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.