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

Eric Sampson

Effective Altruism: "Effective altruism is the project of using evidence and reason to figure out how to best contribute to helping others, and taking action on that basis." – Giving What We Can's answer to "What is Effective Altruism?"

Inspiration: Peter Singer (inspired by Bentham, Mill, Sidgwick), Will MacAskill, Toby Ord, Nick Bostrom

Organizations: Giving What We Can, Give Directly, GiveWell, 80,000 hours, Centre for Effective Altruism, The Future of Humanity Institute, Global Priorities Institute, The Centre for Long-Term Resilience

Tools: Expected value reasoning, QALYs, cost-benefit analysis

Old Emphasis: Global Poverty, Factory farming

- **Justification:** Tackling these problems does the most (expected) good.

New Emphasis: Catastrophic Risk and the Future of Humanity

- **Catastrophic Risks:** catastrophic climate change, killer AI, nuclear winter, engineered pandemics, super-volcanoes, killer asteroids, killer aliens, scientific experiments gone awry
- **Justification:** Tackling these problems does the most (expected) good *in the long term*.
- **Longtermism:** Positively influencing the long-term future is a key moral priority of our time since we can realize unimaginably more value over that span than we can realize in the near term
- **Objection:** For each of these risks, the probability that they'll occur is tiny! Why worry about them?
- **Reply:** *If* one occurs, it would be a *catastrophe*. So, despite these risks' extraordinarily low probability, we have excellent reason to devote intellectual and monetary resources to mitigating them.
    - This is why we lock our doors, buckle our seatbelts, buy insurance, etc. (And we should.)

3 Goals in the Paper:

1. Identify a "new" catastrophic risk EAs have entirely neglected.
    - **Religious Catastrophe:** Trillions of people (or more) go to hell (or something hell-like) for all eternity for rejecting the one true God or religion.
2. Argue that, even by secular EA lights, religious catastrophe is *at least* as bad, *at least* as probable, and therefore *at least* as important as the standard EA catastrophic risks.
3. Argue that EAs (who want to live consistently with their beliefs and values) face a 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). Business as usual—ignoring religious catastrophe while championing the usual EA causes—is inconsistent with secular longtermist principles.

The Threat: Religious Catastrophe

Here's Jesus:

> "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*." (Matt 25:31-46)

Here's the Quran:

> "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).

Religious Catastrophe's Evaluative and Probabilistic Similarity to Standard Longtermist Causes

Badness: Religious catastrophe is *at least* as bad as the worst catastrophic risks

- **One-Shot Argument:** Finite vs. Infinite
    - Each catastrophic risk is finitely bad since a finite number of people will die and be prevented from enjoying a valuable life. (The universe must die a "heat death" eventually.)
    - The disvalue of a religious catastrophe is either infinite or finite but ever-increasing (because it lasts for eternity). So, its disvalue far exceeds the disvalue of any catastrophic threat.
- **Piecemeal Argument:** Which is worse?
    - Eternal hell for trillions (or more) or catastrophic climate change and extinction?
    - Eternal hell for trillions (or more) or nuclear war and extinction?
    - Eternal hell for trillions (or more) or killer AI and extinction?
    - And so on…

Probability: Religious catastrophe's probability is *at least comparable* to EA's catastrophic risks' probability.

Existential Catastrophe via	*Chance within the 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**

Considerations Bearing on Religious Catastrophe's Evidential Probability

1. Standard arguments from natural theology (e.g., Fine-tuning, Kalam, Contingency, Ontological, Resurrection, Moral Knowledge, Psychophysical Harmony).
2. At least 57% of humans on this planet believe in a religion with heaven-and-hell-type stakes (33% Christianity, 24.1% Islam).
3. Literally millions of people, over the ages, have claimed to have religious experiences associated with these religions.
4. Professional philosophers are among the most educated and skeptical people on the planet. Yet, according to the 2020 PhilPapers survey, 18.83% accept or lean toward theism (too low due to selection effects?). 7.21% were agnostic. If we play it safe and suppose that only a third of the theist philosophers believe in hell, that's about 6%. Thus, (on a very conservative estimate) about 6% of the most skeptical people on the planet believe in hell.
5. 77.77% of respondents specializing in Philosophy of Religion were theists. So, among the population most acquainted with the arguments for and against God's existence, 3 out of 4 believe in God.
6. What could justify assigning a zero (or near-zero) probability to every heaven-and-hell -stakes religion?
    i. Philosophical arguments (e.g., Evil, Hiddenness, Evil God Challenge, Religious diversity)? Are these atheological arguments *that much better* than the theistic ones? Are they evidentially *decisive*?
    ii. Commitment to metaphysical naturalism?
    iii. In what other context do philosophers think philosophical arguments provide justified certainty (or near-certainty) that a widely believed philosophical thesis is false?

Objections to Pascal's Wager I Can Easily Sidestep:

- **Impossible:** I can't voluntarily believe in God—it's literally impossible!
- **Morally Bad:** It's morally wrong to believe in God just for heavenly goodies!
- **Ineffective:** Believing in God for the goodies won't work. He doesn't accept for-profit belief!

Reply-to-All: I'm not suggesting that anyone ought to believe, or get themselves to believe, in God.

Objections Meriting a Response:

"I'm an annihilationist. I don't believe in eternal hell."

- Doesn't matter.
    - If you're not justifiably certain there's no eternal hell, you face this problem.
    - Even on annihilationism, an infinite (or indefinitely large) amount of value is lost for each person who experiences Religious Catastrophe.

"A good God wouldn't send people to eternal hell, so we don't have to worry about it."

- Again: If you're not justifiably certain there's no eternal hell, you face the problem I describe.
- The argument doesn't depend on an Eternal Conscious Torment (ECT) conception of hell
    - Could be: ECT, C.S. Lewis-like view, Eastern Orthodox view, eternal disappointment, etc.
- For each conception of hell, you can't be justifiably *certain* God wouldn't "send" people there.

Many-Gods Objection: Each religion has infinite stakes, so the expected (dis)value of each is equal.

- Suppose I offer you one of two lottery tickets with the same payoff:

  Ticket 1: Provides a 1/10,000 probability of infinite bliss, or
  Ticket 2: Provides a 1/3 probability of infinite bliss.

    - The expected value of selecting each ticket is infinite (therefore, equal). Are you indifferent? No.
    - Lesson: When payoffs are equal, choose the most *probable* option.

- EAs *already* do this with catastrophic risks. They prioritize based on probabilities.
- **Practical Upshot:** Devote resources to religions in proportion to their probabilities. Devote the most resources to the most probable religion, second-most resources to the second-most probable religion…

Pascal's Mugger: I'm being held hostage to infinite (dis)utilities!

- **Bostrom's Lesson:** You can rationally ignore threats with vanishingly small probabilities.
    - Problem: Doesn't apply to Religious Catastrophe. The probability isn't vanishingly small.
    - Problem: If you ignore Religious Catastrophe, you must ignore every EA cause whose mitigation has a similar or lower expected value (e.g., nuclear war, runaway climate change, pandemics, killer AI).
    - Problem: Whatever threshold you select for neglectable probabilities, it must be (1) non-arbitrary, (2) high enough to render Religious Catastrophe neglectable, and (3) low enough not to render other EA causes neglectable.
- **Greaves's Lesson:** You should pay the mugger.
    - Upshot for Us: You should devote resources to mitigating the risk of Religious Catastrophe.
- **General Lesson:** Religious Catastrophe is just as (non-)problematic as the standard EA causes.

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). Business as usual—ignoring religious catastrophe while championing the usual EA causes—is inconsistent with secular longtermist principles.

Conclusion: By secular EA's own lights, they ought to devote resources to mitigating the risk of Religious Catastrophe (e.g., giving money to missionaries to convert people to some religion).