of serious concern and resources. But once we see how low the probability of religious catastrophe would need to be in order to ignore it while preserving concern for standard EA risks—roughly one in a billion or less—it becomes highly implausible that religious catastrophe falls below that threshold. As argued earlier, the probability of religious catastrophe may indeed be low. But if it were truly one in a billion or lower, we would not expect so many ordinary people, philosophers, and philosophers of religion (specialists about the relevant questions) to remain theists or agnostics.

One might instead raise the threshold of probabilities that can rationally be ignored, setting it at, say, one in ten thousand. But this move comes with substantial burdens. First, one would need to justify that particular threshold as the point separating catastrophes that may be rationally ignored from those that may not. (Why place the threshold *there* rather than anywhere else?) Second, one would need to argue that religious catastrophe falls on the "rationally ignorable" side of the line. Third, one would need to accept the abandonment of any EA risk causes that also fall below that threshold. On Ord's estimates, that would include naturally arising pandemics, supervolcanic eruptions, asteroid or

comet impacts, and stellar explosions. Perhaps such a threshold could be defended, and perhaps one could show that religious catastrophe falls below it. But doing so requires substantial argumentation that no EA has provided. And if it ever is provided, EAs would still need to decide whether they're really willing to abandon all these causes just to keep religious catastrophe off their agenda.

7. Conclusion

Nuclear war, pandemics, unaligned AI, and catastrophic climate change leading to human extinction would all be extraordinarily bad. Although each is unlikely, it's worth taking steps now to mitigate their risk. Eternal damnation for one person, let alone billions, would be extraordinarily bad. If the threat is sufficiently probable, it's worth taking steps now to mitigate the risk. I've argued that it is sufficiently probable—*at least* in the neighborhood of one in one thousand. If so, the expected value of mitigating religious catastrophe is at least as high as that of the standard EA catastrophic risks. Secular longtermists therefore face a choice: adopt religious catastrophe as an EA cause, or don't, but then drop most of the other EA catastrophic risk causes, too. Either way, business as usual—ignoring religious catastrophe while championing the standard EA causes—is incompatible with secular longtermist principles.

References

Beckstead, N. and Thomas, T. (2021). "A Paradox for Tiny Probabilities and Enormous Values". *Global Priorities Institute Working Paper* No. 7: 1-39.

Bostrom, N. (2009). "Pascal's Mugging." *Analysis*. 69 (3): 443-445.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Christensen, D. (2007). "Epistemology of Disagreement: The Good News." *Philosophical Review*, 116 (2):187-217.

Elga, A. (2007). "Reflection and Disagreement." *Nous*, 41 (3):478–502.

Isaacs, Y. (2016). "Probabilities Cannot Be Rationally Neglected". *Mind* 125 (499):759-762.

Jackson, L. and Rogers, A. (2019). "Salvaging Pascal's Wager." *Philosophia Christi* 21 (1): 59-84.

Kelly, T. (2005). "The Epistemic Significance of Disagreement", in John Hawthorne & Tamar Gendler (eds.), *Oxford Studies in Epistemology*, Volume 1.

Kelly, T. (2010). "Peer Disagreement and Higher Order Evidence", in Alvin I. Goldman & Dennis Whitcomb (eds.), *Social Epistemology: Essential Readings*. Oxford University Press.

Ord, T. (2020). *The Precipice*. New York: Hachette Books.

Rota, M. (2016). *Taking Pascal's Wager: Faith, Evidence, and the Abundant Life*. Downer's Grove: IVP Academic.

Rota, M. (2017). "Pascal's Wager". *Philosophy Compass*. 12 (4): 1-11.

Singer, P. (1972). "Famine, Affluence, and Morality." *Philosophy and Public Affairs* 1: 29-43.

Singer, P. (1975). *Animal Liberation.* New York: Avon Books.