X-Risks Institute
www.xrisksinstitute.com

Who Would Destroy the World?

Phil Torres
Founding Director of the X-Risks Institute

1. **Introduction**

Consider a seemingly simple question: if the means were available, *who exactly would destroy the world*? There is surprisingly little discussion of this question within the nascent field of existential risk studies. But it's an absolutely crucial issue: what sort of *agent* would either intentionally or accidentally cause an existential catastrophe?

The first step forward is to distinguish between two senses of an existential risk. Nick Bostrom originally defined the term as: "One where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential." It follows that there are two distinct scenarios, one endurable and the other terminal, that could realize an existential risk. We can call the former an *extinction risk* and the latter a *stagnation risk*. The importance of this distinction with respect to both advanced technologies and destructive agents has been previously underappreciated.

So, the question asked above is actually two questions in disguise. Let's consider each in turn.

2. **Terror: Extinction Risks**

First, the categories of agents who might intentionally cause an extinction catastrophe are fewer and smaller than one might think. They include:

(1) *Idiosyncratic actors*. These are malicious agents who are motivated by idiosyncratic beliefs and/or desires. There are instances of deranged individuals who have simply wanted to kill as many people as possible and then die, such as some school shooters. Idiosyncratic actors are especially worrisome because this category could have a large number of members (or token agents). Indeed, the psychologist Martha Stout estimates that about 4 percent of the human population suffers from sociopathy, resulting in about 296 million sociopaths. While not all sociopaths are violent, a disproportionate number of criminals and dictators have, or very likely have, had the condition.

(2) *Future ecoterrorists*. As the effects of climate change and biodiversity loss (resulting in the sixth mass extinction) become increasingly conspicuous, and as destructive technologies become more powerful, some terrorism scholars have speculated that ecoterrorists could become a major agential risk in the future. The fact is that the climate is changing and the biosphere is wilting, and human activity is almost entirely responsible. It follows that some radical environmentalists could attempt to use technology to cause human extinction, thereby "solving" the environmental crisis. So, we have some reason to believe that this category could become populated with a growing number of token agents in the coming decades.

(3) *Negative utilitarians*. Those who hold this view believe that the ultimate aim of moral conduct is to minimize misery, or "disutility." Although some negative utilitarians like David Pearce see existential risks as highly undesirable, others would welcome annihilation because it would entail the elimination of suffering. It follows that if a "strong" negative utilitarian had a button in front of her that, if pressed, would cause human extinction (say, without causing pain), she would very likely press it. Indeed, on her view, doing this would be the morally right action. Fortunately, this version of negative utilitarianism is not a position that many non-academics tend to hold, and even among academic philosophers it is not widespread.[1]

(4) *Extraterrestrials*. Perhaps we are not alone in the universe. Even if the probability of life arising on an Earth-analog is low, the vast number of exoplanets suggests that the probability of life arising somewhere may be quite high. If an alien species were advanced enough to traverse the cosmos and reach Earth, it would very likely have the technological means to destroy humanity. As Stephen Hawking once remarked, "If aliens visit us, the outcome would be much as when Columbus landed in America, which didn't turn out well for the Native Americans."

(5) *Superintelligence*. The reason *Homo sapiens* is the dominant species on our planet is ultimately due to our intelligence. It follows that if something were to exceed

our intelligence, our fate would become inextricably bound up with its will. This is worrisome because recent research shows that even *slight* misalignments between our values and those motivating a superintelligence could have existentially catastrophic consequences. But figuring out how to upload human values into a machine poses formidable problems — not to mention the issue of figuring out what our values are in the first place.

Making matters worse, a superintelligence could process information at about 1 million times faster than our brains, meaning that a minute of time for us would equal approximately 2 years in time for the superintelligence. This would immediately give the superintelligence a profound strategic advantage over us. And if it were able to modify its own code, it could potentially bring about an exponential intelligence explosion, resulting in a mind that's many orders of magnitude smarter than any human. Thus, we may have only *one chance* to get everything just right: there's no turning back once an intelligence explosion is ignited.

A superintelligence could cause human extinction for a number of reasons. For example, we might simply be in its way. Few humans worry much if an ant genocide results from building a new house or road. Or the superintelligence could destroy humanity because we happen to be made out of something it could use for other purposes: atoms. Since a superintelligence need not resemble human intelligence in any way — thus, scholars tell us to resist the dual urges of anthropomorphizing and anthropopathizing — it could be motivated by goals that appear to us as utterly irrational, bizarre, or completely inexplicable.

3. **Terror: Stagnation Risks**

Now consider the agents who might intentionally try to bring about a scenario that would result in a stagnation catastrophe. This list subsumes most of the list above in that it includes idiosyncratic actors, future ecoterrorists, and superintelligence, but it probably excludes negative utilitarians, since stagnation (as understood above) would likely induce more suffering than the status quo today. The case of extraterrestrials is unclear, given that we can infer almost nothing about an interstellar civilization except that it would be technologically sophisticated.

For example, an idiosyncratic actor could harbor not a death wish for humanity, but a "destruction wish" for civilization. Thus, she or he could strive to destroy civilization without necessarily causing the annihilation of *Homo sapiens*. Similarly, a future ecoterrorist could hope for humanity to return to the hunter-gatherer lifestyle. This is precisely what motivated Ted Kaczynski: he didn't want everyone to die, but he did want our

technological civilization to crumble. And finally, a superintelligence whose values are misaligned with ours could modify Earth in such a way that our lineage persists, but our prospects for future development are permanently compromised. Other stagnation scenarios could involve the following categories:

(6) *Apocalyptic terrorists*. History is replete with groups that not only believed the world was about to end, but saw themselves as active participants in an apocalyptic narrative that's unfolding in realtime. Many of these groups have been driven by the conviction that "the world must be destroyed to be saved," although some have turned their activism inward and advocated mass suicide.

Interestingly, no notable historical group has combined both the omnicidal and suicidal urges. This is why apocalypticists pose a greater stagnation terror risk than extinction risk: indeed, many see the survival of some believer-community beyond Armageddon as integral to the eschatological beliefs they accept. There are perhaps more than a million radical, active apocalyptic believers in the world today, although emerging environmental, demographic, and societal conditions could cause this number to significantly *increase* in the future, as I've outlined in detail elsewhere (see Section 5 of "Agential Risks: A Comprehensive Introduction").

(7) *States*. Like terrorists motivated by political rather than transcendent goals, states tend to place a high value on their continued survival. It follows that states are unlikely to intentionally cause a human extinction event. But rogue states could induce a stagnation catastrophe. For example, if North Korea were to overcome the world's superpowers through a sudden preemptive attack and implement a one-world government, the result could be an irreversible decline in our quality of life.

So, there are numerous categories of agents that could attempt to bring about an existential catastrophe. And there appear to be fewer agent types who would specifically try to cause human extinction than to merely dismantle civilization.

4. **Error: Extinction and Stagnation Risks**

There are some reasons, though, for thinking that error (rather than terror) could constitute a more significant threat in the future. First, almost every agent capable of causing intentional harm would also be capable of causing accidental harm, whether this results in extinction or stagnation. For example, an apocalyptic cult that wants to bring about Armageddon by releasing a deadly biological agent in a major city could, while preparing for this terrorist act, inadvertently contaminate its environment, leading to a global pandemic.

The same goes for idiosyncratic agents, ecoterrorists, negative utilitarians, states, and perhaps even extraterrestrials. Indeed, the large disease burden of Europeans was a primary reason Native American populations declined. By analogy, perhaps an extraterrestrial will destroy humanity by introducing a new pathogen that quickly wipes us out. The case of superintelligence is unclear, since the relationship between general intelligence and error-proneness has not been adequately studied (although studies do suggest a link).

Second, if powerful future technologies become widely accessible, then virtually everyone could become a potential cause of existential catastrophe, even those with absolutely no inclination toward violence. To illustrate the point, imagine a perfectly peaceful world in which not a single individual has malicious intentions. Further imagine that everyone has access to a doomsday button on her or his phone; if pushed, this button would cause an existential catastrophe. Even under ideal societal conditions (everyone is perfectly "moral"), how long could we expect to survive before someone's finger slips and the doomsday button gets pressed?

Statistically speaking, a world populated by only 1 billion people would almost certainly self-destruct within a 10-year period if the probability of any individual accidentally pressing a doomsday button were a mere 0.00001 percent per decade. Or, alternatively: if only 500 people in the world were to gain access to a doomsday button, and if each of these individuals had a 1 percent chance of accidentally pushing the button per decade, humanity would have a meager 0.6 percent chance of surviving beyond 10 years. Thus, even if the likelihood of mistakes is infinitesimally small, planetary doom will be virtually guaranteed for sufficiently large populations.[2]

5. **The "Two Worlds" Thought Experiment**

The good news is that a focus on *agential risks*, as I've called them, and not just the technological tools that agents might use to cause a catastrophe, suggests *additional ways* to mitigate existential risk. Consider the following thought-experiment: a possible world A contains thousands of advanced weapons that, if in the wrong hands, could cause the population of A to go extinct. In contrast, a possible world B contains only a single advanced "weapon of total destruction" (WTD). Which world is more dangerous? The answer is obviously world A.

But it would be foolishly premature to end the analysis here. Imagine further that A is populated by compassionate, peace-loving individuals, whereas B is overrun by war-

mongering psychopaths. *Now* which world appears more likely to experience an existential catastrophe? The correct answer is, I would argue, world B.

In other words: agents matter as much as, or perhaps even more than, WTDs. One simply can't evaluate the degree of risk in a situation without taking into account the various agents who could become coupled to potentially destructive artifacts. And this leads to the crucial point: as soon as agents enter the picture, we have another variable that could be manipulated through targeted interventions to reduce the overall probability of an existential catastrophe.

The options here are numerous. One possibility would involve using "moral bioenhancement" techniques to reduce the threat of terror, given that acts of terror are immoral. But a morally enhanced individual might not be less likely to make a mistake. Thus, we could attempt to use cognitive enhancements to lower the probability of catastrophic errors, on the assumption that greater intelligence correlates with fewer blunders. Furthermore, implementing stricter regulations on CO2 emissions could decrease the probability of extreme ecoterrorism and/or apocalyptic terrorism, since environmental degradation is a "trigger" for both.

Another possibility, most relevant to idiosyncratic agents, is to reduce the prevalence of bullying (including cyberbullying). This is motivated by studies showing that many school shooters have been bullied, and that without this stimulus such individuals would have been less likely to carry out violent rampages. Advanced mind-reading or surveillance technologies could also enable law enforcement to identify perpetrators before mass casualty crimes are committed.

As for superintelligence, efforts to solve the "control problem" and create a friendly AI are of primary concern among many researchers today. If successful, a friendly AI could itself constitute a powerful mitigation strategy for virtually *all* the categories listed above.

(Note: these strategies should be explicitly distinguished from proposals that target the relevant tools rather than agents. For example, "differential technological development" aims to neutralize the bad uses of technology by strategically ordering the development of different kinds of technology. Similarly, the idea of police "blue goo" to counter "grey goo" is a technology-based strategy. Space colonization is also essentially a tool intervention because it would effectively reduce the *power* (or capacity) of technologies to affect the entire human or posthuman population. Finally, in a forthcoming paper, I explore the possibility of *space bunkers*, which are "technologically minimalist, causally isolated spacecraft" designed to preserve the human species in the case of planetary

catastrophes until exoplanet colonies are widely established. This could provide an extra layer of insurance while our species passes through a period of heightened existential hazards described by the "bottleneck hypothesis.")

6. **Agent-Tool Couplings and the "Hardware Bias"**

Devising novel interventions and understanding how to maximize the efficacy of known strategies requires a careful look at the unique properties of the agents mentioned above. Without an understanding of such properties, this important task will be otiose. We should also prioritize different agential risks based on the likely membership (token agents) of each category. For example, the number of idiosyncratic agents might exceed the number of ecoterrorists in the future, since ecoterrorism is focused on a single issue, whereas idiosyncratic agents could be motivated by a wide range of potential grievances. [3] We should also take seriously the formidable threat posed by error, which could be more significant than that posed by terror, as the back-of-the-envelope calculations above show.

Such considerations, in combination with technology-based risk mitigation strategies, could lead to a comprehensive, systematic framework for strategically intervening on *both sides* of the agent-tool coupling. But this will require the field of existential risk studies to become less technocentric—to overcome the *hardware bias* that leads many scholars to focus exclusively on superintelligence. There are other, flesh-and-bone agential risks that could bring about an existential catastrophe long before humanity creates the first superintelligent machine. These human-agent risks must be understood and neutralized.

Footnotes:

[1] Thanks to Daniel Kokotajlo for bringing this possibility to my attention.

[2] Thanks to Dr. Shankar Bhamidi for helpful feedback on these calculations.

[3] Although the stimulus of environmental degradation would be experienced by virtually everyone in society, whereas the stimuli that motivate idiosyncratic agents might be situationally unique. It's precisely issues like these that deserve further scholarly research.