Agential Risks: A New Direction for Existential Risk Scholarship

Technical Report

X-Risks Institute

www.xrisksinstitute.com

Authors: Phil Torres
philosophytorres@gmail.com
@xriskology

Phil Torres is an author, blogger at the Future of Life Institute, Affiliate Scholar at the Institute for Ethics and Emerging Technologies, founder of the X-Risks Institute, and scholar whose work focuses on emerging technologies, existential risks, and apocalyptic terrorism. He's been published in *Skeptic*, *Free Inquiry*, *The Humanist*, *Bulletin of the Atomic Scientists*, *Humanity+*, *Foresight*, *Metaphilosophy*, *Erkenntnis*, *Journal of Future Studies*, *Journal of Evolution and Technology*, *Salon*, *Truthout*, and *Common Dreams*, among others. His most recent book is called *The End: What Science and Religion Tell Us About the Apocalypse* (Pitchstone Publishing).

X-Risks Institute Mission Statement:

Emerging technologies are introducing brand new risk scenarios that humanity has never before encountered. Some of these technologies will likely place unprecedented power in the hands of a large number of people.

Not only must we examine the tools of mass destruction—nuclear weapons, biotechnology, synthetic biology, nanotechnology, and artificial intelligence—but we also need to study the agents—lone wolves, ecoterrorists, and apocalyptic ideologues—that might attempt to use advanced technologies to destroy civilization.

The central aim of the X-Risks Institute is to understand the unique technological and agential threats facing humanity this century.

Agent-Tool Couplings and Risk Potential

Many existential risk scholars argue that advanced technologies pose the most significant existential risks to humanity this century. Examples of such technologies include nuclear weapons, biotechnology, synthetic biology, nanotechnology (especially molecular manufacturing), and artificial intelligence (AI). But these tools aren't going to initiate a catastrophe on their own. They require a suitable agent to exploit them for harmful ends.[1]

To illustrate this point, imagine two worlds: X contains a large number of species-destroying technologies and a population of peaceable, compassionate individuals, whereas Y contains only a single species-destroying technology and a population of violent, warmongering peoples. Which world is more likely to self-annihilate?

If one looks only at the *technological risks*, then one would likely pick world X. But picking X is problematic because (what could be called) the *risk potential* of advanced technologies can only be realized by a complete *agent-tool coupling*. In other words, an engineered pandemic requires an engineer, just as a nuclear missile launch requires a nuclear missile launcher. It follows that one must consider the relevant population of agents as well. If one does this, world Y appears more dangerous than X, because the *agential risk* of Y is greater than that of X.

One of the central topics of existential risk studies today is, in fact, an agent: artificial superintelligence. Many scholars overlook this fact because artificial superintelligence is also a technology, and agents like states and terrorist groups can use narrow AI systems as *tools* to attack their enemies. But it's important to resist the conclusion that because artificial superintelligence is a technology, it constitutes a technological risk. This is not necessarily true. When an AI acquires human-level intelligence or beyond, it becomes an agent in its own right, capable of making decisions in pursuance of its own goals. The result is an agential risk.

If coupled to any of the tools mentioned above (including narrow AI systems), an artificial superintelligence could pose an existential risk to humanity. For example, it could control nanofactories to create swarms of lethal insect-sized drones, exploit automated processes in biological laboratories to synthesize a designer pathogen, or initiate a nuclear war by producing false alarms in early-warning systems.

But the primary reason that risk scholars worry about superintelligence is because of its unique *agential properties*, rather than the tools that might be available to it. The past decade of research has shown that a superintelligence's values might not align with ours (the orthogonality thesis), it need not hate humanity to harm us (instrumental convergence), and it could rapidly gain new capabilities through an intelligence explosion (or recursive self-improvement). Only by understanding these agential properties can scholars accurately assess the potential dangers of superintelligence, and therefore avoid an AI disaster.

According to a survey of experts conducted by the Future of Humanity Institute, the average American is more than one thousand times more likely to die in a human extinction event than a plane crash. But this survey focuses almost entirely on technological and natural risks, rather than agential risks. Without taking agential risks into account, it's impossible to devise an accurate overall probability estimate of disaster.[2]

Other Types of Agential Risks

Yet hardly any existential risk scholars today take seriously the other agents that might, through error or terror, employ advanced technologies to inflict harm on society. This is a mistake. As the thought experiment above shows, it's impossible to accurately evaluate the potential risks to our species without an understanding of *both sides* of the agent-tool coupling.

What other agents might use advanced technologies to induce an existential catastrophe? Obvious candidates are rogue states and terrorists. But in most cases, the value of *self-preservation* constrains the behavior of such agents. Consequently, rogue states and terrorists are unlikely to bring about a worst-case scenario, except by accident (that is, through error).

For example, North Korea might (someday) use nuclear weapons against its perceived enemies to either defend its territory or expand its borders. Annihilating *Homo sapiens* would, of course, interfere with this goal. Similarly, it's unlikely that a terrorist group motivated by political ideologies, such as the Irish Republican Army (IRA) or the Tamil Tigers, would use advanced technologies to destroy civilization, since their goals are often predicated upon its continued existence. They want to change the world, not destroy it.

Nonetheless, there are some agents who might intentionally try to cause an existential catastrophe.

1. *Radical environmentalists*. Consider the case of Ted Kaczynski, also known as the Unabomber. He wanted to dismantle technological society and replace it with a primitive system in which localized communities live harmoniously with the natural world. Kaczynski sent a number of homemade bombs to people around the country, killing three and injuring twenty-three. But what if he had sufficient means available to achieve his Luddite goals? What if Kaczynski had access to advanced technologies like synthetic biology and nanotechnology? Would he have used these tools to catapult humanity back into the Stone Age? The answer is, "Yes, probably."

While ecoterrorism poses a negligible risk to civilization today, this may change in the future. As the terrorism scholar Frances Flannery notes, extreme weather, megadroughts, ecological collapse, biodiversity loss, species extinctions, desertification, deforestation, and agricultural failures could fuel the ecoterrorist movement later this century. In her words, "As the environmental situation becomes more dire, eco-terrorism will likely become a more serious threat in the future."[3]

Consequently, it's not implausible to imagine a future ecoterrorist group creating, for example, a population of nanobots that selectively targets *Homo sapiens* without harming the biosphere more generally. Ecoterrorists could accomplish this by designing nanobots that recognize genetic signatures that are unique to humanity. A similar scenario could involve engineered pathogens, or even nuclear weapons, if the long-term benefits to Earth-originating life are deemed to outweigh the short-term costs of a nuclear winter.

2. *Apocalyptic terrorists.* Apocalyptic terrorism also poses a significant threat to our collective future. It's the most dangerous form of religious terrorism, which is far more lethal and indiscriminate than past forms of "secular" terrorism. (One should distinguish here between religion per se and religious extremism. The percentage of religious extremists is small compared to the percentage of law-abiding adherents. Nonetheless, the total population of religious extremists will almost certainly grow in the future.[4] And the increasing accessibility of advanced technologies will make even a single extremist group unprecedentedly dangerous.)

Apocalyptic terrorists believe that *the world must be destroyed to be saved.* Good and Evil are locked in a cosmic struggle at the climax of world history, and the only acceptable outcome is the total destruction of God's enemies. As Jessica Stern and J. M. Berger observe, apocalyptic groups aren't "inhibited by the possibility of offending their political constituents because they see themselves as participating in the ultimate battle." Consequently, they are "the most likely terrorist groups to engage in acts of barbarism."[5] The apocalyptic terrorist doesn't just want a fight, he or she wants a *fight to the death.*

Consider the Christian Identity movement in the US, which holds that the world must be "purified" through catastrophic violence before Jesus returns to Earth. This ideology has influenced groups like the Aryan Nations and CSA (mentioned below). As a result, nuclear terrorism experts have kept a close eye on it.[6]

Similarly, the Islamic State is an apocalyptic terrorist group that has openly fantasized about acquiring nuclear weapons from Pakistan. It has also explored weaponizing the bubonic plague and spreading Ebola via modern transportation systems. As one Islamic State sympathizer wrote, "The advantages of biological weapons is [*sic*] the low cost and high rate of casualties."[7]

And the Japanese doomsday cult Aum Shinrikyo tried to hasten Armageddon in 1995 by releasing sarin in the Tokyo subway. This was the deadliest terrorist attack in Japan's history.

It may come as a surprise to some readers that apocalyptic movements are ubiquitous across cultural space and time. From the Crusades to the founding of Islam, end-times thinking has shaped some of the most significant events in history. Even Nazism and Marxism were infused with eschatological themes borrowed from Christianity, as Daniel Chirot and Clark McCauley convincingly argue.[8] Beliefs about the how the world will end have influenced how the world *is* more than most people realize.

3. *Idiosyncratic agents.* A final type of agent that could be existentially dangerous if coupled to advanced technologies includes lone wolves and groups with idiosyncratic motives. History provides numerous examples from which to extrapolate.[9]

Consider the case of Marvin Heemeyer. A resident of Colorado, he became embroiled in a dispute with his local town. To take revenge, he spent months building a bullet-proof "futuristic tank" out of a bulldozer. He then drove the tank into town and, over the course of two hours, leveled a large number of the town's buildings. But what if Heemeyer had a grudge not with the local government, but with human civilization? What if Heemeyer had access to advanced technologies rather than a bulldozer? Would he have tried to destroy the world? The answer is, “Quite possibly.”

Another example comes from school shooters. Consider the April 20, 1999 Columbine High School massacre. The perpetrators, Eric Harris and Dylan Klebold, brought propane tanks converted into bombs to school and equipped themselves with several guns. Their goal was to kill as many people as possible before committing suicide. Again, we can ask: imagine that Harris and Klebold—or someone with a similar psychological disposition—had access to advanced technologies. Such tools could make “killing as many people as possible” tantamount to “destroying humanity.” Anticipated future technologies will offer a tempting way for deranged individuals to “go out with the ultimate bang.”

Climate change could make ecoterrorism a greater threat in the future. It could also fuel other forms of terrorism, such as apocalyptic terrorism.[10] Furthermore, religion is expected to grow worldwide, with more than sixty percent of humanity identifying as either Christian or Muslim by 2050. This suggests that the extremist fringe may grow in absolute numbers as well, perhaps significantly.[11]

Mitigating Agential Risks

Insofar as these types of risks have their own unique properties, they should be studied on their own, as unique sources of danger. Failing to do this could leave humanity vulnerable to otherwise avoidable catastrophes.

For example, some researchers have suggested that bullying and school shooting-type events are causally connected. If this is true, then we should (*prima facie*) try to reduce the prevalence of bullying. This may sound trivial, but if advanced technologies become widely available in the future, reducing the number of disgruntled members of society could become critical for our continued survival. Similarly, if environmental degradation fuels ecoterrorism, then we have yet another compelling reason to mitigate the "conflict multipliers" of climate change and biodiveristy loss.

(In fact, climate change and biodiveristy loss could nontrivially elevate the probability of nearly every other risk facing humanity this century. Interstate wars, civil wars, and terrorist attacks will very likely become more common as the environmental crisis worsens. It follows that these phenomena ought to be top priorities moving forward.)

With respect to apocalyptic movements, one must understand their motivating ideologies in order to neutralize them.

For example, only by studying the Islamic tradition can one know that the year 2076 will coincide with a spike in apocalyptic fervor. The reason is that 2076 corresponds to 1500 in the Islamic calendar, and the turn of the century is a time of renewal in the Islamic world. In fact, the beginning of past centuries has seen major political revolutions and acts of terrorism. It's not a coincidence that the Iranian Revolution and the Grand Mosque seizure—both of which had apocalyptic significance to some Muslims[12]—happened in 1979, which corresponds to 1400 in the Islamic calendar.

Given the weapons that might be available in 2076, existential risk scholars should be especially cautious as this year approaches.[13]

Similarly, many rightwing extremists project special significance onto April 19. Consider the fact that Timothy McVeigh detonated a bomb outside the Alfred P. Murrah Federal Building in Oklahoma City on April 19, 1995. Two years earlier to the day, the Waco siege between the US government and the apocalyptic Branch Davidians in Waco, Texas, came to a tragic end, with 76 deaths. (This was one of the primary inspirations for McVeigh's terrorist attack.) And exactly eight years before this, the US government confronted The Covenant, The Sword, and Arm of the Lord (CSA), a Christian Identity militia that trained 1,200 recruits in the "Endtime Overcomer Survival Training School."

Even more, centuries before these incidents, the Battles of Lexington and Concord inaugurated the Revolutionary War against the "tyrannical" British Empire on April 19, 1775. Thus, as Flannery observes, the day of "April 19 has come to resonate throughout a construed history of the radical Right as a day of patriotic resistance."[14]

Other dates of importance to rightwing extremists are April 15, the deadline for income tax filings, and April 20, Adolf Hitler's birthday. This is why the Anti-Defamation League implores "law enforcement officers, community leaders and school officials" to be "vigilant, especially during the period April 15 to April 24."[15] Existential risk scholars should heed this warning. The malicious agents of the future will have bulldozers, rather than shovels, to dig mass graves for their enemies.

Conclusion

If the task of existential risk studies is to maximize the probability of an "okay outcome" for humanity, "where an 'okay outcome' is any outcome that avoids existential disaster"—to quote the Oxford philosopher Nick Bostrom[16]—then scholars must take seriously the many risks posed by agents *other than* artificial superintelligence. Without a careful analysis of ecoterrorists, apocalyptic terrorists, idiosyncratic actors, and additional agential risks not here discussed, our collective future will remain less certain than it otherwise could be.

Key Terms

Tool: any technology that enables an agent to manipulate the world.

Agent: any autonomous entity with the capacity to choose its actions. Agents may couple themselves with tools to optimize their capacity to accomplish their goals.

Technological risk: the potential for a technology to enable an agent to cause a catastrophe.

Agential risk: the potential for an agent to intentionally (terror) or accidentally (error) use technology to cause a catastrophe. (This article focuses on terror agential risks.)

Additional reading:

Torres, Phil. 2016. Agential Riskology: A Comprehensive Introduction." *Journal of Evolution and Technology*. 26(2). (forthcoming)
Torres, Phil. 2016. "Apocalypse Soon? How Emerging Technologies, Population Growth, and Global Warming Will Fuel Apocalyptic Terrorism in the Future." *Skeptic*. http://goo.gl/Xh9JqO.
Torres, Phil. 2016. "The Clash of Eschatologies." *Skeptic*. (forthcoming)
Flannery, Frances. 2016. *Understanding Apocalyptic Terrorism*. New York, NY: Routledge.

References:
[1] This report draws heavily from "Agential Riskology: A Comprehensive Introduction." *Journal of Evolution and Technology*. 26(2). Due to space limitations, several important distinctions are ignored above. The aforementioned paper provides a detailed analysis and complete typology of agential risks.
[2] The survey gives a 19% overall probability of human extinction this century. If the average American lives

80 years, this results in a 15.5% probability of extinction in one's lifetime. By comparison, the average American has a 1 in 9737 lifetime risk of dying in an "Air [or] space transport accident." Compare this calculation with the 2016 Global Challenges Foundation Report, which states, "The UK's Stern Review on the Economics of Climate Change suggested a 0.1% chance of human extinction each year. If this estimate is correct, a typical person is more than five times as likely to die in an extinction event as in a car crash." See Bostrom, Nick, and Anders Sandberg. 2008. "Global Catastrophic Risks Survey." *Technical Report*. http://www.fhi.ox.ac.uk/gcr-report.pdf, Insurance Information Institute. 2016. "Morality Risk." http://www.iii.org/fact-statistic/mortality-risk, and Global Challenges Foundation. 2016. "Global Catastrophic Risks." http://www.globalchallenges.org/reports/Global-Catastrophic-Risk-Annual-Report-2016.pdf. For more about different probability estimates of an existential catastrophe, see Torres, Phil. 2016. "Existential Risks Are More Likely to Kill You Than Terrorism." Future of Life Institute. http://futureoflife.org/2016/06/29/existential-risks-likely-kill-terrorism/.

[3] Flannery, Frances. 2016. *Understanding Apocalyptic Terrorism*. New York, NY: Routledge.

[4] Torres, Phil. 2016. "Apocalypse Soon? How Emerging Technologies, Population Growth, and Global Warming Will Fuel Apocalyptic Terrorism in the Future." *Skeptic*. http://goo.gl/Xh9JqO.

[5] Stern, Jessica, and JM Berger. 2014. *ISIS: The State of Terror*. New York, NY: HarperCollins Publishers.

[6] See Ferguson, Charles, and William Potter. 2005. *The Four Faces of Nuclear Terrorism*. New York, NY: Routledge.

[7] See Torres, Phil. 2016. *The End: What Science and Religion Tell Us About the Apocalypse*. Durham, NC: Pitchstone Publishing.

[8] Chirot, Daniel, and Clark McCauley. 2006. *Why Not Kill Them All?: The Logic and Prevention of Mass Political Murder*. Princeton, NJ: Princeton University Press. For a detailed look at the influence of apocalyptic beliefs throughout history, see: Torres, Phil. "The Clash of Eschatologies." *Skeptic*. (forthcoming)

[9] These examples may seem too anecdotal to be scientifically useful. But drawing this conclusion would be a mistake. Given the nature of anticipated future technologies, single individuals or groups could potentially wield sufficient power to destroy the world. The statistically anomalous cases of lone wolves or terrorist groups are precisely the ones we should be worried about, and therefore ought to study.

[10] Torres, Phil. 2016. "Apocalypse Soon? How Emerging Technologies, Population Growth, and Global Warming Will Fuel Apocalyptic Terrorism in the Future." *Skeptic*. http://goo.gl/Xh9JqO.

[11] Pew Research Center. 2015. "The Future of World Religions: Population Growth Projections, 2010-2050." http://www.pewforum.org/2015/04/02/religious-projections-2010-2050/.

[12] See Cook, David. 2011. "Messianism in the Shiite Crescent." Hudson Institute. http://www.hudson.org/research/7906-messianism-in-the-shiite-crescent.

[13] Another date of concern is potential 2039, which roughly corresponds to the 1200th anniversary of the Mahdi's occultation. As Cook writes, "given the importance of the holy number 12 in Shiism, the twelfth century ... could also become a locus of messianic aspirations. In one scenario, either a messianic claimant could appear or, more likely, one or several movements hoping to 'purify' the Muslim world (or the entire world) in preparation for the Mahdi's imminent revelation could develop. Such movements would likely be quite violent; if they took control of a state, they could conceivably ignite a regional conflict." See Cook, David. 2011. "Messianism in the Shiite Crescent." Hudson Institute. http://www.hudson.org/research/7906-messianism-in-the-shiite-crescent.

[14] Flannery, Frances. 2016. *Understanding Apocalyptic Terrorism*. New York, NY: Routledge.

[15] Anti-Defamation League. 2005. "Extremists Look to April Anniversaries." http://goo.gl/q2mH6F.

[16] See Bostrom, Nick. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology*. http://www.nickbostrom.com/existential/risks.html.