ARTICLE IN PRESS

Futures xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Futures

journal homepage: www.elsevier.com/locate/futures

Original research article

Agential risks and information hazards: An unavoidable but dangerous topic?

Phil Torres

Project for Future Human Flourishing, Philadelphia, PA, USA

ARTICLE INFO

Keywords:
Agential risks
Existential risks
Omnicidal agents
Global catastrophic risks

ABSTRACT

This paper proposes an original theoretical framework for understanding anthropogenic existential risks associated with advanced dual-use technologies. It outlines a typology of “agential risks,” of which I argue there are four primary types: apocalyptic terrorists, misguided moral actors, ecoterrorists, and idiosyncratic actors. I then explore the issue of agential risks and information hazards, arguing that although there are nontrivial dangers associated with agential risk scholarship, the benefits currently outweigh the risks. The paper’s primary aim is to establish a conceptual foundation for understanding the range of individuals who might attempt to exploit current and future technologies to bring about an existential catastrophe.

1. The theory of agential risks

Recent scholarly work within the interdisciplinary field of existential risk studies has begun to focus on the various human nonstate actors who might “couple” themselves to advanced technologies and bring about an existential catastrophe. This topic is both unavoidable and increasingly important given (T1) the growing power and (T2) the increasing accessibility of dual-use emerging technologies. Examples include digital-to-biological converters, CRISPR/Cas-9, base editing, SILEX (i.e., separation of isotopes by laser excitation), and anticipated future artifacts like nanofactories, self-replicating nanobots, and autonomous artificial intelligence systems (e.g., lethal insect-sized drones). The result of these dual trends is the rapid distribution of increasingly destructive capabilities across society, thus multiplying the total number of state and—*most importantly*—nonstate actors capable of unilaterally destroying the world. Elsewhere I have termed this the “threat of universal unilateralism” and shown how, following Sotos (2017), it has direct implications for the “doomsday hypothesis” (i.e., that a Great Filter lies ahead), as well as for the contractarian foundations of the modern state system (Torres, 2017a).

It follows that to obviate a worst-case outcome for our species, existential risk scholars ought to focus no less on the various properties of individual agents who might destroy the world than on the various properties of “weapons of total destruction” (WTDs) that could enable them to do this. The importance of this point is underlined by a simple gedankenexperiment, namely, the *two worlds thought experiment*. This asks us to imagine two worlds, A and B, where world A contains a single WTD and world B contains 10,000. The question is which world one would rather inhabit based entirely on security considerations, and the obvious answer is world A. But it would be hasty to choose this world without asking for further information about the kinds of beings who inhabit A and B. Thus, imagine further that world A is run by an alien species of bellicose warmongers whereas world B is run by an alien species of irenic peaceniks. Given this additional information about the moral and psychological characters of each population, I would argue that world B appears less likely to self-destruct, and therefore constitutes the most judicious answer. To dissect this conclusion: for an *agent-artifact coupling* to bring about a global disaster, the necessary and sufficient conditions of *means and motivation* (i.e., of being

E-mail address: philosophytorres@gmail.com.

https://doi.org/10.1016/j.futures.2017.10.004
Received 17 July 2017; Received in revised form 20 October 2017; Accepted 25 October 2017

"able and willing") must be satisfied. Thus, whereas both are satisfied in world A, only one is satisfied in world B, and this is what makes world A more existentially hazardous.

If understanding both sides of the agent-artifact coupling is indeed important, the next question to ask is: *Who exactly would destroy the world if only the means were available?* Here we must follow Rees (2004) in distinguishing between terror agents and error agents, where each could destroy the world if they were to gain access to a WTD, but only the former would do this on purpose. Although the topic of agential error is, I believe, important and neglected, the present paper will focus exclusively on agential terror. Thus, the relevant question becomes: *Who exactly would destroy the world on purpose if only the means were available?* I would contend that the answer to this question is not as obvious as it may appear *prima facie*, and in fact it has received almost no serious scholarly attention in *any* field of intellectual inquiry, including the field to which it is most directly germane, existential risk studies.

Nonetheless, one finds many references to candidate answers to this question scattered throughout the literature—these candidates just haven't been organized in any coherent way, which will be the task of Section 2. For example, scholars have used colorful descriptors like "maniacs," "lunatics," "misanthropes," "sociopaths," "nefarious dictators," "belligerent tyrants," "agents of doom" (Yudkowsky, 2008), "suicidal regimes or terrorists" (Bostrom, 2002), "garage fanatics and psychopaths" (Roden, 2015), "criminal groups, terrorists, and lone crazies" (Wittes & Blum, 2015). A particularly concise example of grasping for clarity on this complicated issue can be found in Sagan (1994) *Pale Blue Dot*:

> Can we humans be trusted with civilization-threatening technologies? [Consider] some misanthropic sociopath like a Hitler or a Stalin eager to kill everybody, a megalomaniac lusting after "greatness" and "glory," a victim of ethnic violence bent on revenge, someone in the grip of unusually severe testosterone poisoning, some religious fanatic hastening the Day of judgment, or just technicians incompetent or insufficiently vigilant in handling the controls and safeguards. Such people exist.

One way to impose some conceptual-ontological order on this jumble of imprecise terminology involves what we can call the *doomsday button test*. This is a simple mechanism for determining which agents, whether real or hypothetical, would intentionally cause an existential catastrophe if they could. It is, in other words, a *filter* that enables one to answer the "who" question posed above. The idea is this: imagine that a "doomsday button" were suddenly placed in front of every person alive on the planet. If pushed, this button would initiate a WTD that would immediately cause either human extinction or the permanent collapse of civilization. Having isolated all sorts of potential confounding factors, one can then consider and analyze individual cases one by one, ultimately yielding a list of token individuals who would possibly, probably, or almost certainly "pass" the test.

For example, imagine a doomsday button suddenly presented to members of the Provisional Irish Republican Army (PIRA) during the height of conflict with the British government. Would any terrorist fighting for PIRA push it? Almost certainly not, since destroying the world would interfere with PIRA's provincial political goals of chasing the British out of Northern Ireland. This answer can be generalized to nearly all forms of political, nationalist-separatist, Marxist, anarchist, anti-government, and single-issue terrorism: individuals motivated by the corresponding ideologies are unlikely to willingly destroy the world even if the opportunity were presented. The same goes for most forms of religious terrorism, which the Global Terrorism Index now identifies as the primary manifestation of global terrorism today (see Torres, 2016a). For example, Osama bin Laden didn't harbor fantasies of killing every human on Earth or causing the total collapse of civilization. Rather, his religio-political goals were more focused on crippling Western civilization because of its religious infidelity and jingoistic foreign policy. In particular, bin Laden's campaign of terror that culminated in the 9/11 attacks were motivated by the US military presence in Saudi Arabia and devastating sanctions on Iraq—which resulted in immense human suffering—and his ultimate goal was to establish a global Caliphate before the Last Hour. Thus, it seems highly unlikely, in my view, that he would have pushed a doomsday button if one had been placed in front of him at any moment from, say, the late 1980s until his death in 2011. Similar claims can be made about most world leaders, even the most grandiose, megalomaniacal, militaristic autocrats. Simply put, one cannot rule the world if the world doesn't exist, and this provides a strong incentive for rational actors at the helm of states not to bring about global-scale catastrophes.

Thinking about such examples in the context of the doomsday button test might initially lead one to concur with Eliezer Yudkowsky (2008) that "all else being equal, not many people would prefer to destroy the world." In fact, I would argue that this statement is true, but only if the ambiguous word "many" is understood in *relative* rather than *absolute* terms.[1] That is to say, the total number of people who would pass the doomsday button test is indeed small when *compared* to the human population of 7.6 billion going on 9.3 billion, yet I would also argue that the total number of malicious agents is nonetheless alarmingly large. This is perhaps the most relevant issue—the absolute number—given the trends of (T1) and (T2), because as Rees (2004) and so many other scholars have emphatically argued, it could take only a *single* lone wolf or small group in the future to bring about ruinous consequences for humanity. Since I provide a detailed examination of actual individuals who would almost certainly pass the doomsday button test elsewhere, the present paper will embrace a more theoretical approach (see Torres, 2017b). Thus, the next section will outline an abstract typology of human agents who would almost certainly destroy the world if only they could. I will illustrate these types with a few real-world examples, but the primary aim will be to establish a conceptual foundation for understanding the "agent" side of the agent-artifact coupling, which gives rise to a specific kind of risk that we can call an "agential risk," defined as follows:

> *Agential risk*: the risk posed by any agent who could initiate an existential catastrophe in the presence of sufficiently powerful dual-use technologies either on purpose or by accident.

[1] Thus, I am not confident in Yudkowsky's (2008) conclusion that "if the Earth is destroyed, it will probably be by mistake." This requires more scholarly attention before taking sides on the issue.

Before proceeding to the next section, we can make one more distinction within the category of agential terror between *omnicidal agents* and *anti-civilizational agents*. This distinction not only captures a real difference among agential risks in the world (some of whom fall within the same overarching type), but it roughly tracks the canonical definition of an existential risk (alluded to above) as "one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development" (Bostrom, 2013). Omnicidal agents wish to actualize the former disjunct whereas anti-civilizational agents wish for the latter. This being said, there are also extremely dangerous agents who don't clearly fall into either category, such as the Finnish eco-fascist Pentti Linkola, who (a) doesn't want humanity to go extinct, (b) doesn't target civilization the way, for example, anarcho-primitivists and neo-Luddites typically do, yet (c) does advocate a global-scale involuntary genocide that significantly reduces the human population. There are also some agents whose ambitions exceed omnicide: they wish to annihilate not merely all humans but the entire biosphere—or, in some cases, every instance of life in the universe.

With these nuances stated, we will now outline a four-part typology of agential risks. The penultimate section will then focus on a number of possible information hazards associated with agential risk studies, arguing that these hazards are real and significant but do not currently outweigh the benefits of exploring this topic.

2. A typology of agential risks

2.1. Types of omnicidal and anti-civilizational agents

(1) *Apocalyptic terrorists*: Let's consider two models for understanding this threat. The first involves the following tripartite distinction: (i) *Doxastic apocalypticists*: Individuals of this sort hold a passive belief that the end of the world, the "eschaton," is rapidly approaching, but they do not attempt to alter the timing or chronology of eschatological events. Extrapolating from recent history, we should expect a staggering 3.5 billion religious people to hold doxastic apocalyptic views by 2050.[2] (ii) *Activist apocalypticists*: Individuals of this sort take mostly indirect actions to alter the onset or chronology of eschatological events by, for example, fomenting the conditions necessary for Armageddon to occur. An example comes from the "Armageddon lobby" in the US, a sizable demographic of dispensationalist Christians who, for partly eschatological reasons, support a Jewish state in Palestine and often interpret wars and natural disasters with "a certain grim satisfaction," since such events are harbingers of the rapture (Haija, 2006; Walls, 2008). And (iii) *febrile apocalypticists*: Individuals of this sort take direct action to catalyze the apocalypse. They see themselves as divinely ordained participants in an eschatological narrative that is unfolding in realtime and will ultimately culminate in an epic clash between the cosmic forces of Good and Evil. Somewhat crudely speaking, there are inward-facing and outward-facing instances of this type: suicide cults can constitute the former while Aum Shinrikyo, the Islamic State, and groups influenced by Christian Identity constitute the latter. It is outward-facing febrile apocalypticists—agents who believe that "the world must be destroyed to be saved"—that concern the present paper.

Another model comes from Landes (2011). According to Landes, some eschatologies posit a *transformative* "call for a change of the heart," whereas others anticipate *cataclysmic* violence at the end of time. Furthermore, some religious believers identify the causal agents responsible for bringing about the apocalypse to be *supernatural* (e.g., God or a messianic figure like the Mahdi), whereas others believe that God has delegated this all-important task to *them*. This task is "all-important" because nothing means anything within the religious worldview if there is no divine justice to "set things right" at the end of time; indeed, eschatology constitutes the ultimate *theodicy*, or vindication of the existence of evil given God's supposed omnipotence and omnibenevolence. Even more, Pinker (2011) notes that the utopian aspect of most apocalyptic ideologies is especially dangerous because it sets up a "pernicious utilitarian calculus" whereby present suffering, however immense, can always be justified by the infinite moral value of paradise. Stern and Berger (2015) offer a similar observation, writing that apocalyptic groups aren't "inhibited by the possibility of offending their political constituents because they see themselves as participating in the ultimate battle." It follows that they are "the most likely terrorist groups to engage in acts of barbarism" (Stern & Berger, 2015).

With respect to agential risks, it is the active-cataclysmic mode of belief in the top-left triangle of Fig. 1 that constitutes the gravest danger. Simply put, if the world must be destroyed to be saved and if pressing a doomsday button would destroy the world, then active-cataclysmic extremists would press a doomsday button. Note here that many apocalyptic ideologies are *plastic* in that they can undergo significant changes over time from one quadrant of Fig. 1 to another. In fact, history is replete with apocalyptic groups that oscillated between the active and passive Gestalts as a result of endogenous (within the group) and exogenous (outside of the group) factors. In some cases, the evolution from a thoroughly passive eschatology to a violently active one has occurred quite quickly, meaning that anti-risk enforcement operations will need to keep their eyes on a wide range of apocalyptic movements in the future, not just those that currently embrace an active mode of belief.

Consider a few individuals who would very likely have destroyed the world if only the means had been available. (i) Shoko Asahara was the nearly blind leader of the Japanese doomsday cult Aum Shinrikyo. Although he initially embraced a passive-cataclysmic worldview according to which Armageddon was predicted to occur in 1999, his beliefs began to shift toward an active mode in the late 1980s and early 1990s. This ultimately lead to the 1995 Tokyo subway sarin attack that the group hoped would

[2] I calculate this number as follows: according to Pew (2015), some 8 billion of the 9.3 billion people who will be alive by the middle of this century will be religious. The two biggest religions will be Christianity and Islam, each with nearly 3 billion adherents. Current surveys show that roughly half of Christians and Muslims alike believe that the end of the world will occur within their lifetime, or in the near future. It follows that 3 billion plus 3 billion equals 6 billion, divided by 2 gives 3 billion. Conservatively adding another half billion for apocalypticists in other faith traditions gives 3.5 billion.

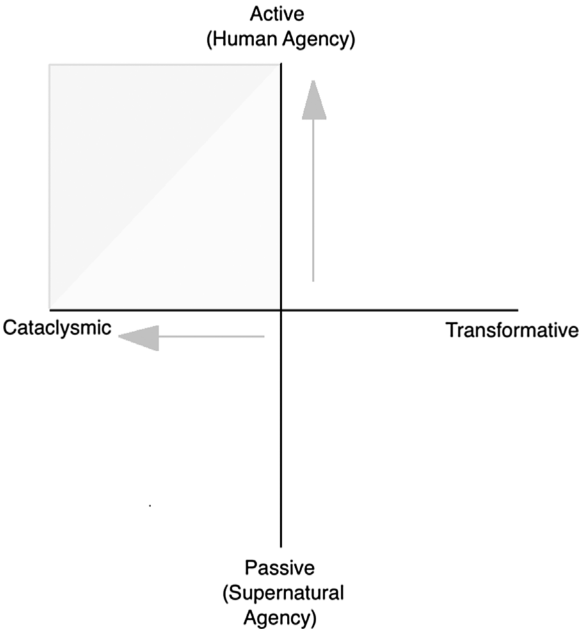

Fig. 1. Following the arrows leads to the most dangerous kind of apocalyptic movements. (Based on a figure in Landes, 2011).

initiate World War III, or Armageddon. (ii) Abu Musab al-Zarqawi was the grandfather of the Islamic State, or Daesh. A sadistic psychopath and sex offender, al-Zarqawi maintained in the mid-2000s that the Islamic version of Armageddon between the Muslim and "Roman" (i.e., Western) forces would soon occur in the small northern Syrian town of Dabiq, near Aleppo. Thus, after the 2003 US-led preemptive invasion of Iraq that many Muslims in the region, both Sunni and Shi'ite, saw as an apocalyptic event, al-Zarqawi proclaimed that "the spark has been lit here in Iraq, and its heat will continue to intensify—by Allah's permission—until it burns the Crusader armies in Dabiq." This later became a rallying cry for Daesh and its sizable army of Salafi Jihadists, and indeed the quote opens every issue of the propaganda magazine *Dabiq*. (iii) Finally, consider The Covenant, The Sword, and the Arm of the Lord (CSA). The group's founder, white supremacist James Ellison, gradually transformed its ideology from a passive-cataclysmic to an active-cataclysmic mode according to which a race war between the "two seedlines" of all non-white people and white Europeans was imminent, and it was CSA's duty to knock over the first domino of the apocalypse. In preparation for this war, CSA established an "urban setting" called *Silhouette City*, where it trained between 1200 and 1500 recruits in the "Endtime Overcomer Survival Training School." They also made plans to poison the Chicago and Washington D.C. public water supplies, destroy power grids in Oklahoma and Arkansas, and detonate a bomb near the Alfred P. Murrah building in Oklahoma City (an attack later perpetrated by Timothy McVeigh). Although the group never carried out these plans, a prominent member named Randall Rader, who later joined the Christian Identity organization known as "The Order," declared on video that "he was getting impatient because of how bad things were getting in the world" and that "if the Lord didn't hurry up and start Armageddon, he was determined to start it himself" (Flannery, 2016).

A book's-worth of examples could be adduced to show just how common across space and time, geography and history, violent apocalyptic movements have been. Indeed, I have elsewhere argued that a nontrivial yet largely invisible—for reasons that frustrate many scholars of apocalypticism—causal factor behind some of the most significant, world-altering events has been febrile apocalyptic hallucinations of what the world is like and, more importantly, *how it ought to be*. I call this the "clash of eschatologies" thesis (Torres, 2016b). Extrapolating the historical incidents upon which this thesis is based into the future—that is, into a milieu cluttered with world-destroying dual-use technologies—yields a frightening picture of the global security predicament.

(2) *Misguided moral actors*. First of all, let's be clear about terminology: the word "misguided" is defined from the particular perspective of transhumanism, since transhumanism provides the axiological framework in which Bostrom (2002, 2013) proposes the concept *existential risks*, and the concept of *agential risks* is defined in relation to the concept of existential risks. It follows that relative to another axiological framework, the moral views here discussed might not be considered "misguided"—and indeed many philosophers will argue that they are not in fact misguided at all. With this caveat out of the way, let's consider two forms of consequentialism, namely, negative utilitarianism and classical utilitarianism.

Ethicists have distinguished between many types of negative utilitarianism (NU), including strong or absolute, lexical, lexical threshold, weak, negative ideal preference, negative hedonistic, consent-based, and negative average preference utilitarianism (Chao, 2012; Ord, 2013; Pearce, 2017; Tomasik, 2016). The version relevant to the present discussion is strong/absolute NU, or the view that reducing suffering is *all* that matters. This leads to some dubious conclusions, such as that a world full of near-infinite pleasure and a single pinprick is less good than a world that doesn't exist at all (the "pinprick argument") and that there is no fundamental difference between (a) a world with zero suffering and near-infinite joy, and (b) a world that contains neither suffering nor joy (the "indifference argument"; Ord, 2013). But the most famous objection is R.N. Smart's claim that exponents of strong NU should endorse a "world-exploder" who simply destroys the universe, and therefore eliminates all possible suffering (Smart, 1958). In fact, there are some NUs who accept this conclusion but argue that one should not attempt to destroy the world for the merely *practical reason* that doing so would likely fail and, in the process, cause sentient life even more pain and misery.

Given the metaethical distinction between moral judgment and moral motivation, we could be even more precise about this agential risk subtype. First, if moral internalism is true and moral judgment subsumes moral motivation, then adopting the theory of strong NU could be sufficient for one to qualify as an agential risk on the above definition. But second, if moral internalism is false—which most philosophers believe is the case (see Bourget & Chalmers, 2014)—then we can distinguish between *strong* NU and *radical* NU. The former is simply the conviction that nothing matters morally except the reduction of suffering, whereas the second adds the "moral ambition" needed to actively pursue the annihilatory prescriptions of this position. In this case, it is the combination of judgment plus motivation that could lead one to pass the doomsday button test and thus destroy the world.

But radical NU isn't the only form of utilitarianism that could pose an existential threat to humanity. Consider David Pearce's argument that "a thoroughgoing classical utilitarian is obliged to convert your matter and energy into pure utilitronium, erasing you, your memories, and indeed human civilisation." Here "utilitronium" denotes a matter-energy configuration that is specifically designed to maximize (to the physical limits) whatever property one believes that a "util" measures; in other words, it is an organized state of physical stuff that is capable of realizing "the good" better than any human brain possibly could. It follows that *thoroughgoing* classical utilitarians (TCUs) are

> obliged to erase . . a rich posthuman civilisation with a utilitronium shockwave. . . The "shockwave" in utilitronium shockwave alludes to our hypothetical obligation to launch von Neumann probes propagating this hyper-valuable state of matter and energy at, or nearly at, the velocity of light across our Galaxy, then our Local Cluster, and then our Local Supercluster.

Although I know of no philosophers who explicitly endorse the creation of a utilitronium shockwave, TCU does appear to prescribe doing this. Thus, insofar as there exist defenders of TCU in the future, such agents could pose a threat to the perpetuation of our lineage.

(3) *Ecoterrorists*. The present typology adopts a liberal semantics of this term, using it to denote a range of overlapping but distinct agent subtypes, i.e., deep ecology extremists, radical environmentalists, eco-fascists, anti-civilization fanatics, anarcho-primitivists, violent technophobes, militant neo-Luddites, and fringe eco-anarchists (or green anarchists). This is where the cross-cutting distinction between omnicidal and anti-civilizational agents is perhaps most applicable: at one extreme are individuals who believe that the "Gaian system" would be better off without *Homo sapiens*, while at the other extreme one finds individuals primarily concerned with the negative externalities of advanced technological civilization.

For example, the neo-Luddite Ted Kaczynski argues in his 1995 manifesto that the megatechnics of industrial civilization have stripped humanity of essential freedoms. The only way to regain these freedoms is to dismantle and discard the "organization-dependent technologies" upon which contemporary society depends and return to local communities in which "small-scale technologies" satisfy our physiological and social needs (Kaczynski, 1995). For Kaczynski, the destruction of industrial civilization may or may not result in human casualties—the ultimate target isn't *Homo sapiens* but the large-scale technical structures that we have built up around ourselves. In recent years, Kaczynski has inspired a terrorist group called "Individuals Tending to the Wild (or Savagery)" (ITS), which has a nominal presence across Central and South America. ITS has specifically targeted scientists working on biotechnology and nanotechnology, claiming that "they must pay for what they are doing to the earth" (Beckhusen, 2013). In particular, ITS has voiced concerns that ecophagic nanobots could be accidentally released from a laboratory—a fear that is unfounded. The result has been a number of attacks, injuries, and deaths.[3]

In contrast to the anti-civilizational focus of Kaczynski and ITS, the deep ecology movement and its principle of *biospheric egalitarianism* has spawned a number of ecoterrorist groups that see humanity itself as the problem. According to biospheric egalitarianism, all living organisms have moral value in their own right; according to some radical environmentalists, all living organisms have the exact *same* intrinsic moral value. If one combines this view with the eschatological narrative that *Homo sapiens* is ruining the biosphere, one gets an anti-humanistic ideology that sees human extinction as desirable. Whereas some who accept this line of reasoning maintain that humanity should die out *voluntarily*—such as the Voluntary Human Extinction Movement and the Church of Euthanasia—others endorse a form of *involuntary* omnicide. For example, the Toronto-based Gaia Liberation Front (GLF) argues that humanity should be understood as an "alien species" that must be exterminated for the good of Earth-originating life (GLF, 1994). They suggest that a designer pathogen—or several different pathogens, released sequentially—could be uniquely suited to this task. This echoes an idea put forth in a 1989 article published in the *Earth First! Journal*, which states that "contributions are urgently solicited for scientific research on a species-specific virus that will eliminate Homo shiticus from the planet. Only an absolutely species-specific virus should be set loose" (Anonymous, 1989). According to Lee (1995), the *Earth First!* group that published the journal was initially "a small, tightly-knit millenarian movement," but when its co-founder David Foreman left, he was followed by his "apocalyptic biocentrist" acolytes. Finally, the founder of the Church of Euthanasia, Chris Korda, quotes her friend "Pete" in a "sermon" as follows:

> It becomes more and more clear to me every day that mass sterilization is the only answer to our environmental problems. . . I'm ready to hop in a B-52 with a payload of genetically-tailored-virus smart-bombs, enough to sterilize 99% of the world's population

[3] Note that ITS appears to have adopted a more pro-omnicide ideology recently (Torres, 2017d).

> in one *trans*-globe flight. Someone need only invent the hardware, train me, and present me with the opportunity. Maybe in 10 years it will be possible (see Torres, 2017b).

The point is that a plethora of groups and individuals exist at both extremes within this category: some harbor fantasies of civilizational collapse, which Rees (2004) estimates has a 50 percent chance of occurring before 2100; others dream about the total annihilation of humanity, given our deleterious impact on the natural environment. As the capacity for unilateral destructive action becomes more ubiquitous, ecoterrorists could become a major agential threat to our collective prosperity and survival.

(4) *Idiosyncratic actors*. This neologism refers to agents motivated by idiosyncratic beliefs and desires that do not clearly place them within any of the above categories. The paradigmatic case is rampage shooters, a demographic of violent malefactors that Peter Langman divides into three groups. (i) The *psychopaths* (or sociopaths). Estimates suggest that between 1 and 4 percent of the population suffers from psychopathy, meaning that there could exist some 300 million psychopaths in the world today and about 372 million by 2050 (Stout, 2005). Not all psychopaths are violent, of course, but they do comprise a disproportionate percentage of the prison population (~20 percent). The primary feature of psychopathy is the lack of a conscience, or "an inner feeling or voice viewed as acting as a guide to the rightness or wrongness of one's behavior" (Oxford, 2017). (ii) The *psychotics*. Individuals who suffer from psychosis, such as schizophrenia or schizotypal personality disorder, typically suffer from hallucinations or delusions that can compel them to lash out violently. While few people diagnosed with schizophrenia are dangerous (an important point to note given the risk of stigmatizing the condition), some rampage shooters have killed because the "voices in their head" commanded them to. And (iii) the *traumatized*. Individuals in this group have suffered from tragic histories of severe physical and/or psychological abuse.

Many rampage shooters have simply wanted to kill or injure as many people as possible before dying, thus leading them to shoot into crowds at random in hopes of maximizing casualties. Langman argues that rampage shooters are often motivated by what he calls "existential rage," meaning that "these were young men who were raging against the conditions of their existence. They were not just angry with a person or a group of people; they were angry at life, angry at the world" (Langman, 2009a, 2009b). This implies that rampage shooters would be good candidates for passing the doomsday button test. In fact, the personal journals and other records left behind by actual rampage shooters suggests that many had disturbingly omnicidal ideations. For example, the mastermind behind the 1999 Columbine High School massacre, Eric Harris, wrote about his urge to "burn the world." As he scribbled in his journal, "if you recall your history the Nazis came up with a 'final solution' to the Jewish problem. Kill them all. Well, in case you haven't figured it out yet, I say 'KILL MANKIND' no one should survive." He also wrote that "I think I would want us to go extinct," adding, "I just wish I could actually DO this instead of just DREAM about it all." Elsewhere he declared,

> if I can wipe a few cities off the map, and even the fuckhead holding the map, then great. Hmm, just thinking if I want all humans dead or maybe just the quote-unquote "civilized, developed, and known-of" places on Earth, maybe leave little tribes of natives in the rain forest or something. Hmm, I'll think about that.

Similarly, the rampage shooter Pekka-Eric Auvinen wrote, "I wish that death to mankind comes soon." In an online manifesto, he claims that "death and killing is not a tragedy" and that "human life is not sacred. Humans are just a species among other animals and [the] world does not exist only for humans." Auvinen further limned his homicidal attack in Finland as "an operation against humanity with the purpose of killing as many people as possible" and "one man's war against humanity." Yet another shooter, Matti Saari, penned a suicide note in which he opined, "I hate the human race, I hate mankind, I hate the whole world and I want to kill as many people as possible." Finally, Elliot Rodger declared in a video just days before his mass shooting:

> I hate all of you. Humanity is a disgusting, wretched, depraved species. If I had it in my power, I would stop at nothing to reduce every single one of you to mountains of skulls and rivers of blood. And rightfully so. You deserve to be annihilated. And I'll give that to you.

What's notable about these examples—a small handful sampled from a much larger set—is that they involve individuals who both (a) repeatedly expressed ghoulish visions of total human annihilation, and (b) actually engaged in catastrophic acts of violence. It follows that if such agents had access to WTDs rather than conventional weapons, the result could have been an existential disaster.[4]

Two additional subtypes of idiosyncratic actors are also worth mentioning. First, whereas a dictator cannot rule the world if the world doesn't exist (as mentioned above), there may be some militaristic autocrats who embody the following preference ordering: (i) total victory resulting in global domination, (ii) the complete destruction of the world, and (iii) defeat. In other words, there could be a "madman" who would rather live than die, controls WTDs, comes to believe that his military operations are doomed to failure, and thus presses a doomsday button to achieve (ii) rather than (iii), given the infeasibility of (i). A future Hitler could potentially instantiate such a person—after all, Hitler directed violence toward his own people as it became clear that the war was lost (for him). To quote Sagan (1994),

> in the winter and spring of 1945, Hitler ordered Germany to be destroyed—even "what the people need for elementary survival"—because the surviving Germans had "betrayed" him, and at any rate were "inferior" to those who had already died. If Hitler had nuclear weapons, the threat of a counterstrike by Allied nuclear weapons, had there been any, is unlikely to have dissuaded him. It might have encouraged him.

Another subtype of idiosyncratic actor is in certain respects the mirror opposite of this. Imagine that at some future time an

[4] Imagine, for example, if Eric Harris had opted to graduate from high school, get an undergraduate and master's degree in microbiology, and then plan his attack. The consequences could easily have been *global* and *transgenerational* in scope.

altruistic agent who genuinely cares about the well-being of humanity comes to believe that a "suffering risk," or "s-risk," is about to occur, thus resulting in immense pain and misery on Earth until the sun becomes a bloated red giant (see Tomasik, 2017). If some conditions of life are worse than death—presumably an uncontroversial idea—then this individual might attempt to preempt the s-risk event by using some potent WTD to euthanize our species. This is a somewhat complicated case to analyze because, on the one hand, if the information about an impending s-risk is veridical, then at least some moral theories would prescribe the "altruistic" agent's actions. But if the information turns out to be inaccurate, then this individual would have brought about an avoidable existential catastrophe, thus foreclosing the possibility of our descendants realizing astronomical amounts of value. This subtype of idiosyncratic actor directly relates to what Nick Bostrom, Tom Douglas, and Anders Sandberg (2016) call the "unilateralist's curse." No doubt there are other subtypes of idiosyncratic actors, but I will leave this topic for future papers.[5]

2.2. Additional issues

Before proceeding to the next section, we should consider a few additional issues concerning the typology and importance of this topic. First, the typology outlined above includes only human agents. There are at least two agential risk types that are non-human in nature: (a) an artificial general intelligence (AGI) or artificial superintelligence (ASI) could instantiate the "agent" side of the agent-artifact coupling no less than any human. Indeed, one of the primary reasons for existential anxiety about AGI and ASI is precisely because of their *agential status*—thus the "*control* problem." (b) An extraterrestrial intelligence of some sort could also instantiate the "agent" side of the coupling. As Susan Schneider (2016) argues, a spacefaring alien is much more likely to have an "artificial," hardware-based material substrate than a "biological" one, although it is conceivable that an alien could have the latter, where this substrate is perhaps non-carbon-based. I discuss these agential risk types elsewhere and so will mostly ignore them here (see Torres, 2017d).

Second, it may help to motivate the general goal of this paper to identify a number of unique insights associated with the agential risk framework:

(i) The most obvious advantage of thinking about the class of anthropogenic risks associated with emerging technologies in terms of the agent-artifact ontology is that this makes explicit that there are two variables that one can intervene upon to reduce the overall level of existential risk. Thus, I have elsewhere distinguished between *agent-oriented* and *technology-oriented* risk mitigation strategies, where moral bioenhancement exemplifies the former and differential technological development exemplifies the latter (see Torres, 2017d). Yet it is precisely because of the agential risk framework that recent proposals to use mostropics (moral enhancers) to overcome the game theoretic trap of the tragedy of the commons encounters serious problems, since a closer look at the likely effects of this intervention suggests that it could nontrivially exacerbate certain types of risky agents (Torres, 2017d, 2017e).

The point is that while most discussions to date about existential risks have narrowly focused on the relevant technologies and how to neutralize their potential harmful uses, the present approach emphasizes that there is another realm of possible mitigation strategies that target not the technologies but the users who would exploit them for bad ends. A combination of agent-oriented and technology-oriented strategies could even have synergistic effects that significantly improve humanity's odds of surviving the present bottleneck of existential risk.

(ii) Some types of agential risks pose a threat in and *only in* the specific context of "sufficiently accessible WTDs," meaning that the existence of such risks can only be seen through the prism of "agential risks." The most obvious example is (2) above: only once a radical NU gains access to a WTD does she or he pose a threat. Otherwise, more radical NUs in the world would likely yield a better world, all things considered, since NUs are dedicated to the alleviation of suffering, and most people would agree that less suffering is good. Similarly, until the first self-replicating von-Neumann probes capable of converting matter and energy into utilitronium becomes a reality, TCUs will pose no active dangers to the perpetuation of our species. These agents would never bleep the radar of anti-risk enforcement operations under the current *status quo* paradigm, which suggests that we need a new paradigm.

(iii) Furthermore, there may be certain agent-artifact configurations that are more likely to obtain than others, given the specific aims of different agential risks. Understanding this could help anti-risk enforcement operations focus their energy on the most probable configurations. For example, a radical NU would almost certainly never detonate a flurry of nuclear weapons around the world because this probably would not cause human extinction, but it could significantly increase human suffering. Thus, allocating large amounts of resources to ensure that radical NUs don't acquire weapon-grade uranium, for example, would be misguided. Instead, anti-risk programs should focus on agents of this type gaining the knowledge and skills necessary, say, to design self-replicating nanobots or, even more speculatively, weaponize a high-powered particle collider to induce a vacuum bubble. What other probable/improbable agent-artifact configurations might there be? The question has received almost no scholarly attention, so no one knows.

(iv) There could also be *triggers*, or information/events that acutely increase the probability of an intentional attack, that are only visible through the agential risk lens. Since this phenomenon falls under the umbrella of information hazards, we will delay discussing it until Section 3.2 below.

(v) A related phenomenon is what we can call *exacerbatory factors*. This refers to phenomena like climate change, the Anthropocene

[5] For a previous and less developed discussion of these categories, see Torres (2016a).

mass extinction, population growth, and globalization, to name a few, that can nontrivially elevate the overall risk associated with agent-artifact couplings by increasing the abundance and intensifying the motivation of risky agents in the world.[6] Consider that both terrorism scholars and scientists have posited a causal link between climate change and the formation of febrile apocalyptic movements: as one scholar puts it, "radical times will breed radical religion," to which he adds that climate change will almost certainly satisfy the "radical times" antecedent (Juergensmeyer, 2017; see also Kelley, Mohtadi, Cane, Seager, & Kushnir, 2015). Making matters worse, the global population of non-religious people is shrinking as a percentage of the total population, and about 8 billion out of 9.3 billion people alive in 2050 are projected to adhere to one or another faith-based belief system (Pew, 2015).[7] In particular, Islam is the fastest growing religion in the world with an estimated 2.76 billion adherents by the middle of this century. While a *tiny* percentage of Muslims today are violent Islamists—about 0.013 percent—the absolute number is unsettlingly huge—about 202,050. It follows that (a) not only will the instability caused by climate change probably increase the *percentage* of violent Islamists, but (b) even if (a) were not the case, *demographic inflation* alone will likely multiply the total number of extremists at the fringe.

With respect to ecoterrorism, Frances Flannery (2016) and Ackerman (2003) have both suggested that this threat will grow in the coming decades as the environmental plight becomes more dire, and I have elsewhere speculated about how societal disruptions associated with "context risks" like climate change and biodiversity loss could affect misguided moral actors and idiosyncratic actors (see Torres, 2017d). The point is that only within the agential risk framework does the *true* importance and urgency of these issues become clear—an importance and urgency that goes way beyond the usual worries of coastal flooding, extreme weather, and political instability, and concerns the very possibility of a Great Filter up ahead (again, see Sotos, 2017 and Torres, 2017f).

(vi) This leads to a final point about *prioritization*. If one accepts the "Maxipok rule" according to which "reducing existential risks should be a dominant consideration whenever we act out of an impersonal concern for humankind as a whole," then certain types of dangerous agents should be strongly prioritized over others (Bostrom, 2013).[8] For example, neutralizing active-cataclysmic apocalypticists should take precedence over neutralizing politically-motivated terrorists (e.g., PIRA); surveilling GLF should take precedence over surveilling monkeywrenchers; and far more resources for studying "misguided" ethical systems and rampage shooters should be allocated than is currently the case. In other words, the agential risk framework offers a *strategic guide* for deciding which types of agents should receive the most attention from scholars and anti-risk enforcement operations as the democratization of science and technology distributes civilization-ending power across the human population.

3. Information hazards

3.1. The dilemma of inquiry

Since this paper aims to lay the theoretical foundations for future scholarly work on the increasingly pressing question of *Who would destroy the world if only the means were available*?, this section will explore a range of issues concerning the possible information hazards associated with such scholarship. This task is important because one's views on these issues will influence and constrain one's subsequent work in the field. For example, if one believes that research on specific agential risk topics or particular modes of purveying one's findings (e.g., in the form of academic publications) could have a net negative impact on the overall level of existential risk, then one should avoid these topics and modes, and perhaps even advocate for a moratorium to prevent others from doing this work. My own tentative view is that the net effect of publicly accessible work on agential risks right now is positive, although I expect this to change as the threat of universal unilateralism grows and the "able" component of the "able and willing" formula becomes more easily satisfied by willing agents.

To begin, information hazards arise "from the dissemination or the potential dissemination of (true) information that may cause harm or enable some agent to cause harm" (Bostrom, 2011). For example, the blueprints for an atomic bomb and the genomic sequence of the Ebola virus are informationally hazardous because malicious agents could exploit them for nefarious ends. The question, then, is whether such information should be made available to the public. While there are rumors that nuclear weapon designs and how-to manuals are circulating on the international black market, a quick Google search confirms that anyone with a wifi connection can access the genomes of numerous pandemic viruses. Yet some scholars argue that the potential benefits of releasing this information outweigh the potential harms: since there are far more good actors than bad actors in the world, the good actors can work faster and more effectively to devise epidemiological defenses against bioterrorist attacks than bad actors can synthesize dangerous pathogens. Thus, the net outcome is positive.

Perhaps one could make a similar argument about studying agential risks: the potential benefits outweigh the potential harms. Unfortunately, this situation is not wholly analogous with the case above, since information about technologies will not (directly) make those technologies more dangerous, but information about agential risks could make those agents more risky. Yet ignoring the "agent" side of the agent-artifact coupling is fatuous given the considerations of (i) in section 2.2—i.e., to most effectively mitigate existential risk, we should intervene upon both the agents and the artifacts. This yields what we can call the "dilemma of inquiry":

[6] Essentially, the difference between (iv) and (v) is the timescale: triggers have short-lasting, immediate effects whereas exacerbatory factors increase the probability of agent-artifact-caused catastrophes over the course of months, years, decades, or centuries.

[7] In fact, environmental degradation itself will likely be interpreted through the eschatological hermeneutics of religion, which could increase religious adherence even more.

[8] Note that elsewhere I have argued against this rule; see Torres (2017g).

> *Dilemma of Inquiry*: Neglecting agential risks could leave humanity unnecessarily vulnerable to an existential catastrophe, since doing so would prevent scholars from devising effective agent-oriented mitigation strategies; yet openly talking about agential risks could unnecessarily elevate the associated dangers.

Here we will focus on the second horn of the dilemma.

How could agential risk scholarship exacerbate our existential predicament? The most obvious way is by drawing attention to the burgeoning literature on existential risks, which is full of clever ways to destroy the world. This could give malicious actors new ideas about which technologies could most efficiently realize their omnicidal or anti-civilizational fantasies. It could also motivate them to more actively pursue such fantasies in the first place. For example, consider the psychological phenomenon whereby one lacks an active desire to do X not because one doesn't actually want to do X, but because one doesn't believe that doing X is possible. The infeasibility of X makes it pointless to actively desire X. Thus, one can imagine an agent who isn't actively determined to synthesize a designer pathogen that would wipe humanity off the Earth simply because she or he is unaware of recent breakthrough gene-editing tools like CRISPR/Cas-9 and base editing. Not only could such technologies increase the dangerousness of already motivated agents, but information about these tools could cause an agent to morph from a passive to an active mode. In this sense, new knowledge of "can" could lead to "will" given some prior (and perhaps latent) "ought."

Another possibility is that risky agents are somehow "inspired" by being singled-out by scholars, or by related academic work that bears on their goals. Consider that a book about rampage shooters, written by Langman, had been studied by a German rampage shooter before he killed 9 people and injured 27 others. When Langman was asked about this, he said, "there must be some desire for self-understanding. . . It's common for shooters to want to study other shooters. . . They're having thoughts of doing this and want to read about other people who did this. . .Whether that's to find inspiration—I hate to use the word, 'inspiration'—or a role model, someone to emulate, it's hard to say" (Jamieson, 2016). Similarly, books by Noam Chomsky were found in the personal library of bin Laden (Burke, 2015).

Some agents could also take offense at certain scholarly analyses of agential risks. For example, notice that token agents within every category of agential risk except for (4) believe that destroying the world is *the moral thing to do*. Apocalyptic terrorists are doing God's will; radical utilitarians are following ethical prescriptions; and ecoterrorists are saving the planet.[9] (Only idiosyncratic actors are expressly motivated by immoral or amoral aims.) It follows that an individual who discovers a small pocket of academia that describes her or him as a "bad person" who must be stopped at all costs because she or he threatens to prevent the realization of astronomical amounts of moral value could be quite incendiary—thus making the threat worse (see Bostrom, 2003; Torres, 2017g). Similarly, someone of risk type X could take umbrage at being grouped together with someone of risk type Y. An anarcho-primitivist might be outraged at being positioned alongside apocalyptic terrorists in the above typology, even though both satisfy the relevant criterion associated with the doomsday button test. (In fact, I have encountered at least one leading anarcho-primitivist who expressed frustration with my framework for precisely this reason.) The same goes for a radical NU being located alongside a sadistic psychopath like Eric Harris, even though—once again—both would intentionally push a doomsday button under the right circumstances.

A worrisome consequence is that some risky agents could direct their anger at academia in general or certain academics in particular. There is some precedence for this already: Kaczynski targeted universities (in fact, Chomsky was on his list) and ITS has targeted biotechnologists and nanotechnologists, as mentioned above. Furthermore, given the link between agential risk and existential risks, and existential risks and transhumanism, anger could end up being (unjustifiably) generalized to the transhumanist community as a whole. In other words, transhumanists could become a proxy target for agents who, in the near-term, "couple" themselves with conventional weapons like guns and bullets.

Yet another possibility is that knowledge of *other* agents who would destroy the world could lead to an "arms race" of sorts whereby different types or subtypes of agents rush to acquire WTDs and accomplish their goals before anyone else can. The reason is that many of the motivating goals of different agents are mutually incompatible. The realization of a "future primitive" world in which a small human population survives by hunting, gathering, and fishing would make it nearly impossible for radical NUs to accomplish their ambition of, say, weaponizing a particle collider to tip the universe into a more stable, lower-energy state (see Zerzan, 2012). The reverse situation holds as well. It follows that insofar as token agents are instrumentally rational, they will work not only (a) to acquire WTDs to achieve their goals, but (b) to *prevent* other types of agents from acquiring WTDs to achieve *their* goals. A failure to accomplish (b) entails a failure to accomplish the aims of (a).

One might wonder whether anti-risk enforcement operations could leverage inter-agent tensions to mitigate a disaster: agential risks turning against other agential risks could perhaps lower the overall threat level. Unfortunately, we encounter another information hazard here that could result in constructive rather than destructive interference, as it were: the end-goals of many different agents actually *align*. For example, GLF, Eric Harris, and (some) radical NUs all ostensibly want the exact same thing: human extinction.[10] It follows that if these agents knew about each other, they could potentially join forces to bring about an omnicidal attack, where each on its own might not have the resources to do this.[11]

There is also the possibility that work on agential risks encourages individuals who instantiate one of the four risk types to be less open and more secretive about their ideological predilections, thus making it more difficult for anti-risk enforcement operations to

[9] In other words, all three of these categories could perhaps constitute instances of the unilateralist's curse (Bostrom et al., 2016).

[10] I say "some" radical NUs because some wish not merely for our extinction but the annihilation of the entire biosphere or all life that might exist in the universe.

[11] For an interesting article about "unholy alliances," see Ackerman and Bale (2012).

detect and monitor them—an increasingly crucial task given the observation of Rees (2004) and others, mentioned above, that the trends of (T1) and (T2) could enable *single* lone wolves to unilaterally end the party of life for everyone. In other words, the future global security situation will require the prevention not merely of *most* existential attacks, but *all* existential attacks without exception. This suggests that more clandestine scholarship might eventually become desirable, although the feasibility of this is questionable. Universities are not like governments: they are designed to be open rather than closed. Yet perhaps aspects of government secrecy could be implemented within the relevant research communities, which might someday adopt classification schemes that label documents "top secret," "secret," "confidential," or whatever.[12]

Finally, talk about agential risks could lead irresponsible individuals within or outside of academia to propagate misleading, exaggerated, or false statements that incense large groups of people and end up being counterproductive for anti-risk enforcement operations. For example, I have been careful to distinguish between strong/radical NU and the many other versions of NU on the marketplace of consequentialist theories. Indeed, I myself am inclined to adopt a much weaker version of NU—something along the lines of the "suffering-focused" approach advocated by the Foundational Research Institute. Yet it is entirely possible that loud mouths connected to unsubtle minds could lead to the stigmatizing or ostracizing of *anyone* who adopts the NU label. This could be bad for both the larger NU community that poses no possible existential threats even in the context of accessible WTDs and the much smaller community of potentially "dangerous" agents. The same goes for a popular media figure publishing an article about how anarcho-primitivists are "evil idiots" that society would be better off without. This would do absolutely nothing to make the world a better place. A real-world example comes from the neuroscientist Sam Harris (2004), who wrote in an op-ed that "we are at war with Islam." From a "meta-strategic" point of view, this is precisely the sort of inflammatory clash-of-civilizations rhetoric that must be avoided moving forward.

3.2. Info-hazards as triggers

Before concluding this section, let's fulfill our promise from above and consider the issue of triggers. As we will see, some of these triggers are *only really visible* through the prism of agential risks, while others are known phenomena whose existential importance *only becomes clear* when properly situated within the agential risk framework. For the purposes of this paper, let's consider three examples; the extent to which one finds these surprising should further underline the theoretical value of the present approach.

(1) *Date-specific prophecies*. The year 2076 roughly corresponds to the year 1500 in the Islamic calendar (AH). As Islamic scholars have pointed out, we should expect a spike of apocalyptic fervor around this time, just as apocalyptic fervor spiked in 1979, or 1400 AH, as evidenced by the Iranian Revolution and the Grand Mosque seizure. In Muslim traditions, the turn of the century is a significant event marked by the appearance of the *mujaddid*, or "renewer," and indeed the Iranian Revolution was widely seen as an "apocalyptic occurrence" while the Grand Mosque seizure was perpetrated by a group of extremists claiming to have the Mahdi among them. Along these lines, David Cook (2011) conjectures that the year 2039 could witness a dangerous rise of agential risk within the Shi'ite tradition. In his words,

> the 1000 year anniversary of the Mahdi's occultation was a time of enormous messianic disturbance that ultimately led to the emergence of the Bahai faith. . . [A]nd given the importance of the holy number 12 in Shiism, the twelfth century after the occultation 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.

Now imagine the dual-use technologies that will likely exist by 2039 and 2076. Whereas the group behind the Grand Mosque seizure ultimately caused little damage during the two-week-long standoff with the Saudi government, who knows what they might have done differently if they had access to WTDs. The same can be said about the Iranian revolutionaries: for example, Khomeini declared that "it is Allah who puts the gun in our hand. But we cannot expect Him to pull the trigger as well, simply because we are faint-hearted" (quoted in Landes, 2011). If the "gun" had been a world-destroying technology of some sort, how might the Iranian Revolution have unfolded? Could it have become a global rather than merely regional event?[13]

(2) *Space colonization*. Being on the verge of establishing colonies on Mars, especially if Earth-independent, could lead certain agents to plan an imminent attack. The reason is that obliterating civilization or annihilating humanity is easier to achieve while we have all of our eggs in one basket, so to speak, here on Earth. Once humans are living sustainably on the red planet, malicious agents will have to drastically expand the spatial scope of the consequences of their attacks to encompass both Earth and Mars, and this could pose formidable logistical and pragmatic challenges. It follows that chatter about expanding into space in an era of accessible WTDs could lead to frantic action to induce a catastrophe by one or more types of agents. In fact, while scouring the Internet for evidence of omnicidal agents I found discussions on Reddit that mention the importance of destroying the world *before* spaceships carry humans to other worlds. For example, in agreement with a stunning post about the desirability of human extinction, someone responds, "preferably before we spread to other planets" (Torres, 2017b, 2017c). Although space colonization would be an occasion for *relief* among many people concerned about the survival of our species, it should make existential risk scholars *more nervous*.

[12] For discussion of this and related issues, see Rees (2004), pp. 73–89.

[13] Note also that the risk posed by rightwing Christian Identity extremists is highest each year between April 15–20 (i.e., between income tax filing day and Hitler's birthday), meaning that anti-risk enforcement operations should be especially vigilant during this period.

(3) *A non-existential global catastrophe*. It is easier to destroy an army if that army has already been half-destroyed by another enemy. By the same logic, it might be easier to destroy the world *entirely* if the world has already been destroyed *partially*. It follows that a non-existential global catastrophe could provide an ideal opportunity for malicious agents to use whatever destructive power they possess to push civilization over the edge of collapse, or our species over the precipice of extinction. Thus, anti-risk enforcement operations should be especially vigilant of agential risk phenomena after a global-scale disaster—one or more of which may be *highly probable* within the next 100 years. There several reasons to believe this probability claim; here's one: generally speaking, on the expected utility definition of "risk," the probability and consequences of a risk are inversely related. For example, a pandemic that kills 1 million people is more likely than one that kills 1 billion, just as a nuclear conflict that kills 1 billion is more likely than one that kills 7.6 billion. Now consider that expert estimates of *human extinction* this century tend to hover around, perhaps, 20 percent or so (see Torres, 2017f). Insofar as this figure is accurate, the probability of a global catastrophe that does not cause our extinction or irreversibly damage civilization is almost certainly greater than 20 percent. Moving down the scale of the spatial and temporal scope of an event's consequences, the probability continues to climb. Here one might also note that if global catastrophes are random—and studies suggest they are, even anthropogenic catastrophes (see Pinker 2011)—then we should expect them to cluster in time, a phenomenon associated with the "clustering illusion." This gives us even more reason for anticipating a series of unprecedentedly bad events this century.

4. Conclusion

This paper attempts to lay the theoretical foundations for future research on the question of who exactly would destroy the world if only the means were available. This question is increasingly important/urgent because the dual trends of (T1) and (T2) are making the antecedent "if only the means were available" easier to satisfy by a wide range of state and nonstate actors. The answer presented here invokes a six-part typology, four of which designate human agents with omnicidal or anti-civilizational beliefs/desires. We also reviewed some reasons that this approach offers unique insights regarding what scholars and anti-risk enforcement operations ought to focus on as the threat of universal unilateralism grows. The second half of the paper then reviewed a number of information hazards that could be associated with agential risks scholarship, noting in particular the phenomenon of triggers. I do not claim that this paper's theses are correct, nor that the typology adumbrated is exhaustive, only that this should constitute a robust point of departure for subsequent analyses of the topic.

Funding

None.

Acknowledgements

Thanks to Steven Umbrello and Gary Ackerman for helpful comments on a draft of this paper.

References

Ackerman, G., & Bale, J. (2012). The potential for collaboration between islamists and western left-Wing extremists: A theoretical and empirical introduction. *Dynamics of Asymmetric Conflict: Pathways Toward Terrorism and Genocide, 5*(3), 151–171.

Ackerman, G. (2003). Beyond arson? A threat assessment of the earth liberation front. *Terrorism and Political Violence, 15*(4), 143–170.

Anonymous (1989). *Eco-Kamikazes wanted earth first! journal*.

Beckhusen, R. (2013). In manifesto, mexican eco-Terrorists declare war on nanotechnology. *Wired*. https://www.wired.com/2013/03/mexican-ecoterrorism/.

Bostrom, N., Douglas, T., & Sandberg, A. (2016). The unilateralist's curse and the case for a principle of conformity. *Social Epistemology, 30*(4), 350–371.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology, 9*(1).

Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilities, 15*(3), 308–314.

Bostrom, N. (2011). Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy, 10*, 44–79.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy, 4*(1), 15–31.

Bourget, D., & Chalmers, D. (2014). What do philosophers believe? *Philosophical Studies, 170*(3), 465–500.

Burke, J. (2015). *Osama bin Laden's Bookshelf: Noam Chomsky, Bob Woodward, and Jihad. Guardian.* https://www.theguardian.com/world/2015/may/20/osama-bin-laden-library-noam-chomsky-bob-woodward.

Chao, R. (2012). Negative average preference utilitarianism. *Journal of Philosophy of Life, 2*(1), 55–66.

Cook, D. (2011). *Messianism in the shiite crescent.* Hudson Institutehttp://www.hudson.org/research/7906-messianism-in-the-shiite-crescent.

Flannery, F. (2016). *Understanding apocalyptic terrorism: Countering the radical mindset.* New York. NY: Routledge.

GLF (1994). *Statement of Purpose (A Modest Proposal).* 2017. Accessed on July 2 2017 http://www.churchofeuthanasia.org/resources/glf/glfsop.html.

Haija, R. (2006). The armageddon lobby: Dispensationalist christian zionism and the shaping of US policy towards Israel-Palestine. *Holy Land Studies, 5*(1), 75–95.

Harris, S. (2004). *Mired in a religious war.* Washington timeshttp://www.washingtontimes.com/news/2004/dec/1/20041201-090801-2582r/.

Jamieson, A. (2016). *It's disturbing: Author of book found in Munich shooter's home sees pattern.* Guardianhttps://www.theguardian.com/world/2016/jul/23/munich-shooter-why-kids-kill-book-psychologist-peter-langman.

Juergensmeyer, M. (2017). Radical religious responses to global catastrophe. In R. Falk, M. Mohaty, & V. Faessel (Eds.). *Exploring emerging global thresholds: Toward 2030*. Hyderabad, India: Orient BlackSawn.

Kaczynski, T. (1995). *Industrial society and its future. washington post.* http://www.washingtonpost.com/wp-srv/national/longterm/unabomber/manifesto.text.htm.

Kelley, C., Mohtadi, S., Cane, M., Seager, R., & Kushnir, Y. (2015). Climate change in the fertile crescent and implications of the recent syrian drought. *Proceedings of the National Academy of Sciences, 112*(11), 3241–3246.

Landes, R. (2011). *Heaven on earth: The varieties of the millennial experience.* Oxford, UK: Oxford University Press.

Langman, P. (2009a). Rampage school shooters: A typology. *Aggression and Violent Behavior, 14*, 79–86.

Langman, P. (2009b). *Why kids kill: Inside the minds of school shooters.* New York, NY: Palgrave Macmillan.

Lee, M. (1995). *Earth first!: environmental apocalypse.* Syracuse, NY: Syracuse University Press.

Ord, T. (2013). *Why I'm not a negative utilitarian. a mirror clear.* http://www.amirrorclear.net/academic/ideas/negative-utilitarianism/.

Oxford (2017). *Conscience oxford living dictionaries.* Accessed on October 26, 2017 https://en.oxforddictionaries.com/definition/us/conscience.

Pearce, D. (2017). *Negative utilitarianism FAQ.* Accessed on July 2, 2017 https://www.utilitarianism.com/nu/nufaq.html.

Pew (2015). *The future of world religions: Population growth projections, 2010–2050.* http://www.pewforum.org/2015/04/02/religious-projections-2010-2050/.

Pinker, S. (2011). *The better angels of our nature: Why violence has declined.* New York, NY: Viking Books.

Rees, M. (2004). *Our final century: The 50/50 threat to humanity's survival.* London, UK: Arrow Books Ltd.

Roden, D. (2015). *Posthuman life: Philosophy at the edge of the human.* Abingdon, Oxon: Routledge.

Sagan, C. (1994). *Pale blue dot: A vision of the human future in space.* New York, NY: Random House.

Schneider, S. (2016). *It may not feel like anything to Be an alien. nautilus.* http://cosmos.nautil.us/feature/72/it-may-not-feel-like-anything-to-be-an-alien.

Smart, R. N. (1958). Negative utilitarianism. *Mind, 67*, 542–543.

Sotos, J. (2017). *Biotechnology and the lifetime of technical civilizations.* arXiv.org. https://arxiv.org/abs/1709.01149.

Stern, J., & Berger, J. M. (2015). *ISIS: E state of terror.* New York, NY: HarperCollins.

Stout, M. (2005). *The sociopath next door.* New York, NY: Broadway Books.

Tomasik, B. (2016). *Are happiness and suffering symmetric? Essays on reducing suffering.* Accessed on July 2, 2017 http://reducing-suffering.org/happiness-suffering-symmetric/.

Tomasik, B. (2017). *Risks of astronomical future suffering.* Foundational Research Institutehttps://foundational-research.org/risks-of-astronomical-future-suffering/.

Torres, P. (2016a). Agential risks: A comprehensive introduction. *Journal of Evolution and Technology, 26*(2), 31–47.

Torres, P. (2016b). *The clash of eschatologies: The role of end-Times thinking in world history.* Skeptic. September issue https://goo.gl/tHsVxa.

Torres, P. (2017a). Superintelligence and the future of governance: On prioritizing the control problem at the end of history. forthcoming. In R. Yampolskiy (Ed.). *Artificial intelligence safety and security*. New York, NY: Taylor & Francis Group.

Torres, P. (2017b). *Who would destroy the world? Omnicidal agents and related phenomena. forthcoming in aggression and violent behavior*.

Torres, P. (2017c). Latent agential risks past, present, and future: Assessing the threat environment of tomorrow. *Journal of Future Studies* [in press].

Torres, P. (2017d). *Morality, foresight, and human flourishing: An introduction to existential risks.* Durham, NC: Pitchstone Publishing.

Torres, P. (2017e). Moral bioenhancement and agential risks: Good and bad outcomes. *Bioethics, 31*, 691–696.

Torres, P. (2017f). *Are we doomed? fermi pessimism and the great filter.* [Under review].

Torres, P. (2017g). *Space colonization and suffering risks: Why the maxipok rule could have catastrophic consequences.* [Under Review].

Walls, J. (2008). Introduction. In J. Walls (Ed.). *Oxford handbook of eschatology* (pp. 3–22). Oxford, UK: Oxford University Press.

Wittes, B., & Blum, G. (2015). *The future of violence: Robots and germs. Hackers and Drones—Confronting a new age of treat.* New York, NY: Basic Books.

Yudkowsky, E. (2008). Cognitive biases potentially affecting judgment of global risks. In N. Bostrom, & M. Cirkovic (Eds.). *Global catastrophic risks*. Oxford, UK: Oxford University Press.

Zerzan, J. (2012). *Future primitive revisited.* Port Townsend. WA: Feral House.