Disaster Prevention and the Possibility of Hell: A Dilemma for Longtermist Effective Altruists

Abstract: Longtermist Effective Altruists (EAs) are concerned about existential risk. In this paper, I have three goals. First, I identify a catastrophic risk that has been completely ignored by EAs. I call it *religious catastrophe*: the threat that (as Christians and Muslims have warned for centuries) billions of people stand in danger of going to hell for all eternity. Second, I argue that, even by secular EA lights, religious catastrophe is *at least* as bad and *at least* as probable, and therefore *at least as important*, as the standard EA catastrophic risks. Third, I present the following dilemma for secular EAs: Either adopt religious catastrophe as an EA cause, or ignore religious catastrophe but also ignore catastrophic risks whose mitigation has a similar, or lower, expected value (i.e., most, or all, of them). Business as usual—ignoring religious catastrophe while championing the usual EA causes—is not an option consistent with longtermist EA principles.

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

Word Count: 7,551

Who would fardels bear,
To grunt and sweat under a weary life,
But that the dread of something after death,
The undiscovere'd country, from whose bourn
No traveller returns, puzzles the will,
And makes us rather bear those ills we have
Than fly to others that we know not of?

Hamlet (Act3, Scene 1)

1. Introduction

According to effectivealtruism.org, "Effective Altruism is a philosophy and community focused on maximising the good you can do through your career, projects, and donations."[1] In the past, effective altruists (EAs) have emphasized immediate concerns such as alleviating global poverty and eliminating factory farming.[2] EAs continue to share these concerns, but their emphases have shifted a bit. Recently, EAs have turned their attention to the distant future. After all, the potential for realizing value over such a vast stretch of time, though less certain, is unimaginably greater than the more certain gains in value to

[1] https://www.effectivealtruism.org/
[2] Singer (1972) and Singer (1975) have been especially influential to EAs.

be achieved by solving immediate problems. In other words, EAs have noticed that the *expected* value of influencing the distant future greatly exceeds the expected value of achieving the sure-bet good in the here-and-now. EAs call this new outlook *longtermism*.

More and more longtermists are converging on the view that the way to do the most (expected) good is to mitigate *existential risks*—catastrophic events that lead to human extinction. These risks can be divided into two categories. The first is anthropocentric (i.e., human-made) catastrophes, such as nuclear war, engineered pandemics, killer AI,[3] and catastrophic climate change. The second is natural catastrophes such as super-volcanoes, natural pandemics, exploding nearby stars, and asteroids colliding with Earth. Toby Ord, one of the EA movement's founders, has recently published *The Precipice: Existential Risk and the Future of Humanity* (2020)—a nearly five-hundred-page book cataloguing and assessing the probability of every existential catastrophe (that he thinks is) currently known to humans. Ord argues that existential risk is *the* most important moral issue of our time because so much value hangs in the balance. If we destroy ourselves, all future value is lost. If we preserve ourselves, unimaginable value lies in our future.

To address this important issue, EAs have sponsored a host of projects such as research centers, charitable organizations, and podcasts. Here's just a sampling of those projects: The Future of Humanity Institute, The Centre for Long-Term Resilience, Global Priorities Institute, Giving What We Can, 80,000 hours, Centre for Effective Altruism, Centre for the Study of Existential Risk, Future of Life Institute, and Open Philanthropy. Running and sustaining these projects is expensive (in time, talent, and money). When you consider that each existential risk is unlikely to occur and that, at most, only one of them will wipe out humanity, you can't help but wonder: Is all this investment worth it? Longtermists enthusiastically answer: Yes! They happily admit that each catastrophic risk is unlikely and that most of their work will be wasted, since, at most, one existential risk will occur. (Ord estimates that the probability

[3] Bostrom (2014) is the canonical text outlining the danger to humans posed by artificial intelligence.

that *any* existential risk will be realized in the next one hundred years is one in six—16.66%.[4]) But EAs argue that the high cost of these projects is well worth paying because the *expected* value is huge.

I have three goals in this paper. The first is to identify a catastrophic risk that has been completely ignored by EAs. I call it *religious catastrophe*. The threat: as Christians and Muslims have warned us for centuries, millions of people stand in danger of going to hell for all eternity if they do not accept (or live in accord with) the one true religion. My second goal is to argue that religious catastrophe, though perhaps improbable, is *at least* as probable and *at least* as bad, and therefore *at least* as important, as the standard catastrophic risks about which EAs worry. Third, I present the following dilemma for secular EAs: *Either* take on religious catastrophe as an EA cause, *or* ignore religious catastrophe but also ignore catastrophic risks whose mitigation has a similar, or lower, expected value (i.e., most, or all, of them). What is not available (at least, if EAs wish to be consistent with their beliefs and values) is business as usual: ignoring religious catastrophe while continuing to sound the alarm about the standard catastrophic risks.

Readers will no doubt see the similarity of my argument with Pascal's Wager. They might therefore think that the standard objections to Pascal's Wager apply to my argument, too. If so, then EAs could coherently prioritize the standard catastrophic risks over religious catastrophe. But most of the standard objections to Pascal's Wager do not apply to my argument. For instance, I can completely side-step many objections to Pascal's Wager (e.g., moral objections to believing in God because of the goodies to be won, the impossibility of believing in God at will, the ineffectiveness of believing in God when God knows it's merely for the goodies) because my argument does not suggest that anyone ought *themselves* to believe, or get themselves to believe, in God or any religion. I do, however, respond to the catastrophic risk versions of the many-gods objection and Pascal's mugger—standard objections to Pascal's Wager that can be copied-and-pasted from that literature. I argue that neither succeeds in the case of religious catastrophe. If I'm correct, then any plausible EA-friendly principle about how to allocate

[4] Ord (2020: 67)

scarce resources effectively will support devoting resources to religious catastrophe along with the standard catastrophic risks. Again, EA's have a choice: adopt religious catastrophe as an EA cause, or don't, but then drop most (or all) of the other catastrophic risks, too.

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

It is well known that orthodox Christianity and Islam claim that an eternity of bliss awaits those who believe, or live in accord with, (what their adherents regard as) the one true religion. Eternal agony awaits those who don't. There are, of course, universalist versions of Christianity and Islam according to which all will be saved from eternal punishment. But I shall ignore those, since all I need for my argument is that, according to the orthodox versions of Christianity and Islam, not all are saved. For example, here's Jesus in a famous passage from Matthew 25:31-46 (ESV translation, my emphases):

> 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*."

And here are a few passages from the Qur'an (my emphasis):

> "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).[5]

You get the point. In both sacred texts, hell is a place of *eternal* punishment, fire, burning, anguish, and misery. Not good. No doubt, many in both traditions interpret these descriptions as metaphors for some other bad state, such as loneliness, regret, ennui, boredom, self-loathing, and so on—not literal burning. But nothing in my argument depends on these descriptions of punishment being literal. Hell is, I take it, unimaginably bad no matter how we interpret the descriptions of it. The point is that *if* such a state awaits those who do not believe (or do) the correct things, then few things are more morally significant.

In both traditions, a minority of adherents have suggested that a good God would never subject people—even very bad people—to this kind of treatment. They therefore argue that we should interpret these passages creatively to avoid the suggestion that God would do such a thing. I'm inclined to agree that a good God (if there is one) would not subject anyone to such treatment. But I'm not *certain* about that. It's not as though "God wouldn't punish people eternally" has the same epistemic status as "2+2=4" for me. I have some (reasonable, in my view) doubts. Maybe I'm mistaken about the badness of sin or what maximal goodness and justice requires. Maybe, as some theologians have suggested, eternal punishment is unjust for finite infractions, but hell is nonetheless eternal (and just) because the damned *continue* sinning eternally in response to the circumstances of hell (e.g., continue being prideful, resentful, deceitful). For now, all that's necessary is that we can't reasonably be *certain* that this sort of punishment does not await those who fail to believe or act in accord with the correct religion (if there is one).

So how bad would it be if large swathes of humanity end up in hell or something like it? Exceedingly bad, obviously. Indeed, it seems easy to show that eternity in hell for billions (or more) is at least as bad (and probably far worse) than *any* of the standard catastrophic risks about which EAs worry.

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

There are two ways to argue for this. The first is with a "one-shot" sort of argument; the second is with a piecemeal argument.

The one-shot argument that religious catastrophe is at least as bad as the standard EA catastrophic risks simply points out the infinite (or finite-but-ever-increasing) nature of religious catastrophe's badness compared to the finite badness of the EA catastrophic risks' badness. For instance, suppose unaligned AI kills every human being, thereby extinguishing the opportunity for all present and future value for humanity. (We can even suppose that AI performs the killing in an especially cruel and painful way.) Whatever badness this realizes, and whatever goodness it destroys, is finite. A finite number of people will have been killed or prevented from existing. A finite amount of value will have been prevented from being realized (since the universe can only exist for a finite amount of time before it experiences "heat death"). By contrast, when *even one* person is condemned to hell for all eternity, they experience infinite (or finite-but-ever-increasing) suffering and experience infinite (or finite-but-ever-increasing) loss, since they could have had an eternity in heaven. Thus, even one person going to hell for all eternity exceeds the badness of the standard existential risks. If so, then *large swathes* of humanity suffering this misfortune is incalculably worse. And notice that there is nothing special about the existential threat from AI. This finite/infinite disparity holds when comparing religious catastrophe to *any* of the standard EA catastrophic risks. That's the one-shot argument that religious catastrophe is at least as bad as the standard EA catastrophes.

The piecemeal argument that religious catastrophe is at least as bad as the usual EA catastrophes proceeds by listing each EA catastrophic risk and assessing its badness compared to large swathes of humanity going to hell for all eternity. Which is worse: Eternal hell for billions (or more) or catastrophic climate change and extinction for billions (or more)? Which is worse: eternal hell for billions (or more) or nuclear war and extinction for billions (or more)? Which is worse: eternal hell for billions (or more) or a large comet slamming into Earth and extinction for billions (or more)? And so on, for each catastrophic risk. It seems clear that eternity in hell for billions, rather than an eternity of bliss, is far worse every time.

But it's good enough for my purposes if religious catastrophe's badness is at least comparable in badness to the EA catastrophic risks.

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

If eternity in hell for billions is at least as bad as the standard catastrophic risks, then the only thing that could justify EAs in ignoring religious catastrophe, given their commitment to expected value reasoning, is that they assign a zero, or infinitesimal, probability to religious catastrophe. If religious catastrophe had such a low probability, it could be safely ignored, despite its unimaginable badness. In this section, I'll argue that, even if the probability of religious catastrophe is low, assigning it a zero, or infinitesimal, probability is irrational. (If you're already convinced of this, you can skip this section.)

Probability is said in many ways. There are subjective probabilities (i.e., the credences an agent *in fact* has in various propositions), evidential probabilities (i.e., the credences an agent *ought* to have, or is *justified* in having, in various propositions), frequentist probabilities (i.e., the frequency of various outcomes in past cases), and objective probabilities (i.e., the *propensity* for an event to happen whether or not it has happened before). We're interested in evidential probability—the degree of confidence an agent ought to have in a proposition, where that degree of confidence can be represented from 0 (justified certainty that the relevant outcome will *not* obtain) to 1 (justified certainty that the relevant outcome will obtain). It's no doubt true that many EAs assign a vanishingly small (or zero) *subjective* probability to religious catastrophe. But I'll argue that they *should not* since that probability assignment is not rationally permitted by the evidence. I'll argue that religious catastrophe's probability is *at least in the vicinity* of the standard catastrophic risks.

Let's begin by considering the probability of the standard catastrophic risks. This will give us a sense of how probable religious catastrophe would need to be to warrant EA concern. Ord provides a helpful chart with the following probability estimates.[6]

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

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**

While these are Ord's own personal estimates, they were reached in consultation with the central figures of the EA movement. Notice that nuclear war and catastrophic climate change are standard existential risks that virtually everyone—even non-EAs—agrees warrant our attention and resources. And yet, Ord assigns them a probability of one in a thousand (0.001). So, I'm assuming that that's roughly the kind of probability threshold religious catastrophe would need to reach for its probability to be comparable to the other catastrophic risks. This helps precisify the question to be investigated in this section: Is the probability of religious catastrophe somewhere in the vicinity of one in one thousand (0.001)? I'll argue that it is.

We *could* proceed surveying all the first-order evidence for and against Christianity, Islam, and perhaps other religions with heaven-and-hell stakes. But we obviously don't have the space for that. Luckily we don't need it. Since we just need rough probability estimates, we can proceed by surveying the higher-order evidence—i.e., the evidence about what the first-order evidence supports. Here's a quick argument that, for most people in the West and most analytic philosophers reading this paper, Christianity or Islam calls for a non-negligible credence—at least in the vicinity of one in a thousand. It's an argument from the testimony (or disagreement) of both ordinary people and professional philosophers.

At least 57% of the humans on this planet believe in a heaven-and-hell stakes religion (33% Christianity,[7] 24% Islam[8]). Millions of them claim to have had religious experiences associated with one of them. That, by itself, suggests that belief in a heaven-and-hell religion has *something* going for it. That is obviously a very low bar. But a very low bar is all we need to clear to get the conclusion that the possibility of hell warrants a non-zero, non-infinitesimal credence.

But, you might think (even if you'd be reluctant to say it) "Who cares what the folk think? They're idiots." Consider, then, professional philosophers. These are among the most educated and skeptical groups of people on the planet. Yet, according to the 2020 PhilPapers survey, 18.83% of philosophers accept or lean toward theism. (14.25% accept theism; 4.58% lean toward it.) And 7.21% were agnostic.[9] So, over a fourth (26.05%) of philosophers (that took the survey) are agnostic or lean toward theism. That means that 26.05% had a credence in the vicinity of 0.5 or greater—not what you'd expect if the evidential probability of Christianity or Islam were zero or infinitesimal. 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 number is almost certainly higher, but we're keeping it conservative here.) Thus, on a very conservative estimate, about 6% of the most educated and skeptical people on the planet believe in heaven and hell. That, combined with the fact that more than half of the people on the planet endorse Christianity or Islam, suggests that the evidential probability of religious catastrophe is *at least in the vicinity* of 0.001—not the kind of probability that can be safely ignored.

But we're not done. According to the 2020 PhilPapers survey, 77.77% of respondents specializing in Philosophy of Religion were theists.[10] So, *among the highly educated and generally skeptical population most acquainted with the arguments for and against God's existence*, *over three fourths believe in God.* No doubt there is a selection effect here: the kind of person who specializes in philosophy of religion is precisely the kind of

[7] https://worldpopulationreview.com/country-rankings/most-christian-countries
[8] https://www.pewresearch.org/fact-tank/2017/08/09/muslims-and-islam-key-findings-in-the-u-s-and-around-the-world/
[9] https://survey2020.philpeople.org/survey/results/4842
[10] https://survey2020.philpeople.org/survey/results/correlations

person who is probably already inclined to think that the questions in philosophy of religion are interesting and live (as opposed to obviously settled in favor of atheism). Thus, we would expect theists and agnostics to gravitate toward philosophy of religion more than atheists (just as we'd expect Kantians or Utilitarians to gravitate toward ethics at a higher rate than moral nihilists). But the point is that the philosophical arguments for atheism are not so compelling that they convince any professionally trained philosopher, who investigates them carefully, that atheism is true. Quite the opposite: over three fourths of those most acquainted with the arguments against God's existence believe in God.

Now, all I've done here is cite the opinions of ordinary people and philosophers. We all know that the philosophical truth isn't determined by a poll. The significance of the polling data, however, can be seen when we consider the epistemology of disagreement. Views on the epistemology of disagreement fall on a continuum between so-called "conciliationist" views and "steadfast" views.[11] Conciliationists argue that, in the face of disagreement from epistemic peers—people roughly equally informed, intelligent, and epistemically virtuous as you—you ought to suspend judgment about (or significantly reduce your credence in) the disputed proposition. Steadfasters, by contrast, argue that it's at least sometimes permissible to maintain your belief in the face of disagreement with an epistemic peer. But, crucially, virtually everyone in the debate agrees that at least *some* reduction in *credence* is called for when you and an epistemic peer disagree about some proposition. (The one exception is Tom Kelly's (2005) "right reasons" view which he has since abandoned.) And while *peer* disagreement has dominated the literature, most think that the disagreement of epistemic superiors and inferiors has evidential value, too. The opinions of epistemic superiors should be given more epistemic weight than one's own and the opinions of epistemic inferiors should be given less. For instance, if your father is a mechanic and he tells you that your radiator is busted, and you don't even know what a radiator is, your dad is your epistemic superior. You should more or less defer to him entirely. But suppose you lead a team of one hundred

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

scientists investigating the safety of a drug. You have at least five years more experience than all of them and an IQ at least ten points higher. The data from the randomized control trials comes in and you think that the drug has met the criteria for being "safe". The other ninety-nine scientists—all of them your epistemic inferiors—think that you're wrong. Many philosophers think it's obvious that, in a case like this, you ought to *at least* reduce your confidence that your judgment is correct, even if, in the end it's permissible for you to retain your belief. If that's correct, then even the disagreement of epistemic inferiors has *some* evidential value.

Here's the application for our purposes: in the case of religious catastrophe, most secular EAs have many religious epistemic superiors (e.g., theist philosophers of religion), peers (e.g., sharp, well-informed religious philosophers), and inferiors (e.g., the majority of humans on the planet) that disagree with them about whether religious catastrophe is a legitimate threat. All of this is evidence that should boost secular EAs' credence that religious catastrophe is a genuine threat, relative to the probability they assign that threat on the basis of their own assessment of the arguments. (Remember: this is compatible with religious catastrophe still receiving a quite low probability assignment.)

None of these observations about the distribution of opinion among philosophers and ordinary folks is intended as a first-order argument that religious catastrophe is a threat. Instead, it's meant to establish a strong presumption in favor of the following more modest claim: it is not reasonable to assign a zero, or infinitesimal, probability to religious catastrophe. Even if religious catastrophe has a low evidential probability—e.g., roughly the same low probability that we should assign to nuclear war or a supervolcano erupting in the next one hundred years—it does not warrant a credence that would justify our completely ignoring the threat of religious catastrophe.

But perhaps you think I've ignored some extraordinarily powerful reason to rule out religious catastrophe. If so, what could it be? It seems that the only considerations that could justify ruling out, or assigning a vanishingly small probability, to every heaven-and-hell stakes religion are philosophical arguments. We all know that the problems of evil, divine hiddenness, religious diversity, and other

atheological arguments are formidable. But those arguments are opposed by a host of theistic arguments: fine-tuning, cosmological, ontological, and moral arguments among others. Are the atheological arguments *so* powerful, and do they *so* overwhelm the evidential force of theistic arguments, that they justify a credence of zero, or near-zero, in each religion with heaven-and-hell stakes? Almost certainly not. To think otherwise, one would have to think that one knows better than the theistic, agnostic, and (the overwhelming majority of) atheistic philosophers in philosophy of religion that the probability of Christianity or Islam is zero. You'd have to think that "Christianity is false" or "Islam is false" is as certain as simple arithmetic claims (e.g., 5 + 3 = 8). One would also have to think that the atheological arguments are *uniquely* powerful in philosophy. For in almost no other contexts do philosophers think that philosophical arguments alone justify (near) certainty that a widely held (among *both* philosophers and ordinary people) substantive philosophical view is false.

Again, nothing I've argued above is intended to show that religious catastrophe is imminent, or even likely. For all I've said, the probability that some heaven-and-hell stakes religion is correct is quite low—somewhere in the vicinity of one in ten thousand. But that's all I need to get my argument going.

4. Can Anything be Done about Religious Catastrophe?

So far I've argued that religious catastrophe is at least as bad and at least as probable, and therefore at least as important, as the standard catastrophic risks. But nothing yet follows about what we ought *do* about it. If there is nothing we can now do to mitigate the risk of religious catastrophe, then we almost certainly have no obligations to try. "Ought implies can", as they say. If, however, we have means at our disposal to mitigate the risk of religious catastrophe, then it's plausible that we do have an obligation to (try to) mitigate that risk. So, we must ask: can we do anything to mitigate the risk of religious catastrophe?

This one's easy: yes. On Christianity and Islam, the way for a person to avoid suffering religious catastrophe is for them to adopt the beliefs, or practices, (or both) of the relevant religion. And this is plainly possible. Conversions to these religions happen all the time. In fact, we all likely know someone

who has experienced such a religious conversion. Now, fewer of us are ourselves able to effect such a conversion in someone else. Doing so requires a kind of religious knowledge and convincingness that is rare. But we can all *fund* such people—people called "missionaries" or "proselytizers" in the Christian tradition, or those engaging in *dawah* in the Islamic tradition. These people train much of their lives to present the content of their religious views to others with the goal of converting them. So, at least one obvious thing EAs could do to mitigate the risk of religious catastrophe for others is to fund and train missionaries, and to conduct research into the most effective means for effecting religious conversion. Notice that one need not *themselves* believe in Christianity or Islam to mitigate the risk of religious catastrophe for others—just as one need not believe that any of the other catastrophic risk will occur in order to mitigate them. Even if you do not yourself believe (and cannot get yourself to believe) in Christianity, you can fund others who can get others to believe and practice Christianity. Same for Islam. The point for now is that we are not powerless in the face of the threat of religious catastrophe. We can do something about it. So EAs cannot justify ignoring religious catastrophe by appeal to the "ought implies can" principle.

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 standard catastrophic risks, 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 religious catastrophe but also ignore catastrophic risks whose mitigation has a similar, or lower, expected value (i.e., most, or all, of them). What is not available (at least, if EAs wish to be consistent with their beliefs and values) is business as usual: ignoring religious catastrophe, while championing the cause of the standard EA catastrophic risks.

I assume that the first horn—adopting religious catastrophe as an official EA cause—is unattractive to most EAs because it involves taking religion and religious proselytization seriously. Those familiar with the EA movement and its culture will know that, while it is not explicitly anti-religion, it has not given

religious values and concerns a prominent place. The stars of the EA movement are secular academics (e.g., Peter Singer, Will MacAskill, Toby Ord) and being religious in academia has been unfashionable for at least a century now. Among many academics, religion is considered crazy, stupid, evil, or all the above. I suspect that this partly explains the relative silence in the EA movement about religious concerns. Placing religious catastrophe among the standard catastrophic risks thus threatens to be embarrassing to the movement and its members.

The second horn of the dilemma—ignore religious catastrophe but also ignore the catastrophic risks with similar or lower expected values—is unattractive as well for obvious reasons. EAs have long championed and devoted resources to causes such as unaligned AI, nuclear war, pandemics, supervolcanic eruptions, and catastrophic climate change. To stop doing so merely to avoid having to take on the cause of religious catastrophe would also be embarrassing, since it would be transparent that their abandonment of these long-held causes is motivated only by a desire to avoid taking on religious catastrophe. And that seems unprincipled in a way that cuts against the values of the EA movement.

Of course, there is always a third option: embrace inconsistency. EAs could admit that the expected value of mitigating religious catastrophe is at least as high as the expected value of mitigating the other catastrophic risks, but simply refuse to apply their principles consistently. This would *lead to* EAs ignoring religious catastrophe, but it wouldn't *justify* it. This option is unattractive for EAs because it represents an abandonment of a central EA value: following the argument where it leads. Rather than research and fund merely *popular* causes, EAs have (true to their name) routinely researched and funded *effective* causes, regardless of their popularity or flashiness. Mitigating the risk of unaligned AI was, at one time, deeply unpopular—widely regarded as a silly sci-fi possibility. Many still regard it that way. But EAs have never backed down, championing the cause all the same and enduring whatever ridicule they receive. So, ignoring religious catastrophe merely on the grounds that doing so is unfashionable would be a betrayal this central EA value.

My own view is that EAs' least costly option is to embrace the first horn, taking on religious catastrophe as yet another catastrophic risk worth researching, donating to, and sounding the alarm about. But I leave that up to them. For now, I wish only to argue that a choice must be made.

6. Pascal's Wager and Religious Catastrophe

My argument is similar to Pascal's Wager. No surprise: EA's wholeheartedly embrace Pascalian reasoning (i.e., expected value reasoning). Pascal argued that, though the evidence for Christianity is far from conclusive, the expected benefits of believing and practicing Christianity are infinitely positive while the expected consequences of not doing so are infinitely negative (or, at least, not infinitely positive). This makes it rational, he argued, even for non-believers to take steps to *get themselves* to believe in Christianity. Pascal's Wager is subject to a host of objections. You may therefore think that those objections from the philosophy of religion literature can be transferred seamlessly to apply to my argument. This is false for two reasons. First, many objections to Pascal's Wager object to the Wager's recommendation that *you* believe, or try to get yourself to believe, some religious claim. And believing some religious proposition because of the goodies you'll receive is either immoral, or impossible, or destined to fail because God would never accept that kind of for-profit belief. But I'm not arguing that anyone ought *themselves* to believe, or try to get themselves to believe, anything—including any religious claims. I'm taking the standard EA arguments for the importance of mitigating catastrophic risks, such as nuclear war and unaligned AI, and applying them to religious catastrophe. So, these worries don't apply to my argument. It's true that those working to mitigate the risk of religious catastrophe will have to get *other* people to believe some religious claims. But presumably the people EAs support to convert people to some religion or other (i.e., missionaries) will already believe in the relevant religion. Second, since EA reasoning *just is* Pascalian (expected value) reasoning, many objections from the Pascal's Wager literature to my argument apply just the same to the standard EA arguments that we should care about catastrophic risks. So, if

objections to Pascal's Wager defeat the argument I've advanced here (and I'll argue that they don't), then they would equally defeat EA arguments for caring about catastrophic risks.

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

6.1 Catastrophic Risk and the Many-gods Objection

I've argued that EA principles, consistently applied, commit EAs to trying to convert people to religion to help them avoid religious catastrophe. Crucial question: to which religion should EAs try to convert people? It seems the expected utility of converting a person to Christianity is the same as the expected utility of converting them to Islam, which is the same as converting them to any other religion with heaven-and-hell stakes. After all, if the expected benefit is infinitely positive, it doesn't matter what the evidential probability of the religion's being true is. So, there seems to be no reason to prefer converting someone to Christianity rather than Islam or rather than, say, a variant of Old Norse religion according to which Odin grants an eternity of bliss to those who commit their life to him and an eternity of agony to those who don't—a transparently ridiculous suggestion. This is the catastrophic risk version of the so-called "many-gods objection" to Pascal's Wager.

It's true: the expected value of converting someone to each religion with heaven-and-hell stakes is the same. But that doesn't mean we have no way of choosing which religion to favor over others. Suppose I offer you one of two lottery tickets (you're certain the offer is genuine):

Ticket 1: provides you a one in ten thousand probability of infinite bliss, or

Ticket 2: provides you a one in five probability of infinite bliss.

The expected value of selecting each ticket is infinite. Which should you choose? Are you indifferent? Are you unsure which to choose? No. The answer is obvious: you should select Ticket 2. The lesson: if you have multiple options, each with infinite expected utility, you should select the option you're (justifiably) most confident will succeed.[12] If a decision theory says that you should be indifferent between

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

Ticket 1 and 2, or that you should choose Ticket 1, then you should ditch that decision theory. Nothing could be more obvious than that you ought to select Ticket 2. Any plausible decision theory will vindicate that thought.

So we can answer the many-gods objection to religious catastrophe mitigation the same way. You should devote resources to converting people to whichever religion with heaven-and-hell stakes you (justifiably) believe to be most probable. For some that's Christianity. For others it's Islam. For others, maybe something else. (The evidential probability of religions can differ between agents because different agents have different evidence. And evidential probabilities depend on what an agent's evidence supports.) But, in any case, there is a clear, easy way to choose between multiple options with infinite expected utility.

Notice that this is no different from what EAs *already* say about the many catastrophic risks we face. How do EAs decide how to prioritize which of the many catastrophic risks when each of them is capable of killing every living person and extinguishing all future value? Answer: they prioritize the catastrophic risks from most- to least-likely to occur (with some consideration for tractability). This explains why unaligned AI has loomed so large in EA circles. It's regarded as the catastrophic risk most likely to occur. The same reasoning applies to religious catastrophe mitigation. Prioritize religions by evidential probability.

Objection: if we convert someone to Christianity, when Islam is the correct religion, we'll do infinite (or enormously large) harm. Similarly, if we convert someone to Islam when Christianity is the correct religion, we'll do more harm than good. So, it's best not to go around converting people to any religion for fear of harming people.

Reply: Prioritize atheists, or scientologists, or adherents of other worldviews whose probability of escaping religious catastrophe is far lower than the religion you're converting them to. Now there's no, or significantly less, threat of converting someone away from the correct religion.

But it's also worth noting that the same reply to the many-gods objection works here too: If Islam has the highest evidential probability, then, when you convert someone from Christianity to Islam, it is more likely that you've just made a conversion for good than that you've made a conversion for harm. So, it still makes sense to try to convert someone from a lower probability religion to a higher probability religion. To leave them with their current religion when you could get them to a higher probability religion is, given your evidence, riskier than the risk of converting them to the wrong religion. This follows straightforwardly form the fact that the converted-to religion is more likely to be true than the converted-from religion.

6.2 Catastrophic Risk and Pascal's Mugger

One might worry that my argument illicitly holds EAs hostage to causes threatening infinite (or unimaginably large) (dis)utilities. This idea has been captured in Bostrom's (2009) Pascal's Mugger thought experiment. Imagine a person confronts you on the street and tells you that if you give them your wallet now, they'll return tomorrow and give you one hundred quadrillion utils. Though you seriously doubt they'll live up to their end of the bargain, there is *some* chance they're telling the truth. And if the promise of benefits is large enough, it will be rational—according to expected value reasoning—to hand over your wallet, even if you're almost certain not to get your money back. The intuitive worry this case illustrates is that cases where (dis)utilities are enormously high and probabilities are unusually low are weird. So maybe they don't work like standard expected utility cases. But I'll argue that, no matter how one chooses to respond to Pascal's Mugger, religious catastrophe remains in the same boat as the standard catastrophic risks. So, one is not justified in treating religious catastrophe differently, or as less important, than the other catastrophic risks.

The first response to Pascal's Mugger is to pay the mugger. After all, that's what straightforward expected value reasoning says you ought to do. That verdict translated to the case of religious catastrophe

means that you ought to worry about religious catastrophe and devote resources to its mitigation—same as the other catastrophic risks.

The second response is to argue that the lesson of Pascal's Mugger is that vanishingly small probabilities can be rationally ignored. According to this response, once the probability of some outcome reaches some threshold of low probability, it can be treated *as if* the probability of that outcome is zero. In the Pascal's Mugger case, this would mean one can proceed as if the probability of the mugger's payoff is zero. In the catastrophic risk context, it means one can proceed as if the probability of religious catastrophe is zero.

There are two responses to this thought. First, this response assumes that some non-arbitrary probability threshold can be identified and that any outcomes below that probability threshold can be rationally ignored. That has been questioned, even by EAs, since it leads to all sorts of paradoxes in one's decision theory (Isaacs 2016; Beckstead and Thomas 2021). Second, even if some non-arbitrary probability threshold can be identified, that threshold would have to be lower than the probabilities of the standard EA catastrophic risks. Otherwise, one could justify ignoring religious catastrophe, but only at the cost of justifying ignoring most, or all, of the other EA catastrophes. For instance, Ord estimates that the probability of a stellar explosion resulting in the death of every human being on Earth is around one in a billion. And yet, he thinks it's worth devoting resources to mitigating the risk of death for all by means of stellar explosion. But once we see just how low the probability of religious catastrophe would have to be to be safely ignored while preserving concern for the standard EA causes (one in a billion or lower), it becomes highly implausible that religious catastrophe is sufficiently improbable. As I argued in section three, the probability of religious catastrophe may be low, but if it were one in a billion or lower, we wouldn't expect so many philosophers to be theists and agnostics. Now, one might raise the threshold of probabilities that can be safely ignored so that any outcome that is, say, one in ten thousand or lower can be safely ignored. But one would have to justify that specific probability threshold as the one that separates the safely-ignored catastrophes from the unsafely-ignored catastrophes *and* argue that religious

catastrophe falls on the "safely-ignored" side *and* be willing to jettison the cause of any catastrophic risks falling below that threshold (i.e., naturally arising pandemics, supervolcanic eruptions, asteroid or comet impacts, stellar explosions). Perhaps one could do it, but it would require quite a bit of argumentation. And that argumentation would have to justify the fine-grained probability assessments necessary to identify the relevant threshold and sort each catastrophic risk on the desirable sides of that threshold. A tall order, to put it mildly.

7. Conclusion

Nuclear war, pandemics, killer AI, and catastrophic climate change leading to the extinction of human beings are really bad. While each is unlikely to occur, it's still worth taking steps *now* to mitigate their risk. Similarly, religious catastrophe—large swathes of humanity being condemned to hell for all eternity—is really bad. If religious catastrophe is sufficiently probable, it's worth trying to prevent. I've argued that it is sufficiently probable. I haven't given a precise probability assessment, but I've argued that the probability is *at least* in the vicinity of one in ten thousand. So EAs have a choice: *either* adopt religious catastrophe as an EA cause, *or* don't, but then drop most of the other catastrophic risks, too. Business as usual—ignoring religious catastrophe while championing the usual EA causes—is not an option consistent with longtermist EA 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.

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: Hatchette 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.