Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research

Hin-Yan Liu, Kristian Cedervall Lauta* and Matthijs Michiel Maas**

Abstract

In recent years, the study of existential risks has explored a range of natural and man-made catastrophes, from supervolcano eruption to nuclear war, and from global pandemics to potential risks from misaligned AI. What these risks have in common is that they might cause outright human extinction were they to occur. In this approach, such identified existential risks are frequently characterised by relatively singular origin events and concrete pathways of harm which directly jeopardize the survival of humanity, or undercut its potential for long-term technological progress. While this approach aptly identifies the most cataclysmic fates which may befall humanity, we argue that catastrophic 'existential outcomes' may likely arise from a broader range of sources and societal vulnerabilities, and through the complex interactions of disparate social, cultural, and natural processes—many of which, taken in isolation, might not be seen to merit attention as a global catastrophic risk, let alone an existential one.

This article argues that an emphasis on mitigating the hazards (discrete causes) of existential risks is an unnecessarily narrow framing of the challenge facing humanity, one which risks prematurely curtailing the spectrum of policy responses considered. Instead, it argues existential risks constitute but a subset in a broader set of challenges which could directly or indirectly contribute to existential consequences for humanity. To illustrate, we introduce and examine a set of existential risks that often fall outside the scope of, or remain understudied within, the field. By focusing on vulnerability and exposure rather than existential hazards, we develop a new taxonomy which captures factors contributing to these existential risks. Latent structural vulnerabilities in our technological systems and in our (institutional and cultural) societal arrangements (e.g. systemic 'normal accidents'; institutional absence or -failure; cultural distrust of authorities) may increase our susceptibility—the likelihood that we succumb to existential hazards. Finally, different types of exposure of our society or its natural base determine if or how a given hazard can interface with pre-existing vulnerabilities, to trigger emergent existential risks. We argue that far from being peripheral footnotes to their more direct and immediately terminal counterparts, these "boring apocalypses" may well prove to be the more endemic and problematic, dragging down and undercutting short-term successes in mitigating more spectacular risks. If the cardinal concern is humanity's continued survival and prosperity, then focussing academic and public advocacy efforts on reducing direct existential hazards may have the paradoxical potential of exacerbating humanity's indirect susceptibility to such outcomes.

Adopting law and policy perspectives allow us to foreground societal dimensions that complement and reinforce the discourse on existential risks. This holistic taxonomy accordingly enables scholars in the field of existential risk to better recognise the expanded range of existential risks, and helps them to better understand and deploy a more diverse toolbox of law and governance approaches to address these challenges.

Introduction: the definition and framings of existential risk

In recent years, a growing body of scholarship has argued that a new class of risks bears closer study, for their potential extreme impact on the survival of humanity (Bostrom, 2002, 2013; Bostrom & Cirkovic, 2011; Matheny, 2007; Rees, 2004). Prior research has identified a range of

* Associate Professors, Centre for International Law, Conflict and Crisis, Faculty of Law, University of Copenhagen. Correspondence email: hin-yan.liu@jur.ku.dk

** PhD Fellow, Centre for International Law, Conflict and Crisis, Faculty of Law, University of Copenhagen.

such human extinction risks (Bostrom & Cirkovic, 2008; Haggstrom, 2016; Pamlin & Armstrong, 2015), both natural and manmade, including risks from supervolcano eruption, asteroid impact, global warming, nuclear war, as well as more speculative risks from emerging technologies such as biotechnology, high-energy physics experiment disasters, or misaligned artificial intelligence. (C. Sagan, 1983; Asimov, 1981; Smil, 2005; Posner, 2004; Tegmark & Bostrom, 2005; Ord, Hillerbrand, & Sandberg, 2010; Seth D. Baum & Barrett, 2016; Yudkowsky, 2008a; Bostrom, 2014).

While it is encouraging to see greater attention for a critical topic that has long remained understudied, it is relevant to ask how the framing of the field's basic concepts shapes both which problems it identifies and prioritizes, as well as which policy approaches it considers and engages. In his seminal paper, Bostrom defined an existential risk as '[o]ne where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential' (Bostrom, 2002). Thus in Bostrom's view, existential risks are characterised both by their scope (pan-generational) and their intensity (crushing): the size of the group of people who are at risk[1] and how badly each individual within that group is affected, respectively (Bostrom, 2002).

Much prior research on existential risks has thus deployed criteria and methodology which have identified discrete and independent challenges of sufficient severity and pervasiveness to bring about the 'adverse outcome' in a direct causal manner. In this reading, existential risks are an extreme offshoot of global catastrophic risks—disasters which "might have the potential to inflict serious damage to human well-being on a global scale" (Bostrom, 2013; Bostrom & Cirkovic, 2011, p. 2), but which fall short of permanent collapse. While we are not necessarily averse to the Bostromian definition of 'adverse outcomes'—a definition which indeed seems to characterize the space of eventual outcomes to be avoided—we take more issue with the limited range of pathways towards this dreaded outcome-space, which much of the literature has focused on exploring. Specifically, as noted by others in the community, much prior research "has focused mainly on tracing a causal pathway from a catastrophic event to global catastrophic loss of life" (Avin et al., 2018, p. 1). As such there remains an event-focus, in the sense that only discrete events that are causally connected to the demise of humanity within a relatively short time-frame qualify as an existential risk (rather than a 'merely' globally catastrophic one, or a background risk).

Existential risk (re)framings as crucial consideration for law & governance approaches

Distinguishing existential risks as a uniquely threatening outlier along the spectrum of global risks, however, is arguably an unnecessarily narrow framing of the field of study. Indeed, a high-profile 'one-hit-KO' existential risk such as a global nuclear war or a pandemic may constitute only one avenue towards that 'adverse outcome', and concentrating predominately upon (ways to intervene in) its origin and direct pathway, risks overshadowing other potential paths or disaster interaction

[1] Including future generations. Broadly speaking, many scholars in this space share an emphasis on the ethical value of far-future humans (Beckstead, 2013), with some arguing for the absolute prioritization of reducing human extinction risks (rather than risks that destroy civilization but would leave some humans alive) on the grounds that these risks would destroy all future generations (Parfit, 1984, pp. 453–454; C. Sagan, 1983; Ng, 1991; Matheny, 2007). Bostrom himself appears to favour the 'Maxipok' strategy—"Maximise the probability of an 'OK outcome', where an OK outcome is any outcome that avoids existential catastrophe" (Bostrom, 2013, p. 19)—though he takes a slightly broader perspective of mitigating not just 'hard' extinction risks but also 'global catastrophic risks' which could inflict significant, lasting long-term harm to the trajectory of human civilization, and which could thereby end up inflicting other categories of existential risks (including 'permanent stagnation'; 'flawed realisation', or 'subsequent ruination' (Bostrom, 2013, p. 19).

effects[2] that functionally converge towards that same disastrous outcome, even if only indirectly or over longer timescales, with a potentially higher probability. Indeed, as recently noted by scholars in the field, a full mapping of scenarios that lead to catastrophic outcomes "requires exploring the interplay between many interacting critical systems and threats, beyond the narrow study of individual scenarios that are typically addressed by single disciplines" (Avin et al., 2018, p. 2).The precise framing of 'existential risks' is therefore a crucial consideration, informing ethical, strategic, and epistemological (cf. academic) priorities in facing 'adverse outcomes'. This is particularly the case in the context of studying how global political dynamics may interact with certain existential risks, and in formulating meaningfully effective policies and governance approaches to such risks. Of course, this is not to say that the field of existential risk studies has not sought to involve and engage with policy and governance approaches and solutions. Indeed, to its credit, research in the field of existential risks has actively sought to engage with these issues—given that, as Bostrom himself observes (2013:27), global cooperation is critical to mitigating a wide range of existential risks. Likewise, researchers within the 'AI safety' community are beginning to highlight fields such as policy and psychology, as under-represented but potentially promising approaches to addressing risks arising from AI (Brundage, 2017; Kaj Sotala, 2017).

Accordingly, there has been research into the interaction effects between technologies and politics—such as the possibility that arms races might increase the risk that untested, powerful AI systems are deployed rashly or prematurely (Armstrong, Bostrom, & Shulman, 2013; Shulman, 2009). Other work has drawn on cognitive psychology, to study how people might structurally (mis)judge the probability of risks (Yudkowsky, 2008b). Likewise, work exploring policy- and governance approaches to mitigating existential risks has explored policy approaches that include insurance arrangements for large catastrophes (Taylor, 2008); technology taxes and subsidies (Posner, 2008); and work drawing on social (and organizational) psychology to assess ways to motivate AI researchers to choose beneficial AI designs (Seth D. Baum, 2016). Yet, other work has examined the cost-effectiveness of biosecurity interventions (Millett & Snyder-Beattie, 2017); pricing externalities to balance public the risks and benefits of scientific research generally (Farquhar, Cotton-Barratt, & Snyder-Beattie, 2017); and proposing a general international regulatory regime to govern global catastrophic and existential risks from emerging technologies (Wilson, 2013). At present, a majority of existential risk research centres[3] have articulated law and policy research as areas of interest, and scholars in this space have begun to translate such work into concrete proposed policy interventions—notably the 2017 GPP report, which included proposals to develop governance for geoengineering research; establish international scenario plans and exercises for engineered pandemics, and build international attention for existential risk reduction (Farquhar, Halstead, et al., 2017).

Such work is highly encouraging, and the existential risk research agenda has benefited from it. Nonetheless, the risk remains that a too-narrow conception of 'existential risks' prematurely closes down the space of law and governance solutions that are possible—or necessary—in assuring humanity a non-catastrophic future—for instance, a future that, in Bostrom's framing, meaningfully 'maximize[s] the probability of an ok outcome' (Bostrom, 2013, p. 19). However, if human extinction and the persistent and pervasive truncation of technological potential are not completely

[2] Though for some work that examines 'interaction effects' between different global catastrophes, see (S. Baum & Barrett, 2017; S.D. Baum, Maher, & Haqq-Misra, 2013).

[3] While not exhaustive, these include: the Centre for the Study of Existential Risk (CSER), the Global Catastrophic Risk Institute (GCRI), the Global Priorities Project (GPP), the Gothenburg Centre for Advanced Studies, the Global Challenges Foundation, and the Future of Humanity Institute (which has recently announced its 'Governance of AI Program').

homologous, then tailoring our portfolio of policy responses exclusively to closing off the pathways these risks could take—and then calling it a day—would be insufficient. In fact, this might only afford future policymakers with a false sense of security, even as the world continues to reside in an overall state of 'super-risk' (Bermudez & Pardo, 2015).

This is especially the case when there is a narrow 'technological' (re)solution on offer—such as 'improve global vaccine synthesis and production capability', or 'subsidize international technical AI safety research'—which promise to address or prevent the risk at its root. While such direct technological solutions may certainly be indispensable to averting some existential risks, they may not suffice in actually 'plugging all the holes' in our risk space. In a disciplinary context, there is a risk (admittedly self-correcting, given publication incentives) of the research agenda 'halting' early. In a real-world context, the availability of simple, straightforward 'fixes' might even pose a 'moral hazard', if policymakers or global governance systems which lack political will or the attention to explore more complex or costly changes, seize upon the 'symbolic action' of the straightforward, first-order mitigation strategies. Even where this is not the case, certain policy recommendations to mitigate existential risks might depend on too-optimistic a view of institutional rationality or capability.

'Boring Apocalypses': from Existential Hazards, to Existential Risks

While such efforts might mitigate specific existential risks, this might not translate into significantly lowering the overall probability of the 'adverse outcome', if only a part of the problem, or only one problem among many, is addressed. An alternative articulation is that only one path to the 'adverse outcome' is being explored by much research into existential risks: erecting obstacles along that path may indeed reduce the overall likelihood of manifesting these risks, but this might have little impact, or even no effect, upon the manifestation of the 'adverse outcome'.

Thus, our view is that a materialised existential risk (what we call an 'existential hazard') is *sufficient* to lead to an (existentially) 'adverse outcome', but crucially, that this is *unnecessary* to reach that result. If the overarching objective is to lower the probability of human extinction or significant technological curtailment, adopting an array of approaches which complement the mitigation of direct existential risks are required. Within this broad spectrum of aligned approaches, we propose to introduce law, policy, regulatory and governance tools in this paper as an example. The choice of law and policy perspectives is two-fold: on one hand, they make it possible to take second-order considerations, which involve indirect and socially and culturally mediated paths towards 'adverse outcomes' into account; on the other hand, these recognise both the complexity of social organisation and the prospect that civilizational collapse may trigger or possibly instantiate existential outcomes. In this sense, law and policy approaches offer the possibility of complementing and enhancing the narrower approach adopted by contemporary existential risk research, to take into consideration other paths to existentially adverse outcomes; and to better anticipate vulnerabilities, exposures and failure modes in societal efforts to address existential risks.

Exploring the implications of the existential risk framing: risks from AI

An example of this can be drawn from the prospect of super intelligent artificial intelligence (Bostrom, 2014; Yudkowsky, 2008a). Although the landmark research agenda articulated by Russel et al. (2015) does call for research into 'short-term' policy issues, debates in this field of AI risk[4] have—with some exceptions—identified the core problem as one of value alignment, where the divergence between the interests of humanity and those of the superintelligence would lead to the demise of humanity through mere processes of optimisation. Thus, the existential risk posed by the

[4] Cf. (Farquhar, Halstead, et al., 2017). For an excellent overview of recent work (both on technical safety as well as strategy and policy) on mitigating existential risks deriving from artificial intelligence, see (Dawson, 2016, 2017).

superintelligence lies in the fact that it will be more capable than we can ever be; human beings will be outmanoeuvred in attempts at convincing, controlling or coercing that superintelligence to serve our interests. As a result of this framing, the research agenda on AI risk has put the emphasis on evaluating the technical feasibility of an 'intelligence explosion' (Chalmers, 2010; Good, 1964) through recursive self-improvement after reaching a critical threshold (Bostrom, 2014; K. Sotala, 2017; Yudkowsky, 2008a, 2013);[5] on formulating strategies to estimate timelines for the expected technological development of such 'human-level' or 'general' machine intelligence (Armstrong & Sotala, 2012, 2015; S.D. Baum, Goertzel, & Goertzel, 2011; Brundage, 2015; Grace, Salvatier, Dafoe, Zhang, & Evans, 2017; Müller & Bostrom, 2016); and on formulating technical proposals to guarantee that a superintelligence's goals or values will remain aligned with those of humanity—the so-called superintelligence 'Control Problem' (Armstrong, Sandberg, & Bostrom, 2012; Bostrom, 2012, 2014; Goertzel & Pitt, 2014; Yudkowsky, 2008a).[6]

While this is worthwhile and necessary to address the potential risks of advanced AI, this framing of existential risks focuses on the most direct and causally connected existential risk posed by AI systems. Yet while super-human intelligence might surely suffice to trigger an existential outcome, it is not necessary to it. Cynically, mere human level intelligence appears to be more than sufficient to pose an array of existential risks (Rees, 2004; Martin, 2006).

Furthermore, some applications of 'narrow' AI which might help in mitigating against some existential risks, might pose their own existential risks when combined with other technologies or trends, or might simply lower barriers against other varieties of existential risks. To give one example; the deployment of advanced AI-enhanced surveillance capabilities[7]—including automatic hacking, geospatial sensing, advanced data analysis capabilities, and autonomous drone deployment—may greatly strengthen global efforts to protect against 'rogue' actors engineering a pandemic ("preventing existential risk"). It may also offer very accurate targeting and repression information to a totalitarian regimes,[8] particularly those with separate access to nanotechnological weapons ("creating a new existential risk"). Finally, the increased strategic transparency of such AI systems might disrupt existing nuclear deterrence stability, by rendering vulnerable previously 'secure' strategic assets ("lowering the threshold to existential risk") (Hambling, 2016; Holmes, 2016; Lieber & Press, 2017).

Finally, many 'non-catastrophic' trends engendered by AI—whether geopolitical disruption, unemployment through automation; widespread automated cyberattacks, or computational propaganda—might resonate to instil a deep technological anxiety or regulatory distrust in global public. While these trends do not directly lead to catastrophe, they could well be understood as a meta-level existential threat, if they spur rushed and counter-productive regulation at the domestic level, or so degrade conditions for cooperation on the international level that they curtail our collective ability to address not just existential risks deriving from artificial intelligence, but those from other sources (e.g. synthetic biology and climate change), as well.

These brief examples sketch out the broader existential challenges latent within AI research and development at preceding stages or manifesting through different avenues than the signature risk posed by superintelligence. Thus, addressing the existential risk posed by superintelligence is both

[5] For critiques of the 'singularity' claim, see (Brooks, 2014; Dietterich & Horvitz, 2015; Goertzel, 2015; Plebe & Perconti, 2012; Jilk, 2017).
[6] The field of AI safety is particularly active. For a selection of influential papers, see (Amodei et al., 2016; Amodei & Clark, 2016; Christiano et al., 2017; Orseau & Armstrong, 2016; Soares & Fallenstein, 2014).
[7] Notwithstanding interesting developments in 'privacy-preserving' homomorphic encryption configurations, for an interesting exploration of which, see (Trask, 2017).
[8] For a treatment of totalitarianism as a 'global catastrophic risk', see (Caplan, 2008)

crucial to avoiding the 'adverse outcome', but simultaneously misses the mark in an important sense.

Re-examining Existential Risks: Hazard, Vulnerability, and Exposure

While Bostrom's leading typology identifies the general area inhabited by existential risks, it provides little guidance for how to differentiate among the diverse risks within that category (the box marked 'X'), because these risks are not distinguished according to their source, characteristics, or complexity, but only their impact ("crushing") and scope ("pan-generational").[9]

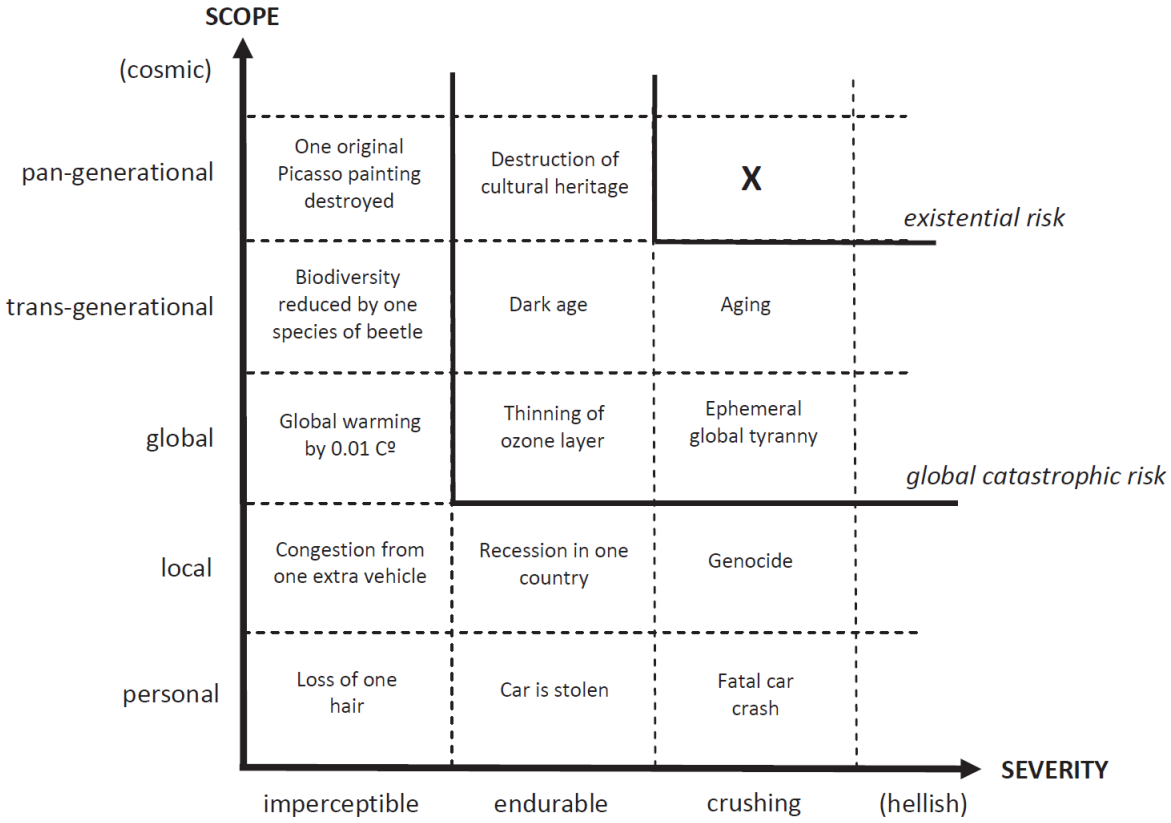

Figure 1. Qualitative risk categories, indicating the relative position of existential risks. (Reproduced from Bostrom, 2013, p. 17)

However, given the range of distinct risks falling within the 'X' box—that is, risks that could cause or feed into an eventual terminal and crushing 'adverse outcome' for humanity—we suggest it relevant to deconstruct existential risks, and instead consider the broader category of 'risks as a function of *hazard*, *vulnerability* and *exposure*':[10]

[9] Of course, Bostrom's objective in setting out this typology is merely to differentiate existential risks from the much larger space of unfortunate occurrences.

[10] This classification schema is distinct from another recently proposed by (Avin et al., 2018), which instead breaks down the analysis of global catastrophic risk scenarios along three different components—(1) a critical system whose safety boundaries are breached by a threat; (2) the mechanisms by which this threat might spread globally to affect the majority of the population, and (3) the manner in which we might fail to prevent or mitigate 1 and 2. While elegant, a discussion of the similarities, differences, and potential (in)commensurability between these two classification taxonomies is out of scope for this present paper.

Existential Risk = Hazard * Vulnerability * Exposure[11]

Here, *hazard* denotes the external source of peril (which is captured within the prevailing agenda studying existential risks)—the 'spark' that threatens the pan-generational/crushing harm.

Vulnerability denotes propensities or weaknesses inherent within human social, political, economic or legal systems, that increase the likelihood of humanity succumbing to pressures or challenges that threaten existential outcomes.

Finally, *exposure* denotes the 'reaction surface'—the number, scope and nature of the interface between the hazard and the vulnerability.

Thus, a hazard is what kills us, and a vulnerability is how we die. Exposure is the interface or medium between what kills us, and how we die. To take an example from disaster studies, a major earthquake only becomes a risk if the built, social or institutional environment can be destabilised during earthquakes of the threatened magnitude ('is vulnerable to"), *and* if such an environment is located in ('exposed to') an earthquake zone. Thus, vulnerability and exposure refer to two different aspects of the affected system: how it breaks, and how it intersects with a given hazard's operating space or pathways of impact.

As a species of global catastrophic risks, the study of existential risks is often conflated with, and perhaps even collapsed into, the identification and mitigation of existential hazards. Where attention is paid to issues of vulnerability and exposure, these are often identified in light of an existential hazard. One of the leading sources and reference points in the field symptomatically organizes the field as a collection of existential hazards (Bostrom & Cirkovic, 2008). A caveat applies for a small subset of hazards of such enormous magnitude that it renders mitigation strategies focussing upon vulnerability and exposure less relevant, or perhaps even irrelevant. The paragon might be the scenarios of 'simulation collapse', or a high-energy physics experiment going awry, altering the astronomical vicinity and rendering life untenable (Ord et al., 2010). Such extreme hazards constitute the archetype of existential risks as a subset of global catastrophic risks and can only be addressed by managing the hazard head-on, with vulnerability and exposure components relegated to marginal roles:

Existential Risk = Existential Hazard * Vulnerability * Exposure

Thus, our claim is not that the field of existential risks research is looking in the wrong places—the emphasis on existential risks has enabled this field to identify a core group of existential hazards which would on their own suffice to bring about the 'existentially adverse outcome'.

Nonetheless, there are also many other, slower and more intertwined ways in which the world might collapse, without being hit by spectacular hazards. To complement the study of existential risks we can draw upon lessons learnt through historical and anthropological studies of civilizational collapse. Thus, while existential risks concentrate upon clear-cut existential hazards, civilizational collapse research infers influential factors that were involved in trajectories of decline. These studies are beginning to challenge the traditional conceptual framework which set out a cyclical

[11] Expressing the interrelationship of several variables, and not a mathematically valid equation. A way of deconstructing risk, common to disaster studies – see e.g. (Wisner, Blaikie, Cannon, & Davis, 2004; Perry, 2007).

history, wherein a civilisations rise and fall, progressing through a predictable pattern of growth, zenith and decline in a gradual manner (Ferguson, 2011). In other words, historically civilizational collapses are boring. Diamond refined this model by recognising that civilizational collapse could be a slow and protracted process emerging from complex interactions (Diamond, 2006).

Beyond Hazards: Vulnerability and Exposure

In this paper, we set out to foreground the other two variables involved in the existential risk equation.

Thus, as noted, 'vulnerability' denotes propensities or weaknesses inherent within human social, political, economic or legal systems that increase the likelihood of humanity succumbing to pressures or challenges that threaten existential outcomes.

'Exposure' indicates the nexus between external hazards and internal vulnerabilities: the interface at which the 'adverse outcome' precipitates from their interaction. Historical studies of civilizational collapses indicate that even small exogenous shocks can destabilise a vulnerable system (Diamond, 2006; Ferguson, 2011). Given this, studying 'exposure' is relevant in systematically analysing interaction effects: a cataclysmic hazard interacting with robust and resilient human systems may be survivable, but conversely, at the interstices at which our human technology, institutions or culture are most vulnerable, even minor (initially 'non-catastrophic') hazards can be the inflection point that tips these susceptible systems towards trajectories of collapse (Gladwell, 2001).[12]

In order to offset the tight coupling between existential risks and existential hazards, we will further dissect the vulnerability and exposure factors introduced in the existential risk calculus. Our proposed taxonomy distinguishes four general categories of vulnerability and exposure (see Table 1).

- **Ontological**: vulnerability through existing in a given location and time in our universe;[13]
- **Passive**: vulnerability through lack of action; 'indirect' exposure;
- **Active**: vulnerability because of insufficient/mis-specified action.
- **Intentional**: vulnerability or exposure knowingly maintained, for that purpose.

Note that for vulnerability, the Passive, Active and Intentional categories correspond to the jurisprudential concepts of 'omission' ('failure to act'), 'negligence' (action, but with failure to exercise the appropriate care to prevent foreseeable future harm) and 'intention' (action with the known purpose to bring about a consequence).

Drawing such distinctions offers the opportunity to be more precise about the features or characteristics which give rise to the existential dimension of the challenge, and thus suggest specific points for targeted intervention, as well as potential failure modes to caution against.

[12] Notably, the resilience of civilization to catastrophes has had some treatment in the field of global systemic risk (Seth D. Baum & Handoh, 2014; Centeno, Nag, Patterson, Shaver, & Windawi, 2015; Helbing, 2013).
[13] Another possible term could be 'anthropic vulnerability'.

	Type of Vulnerability (V) Vulnerability by…	**Type of Exposure (E)** Exposure by…
Ontological (O)	Existence (V-O)	Existence (E-O)
Passive (P)	Omission (V-P)	Indirect link (E-P)
Active (A)	Negligence (V-A)	Direct link (E-A)
Intentional (I)	Intention (V-I)	Intention (E-I)

Table 1: the general categories of vulnerability and exposure, used to structure our taxonomies of existential vulnerability and existential exposure

Below, we combine these categories and their sub-divisions, in twin taxonomies of existential 'vulnerabilities' and 'exposure'. We also seek to give concrete examples. Obviously, not all of these examples are currently unstudied—indeed many feature prominently in the existing literature—though in other cases they remain understudied. While this list is naturally not comprehensive, we hope that such examples enable researchers in the field of existential risks to locate their research in an overarching framework, as well as facilitating links to established scholarly fields which have studied given issues, without considering their bearing on larger existential risks.

A taxonomy of Existential 'Vulnerability'

Our proposed taxonomy for distinguishing between different manifestations of existential vulnerabilities is summarised in Table 2: note that the salience or tractability of these existential vulnerabilities to law and policy approaches increases as one goes down: ontological vulnerabilities appear (at present) highly intractable to mere law and policy—it would be a vain regulator indeed who would try to legislate against physical laws. However, as one proceeds to passive, active, or intended vulnerabilities, the salience of governance approaches increases.

Category	Description	Sub-distinction	Examples of Existential Vulnerabilities
V-O. Ontological vulnerability	Vulnerability that is inherent in being, at present		• Simulation shutdown; • Biological dependence on continuous/frequent energy & resource inputs (including food, water, air, light, …); • Physical dependence on physics integrity; our biochemistry 'works' only within a narrow subset of all possible physical laws (rendering us vulnerable to vacuum decay); • Biological aging.
V-P. Passive vulnerability	Vulnerability existing due to the lack of structures in place. **[OMISSION]**	Built (vulnerability because of the lack of availability of a defence)	• Lack a of super-volcano warning system (technology does not yet exist—lack of global capacity). • Lack of asteroid defence program (existing technology, but not deployed—lack of local capacity at key point);
		Institutional (top-down social vulnerability)	• Lack of effective global institutions, as well as crisis management organisation; • Lack of global coordination on identifying and addressing existential risks. • Lack of public investment in developing critical technologies, e.g. alternate food sources for surviving volcanic winter (Pearce & Denkenberger, 2016) or

			refuges for global catastrophic risks (Haqq-Misra, Baum, & Denkenberger, 2015).
		Cultural (bottom-up social vulnerability)	• Lack of public engagement in confronting existential risks: propensity of public to stereotype/dismiss disaster scenarios ('Terminator headlines'); • Lack of (widely shared) concepts and language to express existential vulnerabilities.
V-A. Active vulnerability	Vulnerabilities existing in spite of / because of the social structures in place. **[NEGLIGENCE]**	Built vulnerability	• Intrinsic path-dependent vulnerabilities in infrastructure *components*: architectural security deficits in universally used components of global (digital) infrastructures (e.g. Spectre and Meltdown exploits in Intel chips); future geo-engineering projects, such as stratospheric aerosol injection, which could backfire heavily if interrupted temporarily, and which might be disrupted (S. Baum, 2015; S.D. Baum, Maher, & Haqq-Misra, 2013). • Intrinsic path-dependent vulnerabilities in infrastructure *configuration*: critical infrastructures (e.g. national electricity grids) are centralized and homogeneous (e.g. rendering society vulnerability to solar flares). • More generally: driven by organizational and competitive optimization. ('Moloch' traps (Alexander, 2014)), globalization homogenizes all solutions across the globe, eroding resilience (e.g. proliferation of homogenized monocultures of staple crops creates vulnerabilities to engineered crop diseases).
		Institutional vulnerability	• Narrow bureaucratic interest and perverse incentives which lock civilization into 'inadequate equilibria' (Yudkowsky, 2017), potentially blocking coordination for known existential risks • Globalised economic and institutional frameworks. Market dependency (Harari, 2015) • Overconfident belief in own ability to foresee risks (Burton, 2008)—risk-based governance and incorrect probabilistic approaches which underestimate fat-tail events.
		Cultural vulnerability	• Spread of pandemics caused by culturally determined interactions (e.g. Ebola); • Ingrained distrust of governmental authorities / public media undercutting disaster response efforts; • Social norms promoting high fertility and unsustainable population growth (Kuhlemann, 2018). • Globalized diets and food demand that can only be met by (unsustainable; vulnerable) monocultures. • Increasingly homogenous global 'monoculture' in practices and ideology creates vulnerabilities, by limiting redundancies and diversity.
V-I. Intended vulnerability	Vulnerability maintained for a direct purpose **[INTENTION]**		• Misaligned, 'apocalyptic' AI (Geraci, 2010); • Nuclear force posture combining centralization of launch command authority, with fallible nuclear early warning systems and 'launch-on-warning' missile force postures (Borrie, 2014). • 'Back-doors' or 'zero-day-vulnerabilities' in critical infrastructure software, knowingly maintained by intelligence services. • Existence of 'omnicidal' agents (Torres, 2016)—including religious groups' faith in end-times, e.g. the Rapture or Yawm ad-Dīn.

Table 2: A taxonomy of vulnerabilities which contribute to existential risks

Ontological Vulnerability

The category of ontological vulnerability denotes intrinsic vulnerabilities associated with human existence. These include the possibility that we inhabit a computer simulation (Bostrom, 2002), which might be terminated or altered at any time. More conceptual and basic vulnerabilities—so fundamental that we often would not even consider them as such—include our existence as biological beings that are dependent (potentially more so than other species such as tardigrades) on continuous or relatively uninterrupted inputs of energy & resources (such as food, water, air, light, …), which renders the human species one comparatively vulnerable to 'extinction' events such as a supervolcano- or meteor-induced global winter. On a deeper level yet, all biochemistry is dependent on the existing laws of physics within which it evolved, rendering us acutely and terminally vulnerable to any processes (e.g. vacuum decay) which would profoundly alter these processes. Biological deterioration due to aging processes, or exterior damages, might also rank amongst these, although that is conditional on whether or not there exists a physical 'hard ceiling' to how far medical senescence research might extend human lifespans and reduce other vulnerabilities.

As these are background conditions at the frontiers of epistemology, we are unlikely to be able to unveil more than a fraction of these vulnerabilities. Also, as inherent features of human existence we have limited abilities to act effectively in this category. Perhaps the most utility we can extract from delimiting ontological vulnerability is to restrict its reach: in other words to leave this as a residual class of vulnerabilities inherent in existence.

Vulnerabilities, Passive and Active; built, institutional and cultural

Passive vulnerabilities are characterised by inaction: the susceptibility to existential outcomes by virtue of failure to take appropriate measures. Conversely, *active* vulnerabilities arise in association with human activities, as by-products or unintended consequences.

Three cross-cutting sub-distinctions can also be made for both passive and active vulnerabilities: built, cultural, and institutional.

Built vulnerabilities are characterised by our (passive) failure to put into place relevant solutions or defences to existential challenges, or by our (active) failure to repair or correct the extant vulnerabilities in the legacy infrastructures we deploy, or the path-dependent ways we deploy them—even if we have such solutions or repairs at our disposal. Such solutions can in fact include some interventions proposed by the existential risk research agenda, such as an asteroid defence programme or the ability to systematically monitor for supervolcano eruptions (Denkenberger & Blair Jr., 2018); they also cover the active existential risks posed by the technologies which humanity has introduced, but which go unfixed—such as architectural deficiencies creating intractable cybersecurity vulnerabilities in universally used computing chips. Because of the technical nature of engineered vulnerabilities, some of these are perhaps closest to the existing (policy) research agenda of the existential risk community—and at present some may consider that law and policy tools have less of a role to play, other than to coordinate efforts aimed at addressing them.

In contrast, top-down vulnerabilities resulting from suboptimal direction and coordination are captured by our sub-category of *institutional vulnerability*. Here, the line between active and passive is admittedly thin, where recklessness can be the distinguishing feature. Active institutional vulnerability may be characterised by failure to coordinate to address a known risk, such as climate change, or cyclical global economic melt-down. Passive institutional vulnerability may then be understood as directional and coordination failures that limit the scope of knowledge related to existential risks—perhaps an implicit 'unwillingness to know', which translates in an unwillingness to fund blue-sky research into charting 'unknown unknowns' (Rumsfeld, 2002).

Cultural vulnerability encompass the bottom-up societal dimensions, reflecting how certain social practices may affect susceptibility to existential challenges. Active cultural vulnerabilities include customary practices that facilitate the spread of pathogens, increasing susceptibility to pandemics, for example integrated commercial travel networks and interpersonal greeting rituals which encourage physical proximity or contact. Passive cultural vulnerabilities include the exclusion or ridicule of existential risks from serious discussion in public forums (let alone the halls of power). This increases collective vulnerabilities insofar as the public and policymakers underrate the prospects for existential risks (cognitive biases exacerbate these effects, Kahneman, 2012) resulting in further marginalisation.

Intended Vulnerabilities

Intended vulnerabilities are those which are created or retained specifically for that purpose, and within the existing research agenda are reflected in the premises of the 'AI risk' or 'Apocalyptic AI' movement (Geraci, 2010). Another salient example can, however, be found in nuclear force postures which (in the US context) features centralization of launch command authority along with a 'launch-on-warning' doctrine that relies on input from fallible early launch warning systems (Borrie, 2014; S. D. Sagan, 1993). Together, this gives rise to the catastrophic risk of an accidental nuclear war (Barrett, Baum, & Hostetler, 2013). Yet far from incidental, this is arguably by design. As the theorist Kenneth Boulding once observed: "if [deterrence] were really stable … it would cease to deter. If the probability of nuclear weapons going off were zero, they would not deter anybody" (Boulding, 1986, p. 32).[14] The nuclear force knowingly renders itself more vulnerable to catastrophic accidents—sacrificing a degree of safety for the sake of strengthening operational readiness and deterrence. While less dramatic, similar intentional vulnerabilities could emerge from a state intelligence service knowingly holding back back-doors or 'zero-day-exploits' which it identifies in critical infrastructure software, in the hope that this may enable more effective cyberattacks against rival states at a later state.

Existential Vulnerability: mitigation and adaptation strategies

This taxonomy of vulnerabilities can provide concrete suggestions for addressing existential risks. While the categories of ontological and intended vulnerabilities may seem superfluous, their treatment as additional classes allow limited resources to be concentrated into the most tractable areas. Perhaps the main contribution of this taxonomy is to highlight how existential risks need not be active and discernible, in the manner of the 'hazards' identified in the field. Instead, many of these risks can be latent, and slow-moving. Moreover, this taxonomy aids in understanding how human activities can impact paths towards 'existential outcomes' in several ways: (1) intent: by directly creating technologies which pose existential hazards (i.e. emerging technologies such as AI, nanotechnology and synthetic biology); (2) (negligence) by establishing complex systems for which failure is unavoidable (Perrow, 2011); (3) and by omission, the failure to take steps to confront existential risks.

Beyond merely refining the sources of existential risks, the contribution of this taxonomy lies in creating a roadmap for the study and integration of risks that have not yet received much or consistent attention in the field of existential risks. In doing so, we emphasise a number of existential vulnerabilities, such as global dependency upon a few species of staple crops, or certain

[14] Indeed, in *Essentials of Post-Cold War Deterrence*, the US Strategic Command recommended a species of potentially risky brinkmanship, arguing that "*[t]he fact that some elements may appear to be potentially 'out of control' can be beneficial to creating and reinforcing fears and doubts in the minds of an adversary's decision makers. This essential sense of fear is the working force of deterrence. That the U.S. may become irrational and vindictive if its vital interests are attacked should be part of the national persona we project to all adversaries.*" (Policy Subcommittee of the Strategic Advisory Group (SAG), 1995)

types of globalised technologies (e.g. SCADA-based systems in critical infrastructure) that are not commonly recognised as sources or failure points of existential risks.

The study of existential 'vulnerability' may suggest that adaptation strategies are preferable to those of mitigation, both because of the inherent complexity underlying both forms of structural vulnerability and because adaptation can now occur simultaneously with mitigation. This is because the vulnerability analysis in effect opens up a parallel system where other trajectories of existential risks are at play. The rough equivalence drawn between traditional existential risks with existential hazards might have the effect of underselling adaptation strategies: it is illogical to conceive of robustness as a defence against the apocalypse, after all. Along with efforts to mitigate or avert existential hazards, however, we can now also plan for adaptation against vulnerabilities. Thus, adaptation strategies are not limited to actions undertaken after 'the Fall': instead they may become rational reactions towards limiting susceptibility to existential risks. In order to explore this potential further, we proceed to examine a taxonomy of exposure.

A taxonomy of Existential 'Exposure'

As a parallel effort to our taxonomy on existential vulnerabilities, we set out a classification system to differentiate between different forms of exposure. It is worth recalling at this point that we use exposure to express the interface between hazards and vulnerabilities—between what kills us, and how we die. Both hazards and vulnerabilities in isolation remain as potentials: exposure is thus a means of actualising such potential into existential risks.

Such exposure can further be directed towards either the societal or the natural environment. This is about what is directly at risk: our (human) society and the common capabilities and support structures preventing existential risks; or nature and its carrying capacity and resilience to future shocks. Thus, we assert that devastating results for humankind can follow from the collapse of both the societal structures we have built, as well as the natural environments within which these constructed systems are embedded. Again, the distinction allows us to single out different examples and trajectories to build alternative strategies for human survival. As is clear from the examples above, it also draws out lessons for existential outcomes which might not be immediately evident from an analysis of existential hazards alone. For example, when 'exposure' is seen from the perspective of the natural environment on which mankind depends, pervasive over-fishing and deforestation, combined with trends in resource demands tracking population growth, may become potentially hazardous activities with the potential to curtail human development in the long run (Diamond, 2006), even if they do not affect most humans directly in the short run.

Category	Description	Sub-distinction	Examples of Existential Exposures
E-O. Ontological exposure	Exposure imposed exclusively by existing (as a human on Earth).	-	• Outer space events; • Super volcanos. • Potential (hostile) alien lifeforms
E-P. Indirect exposure	Exposure indirectly caused by societal arrangements intended for something else.	Exposure of Society	• AI, nuclear power, nanotechnology and synthetic biology. • Experimental scientific curiosity
		Exposure of Nature	• Global extreme climate change. • Over-utilization of nature: unsustainable fishing or hunting

E-A. Direct exposure	Exposure directly caused by societal structures intended for something else.	Exposure of Society	• Lack of political will and institutional inertia leading to 'progress traps' (Wright, 2006). • War, METI, or cultural sentiment. • Unconstrained optimization processes in society, economics, politics (politicians), which pursue originally legitimate goals but become misaligned as they find ways to achieve these in increasingly perverse ways, or with increasing amounts of externalities (cf. 'Moloch' (Alexander, 2014)).
		Exposure of Nature	• Local ecosystem collapse (Kolbert, 2014). • Urbanisation, agriculture and deforestation
E-I. Intentional	Exposure directly imposed by societal arrangements intended precisely for that purpose.	-	• The existence of nuclear and (infectious) biological weapons for strategic purposes such as deterrence. • On a more granular level: the retention of deterrent weapons which risk nuclear winter, over 'winter-safe' deterrent (S.D. Baum, 2015).

Table 3: A taxonomy of modes of exposure which contribute to existential risks

Ontological Exposure

Some exposures are inherent in residing on Earth. Those falling in the category of natural exposure denote existence on earth itself as the exposure, and include our exposure to Near-Earth Objects (NEO) hitting earth or supervolcanoes triggering a protracted volcanic winter. The common denominator underlying this form of exposure is their requirement for measures beyond our present technological capacity to overcome (which admittedly, can be a moving threshold).

Indirect and Direct Exposure

As with the discussion of existential vulnerabilities set out above, the potential of our proposed taxonomy lies in the analysis of indirect and direct exposures. This distinction identifies the exposures that are a direct consequence of human activity, from those that are caused by more complex interactions with other systems.

The theoretical example of high-energy physics research going awry[15] provides an example of societal exposure.[16] A final example of direct exposures are private or unilateral attempts to undertake 'Active SETI'—alternately called METI ('Messaging to Extra-Terrestrial Intelligence') (Zaitsev, 2006)—which might expose the rest of mankind to catastrophic risk, should any future

[15] Cf. (Posner, 2004); for interesting methodological work on estimating the safety of experiments within particle physics, as a particular case of evaluating risks with extremely low probability but very high stakes, see (Ord, Hillerbrand, & Sandberg, 2010).

[16] High-energy physics research can be distinguished from the category of hazards as an example of societal exposure because of the active decision to conduct the relevant experiments. One might also counter that 'nature' would be as much exposed to a potential physics disaster, as our society would be. While that is certainly the case, we treat it as a case of societal exposure insofar as humankind is impacted by such accidents, rather than by the impact of such accidents on the environment's integrity or carrying capacity.

contacted alien species prove hostile and capable of interstellar-scale interdiction. These examples illustrate how surfaces of direct exposure (and ways to reduce it) might be overlooked when concentrating upon the hazard alone.

Beyond direct exposures, an array of arrangements which jeopardise the human societies that have become dependent on them. This category includes any activity or arrangement which might expose the world to extinction through cascading effects. The development of critical common global infrastructures such as the internet, energy markets, and cultural and scientific harmonization might be classified as exposures, rather than vulnerabilities because these reveal new interfaces between hazards and vulnerabilities. Thus collapse of common infrastructures would trigger cascades which jeopardise civilizational sophistication at the global level (Wright, 2006), the edifice upon which humanity's long-term potential has been built. Similarly, developments like urbanisation, intensification of agriculture, and even increasing global inequality (Motesharrei, Rivas, & Kalnay, 2014) appear to be factors that create fault lines and further drive exposures to existential vulnerabilities. Here the exposure perspective shows us that only by certain actions or inactions do risks actually materialise fully against civilization.

Intentional Exposure

Finally, some of these exposures appear to exist intentionally, or at least knowingly or recklessly. The city of New Orleans, Louisiana, provides a microcosm of how dysfunctional behaviour, seen from an existential risk perspective, might be driven by human incentives or rationales operating at different orders. The city is, in design and position, incredibly vulnerable to its natural environment—pinched in between the Mexican Gulf and Lake Pontchartrain and built on the banks of the Mississippi River. Accordingly, some have argued that the most reasonable strategy following hurricane Katrina would have been to abandon the city permanently (Richards, 2011). Instead, the affected populations were given incentives to return, with the US government investing billions in the reconstruction of the city, aware that even with improved defences, the city remains unsafe (Cutter et al., 2014).

Similarly, many populations worldwide, from Tehran and Kathmandu, to San Francisco and Port-au-Prince, persist in known disaster-prone zones, for (legitimate) reasons of culture, history, identity or economy. The purpose of these examples is not to warn the populations of these cities, nor to judge their decision to remain: rather the point is that individuals and societies often make decisions based upon entirely different rationales than a concern for survival. This is an insight that seems to scale to any level of government. In simpler terms, sometimes we choose exposure over safety because of competing considerations, and while this might be productive from a cultural heritage perspective, it remains problematic when seen through the lens of existential risks.

Are Existential Hazards Necessary for Existential Risks?

Having set out taxonomies for differentiating between factors which influence existential risks the question remains whether all components are necessary to bring about an 'adverse outcome'. Our initial claim was that existential hazards could be sufficient existential risks, but that they were not necessary to pose such risks.

Returning to the civilization collapse literature cited above, Ferguson provides a critical insight in contesting the traditional view of cyclical history itself. He posits an alternative conceptual framework by asking the question: 'What if history is not cyclical and slow-moving, but arrhythmic?' (Ferguson, 2011, p. 299). Continuing, he summarises the perspective we adopt succinctly:

> Civilisations… are highly complex systems, made up of a very large number of interacting components that are asymmetrically organised, so that their construction more closely resembles a Namibian termite mound than an Egyptian pyramid. They operate somewhere between order and disorder – on 'the edge of chaos', in the phrase of computer scientist Christopher Langton. Such systems can appear to operate quite stably for some time, apparently in equilibrium, in reality constantly adapting. But there comes a moment when they "go critical". A slight perturbation can set off a "phase transition" from a benign equilibrium to a crisis – a single grain of sand causes an apparently stable sandcastle to fall in on itself. (Ferguson, 2011, pp. 299–300).

Wright echoes this sentiment: 'Civilisations often fall quite suddenly – the House of Cards effect – because as they reach full demand on their ecologies, they become highly vulnerable to natural fluctuations' (Wright, 2006, p. 130). When combined with the observation that hitherto isolated civilizational experiments have now been merged (Harari, 2015), this raises the spectre that existential risks can coalesce from factors that historically brought about only limited civilizational collapses. Thus, the question we need to pose in this regard is whether vulnerabilities themselves contain the seeds of existential risks?

In this context, we should note that vulnerabilities have often been considered mostly as aggravating factors. As aggravators then, vulnerabilities are subsidiary considerations restricted to influencing borderline events: where a potential existential hazard impacts humanity, its susceptibility or resilience could determine whether or not that hazard was transmuted into an existential outcome.

In line with vulnerabilities being developed as a separate sphere where existential risks are at play, this section explores the possibility of removing the existential character of the hazard and thus plausibly reducing the calculus to:

$$\text{Existential Risk} = {}_{\text{Hazard}} * \text{Existential Vulnerability} * {}_{\text{Exposure}}$$

[and/or]

$$\text{Existential Risk} = {}_{\text{Hazard}} * {}_{\text{Vulnerability}} * \text{Existential Exposure}$$

An initial issue is that a catalyst of some sort is required to precipitate the existential risk, because even a system with well-exposed inherent susceptibilities will need something to set it motion. Removing the existential hazard component allows us to explore the possibility that relatively minor occurrences can trigger cascades that emerge as existential risks. But a vulnerability cannot by definition transmute into the existential risk itself absent external input: for this reason we diminish the stature of 'hazard' in the equation to represent our proposition that exogenous shocks need not be the spectacular existential hazards recognised by the study of existential risks. Instead, the external hazards in our revised equation can include insignificant events which go unnoticed (and quite probably involve a large number of minor occurrences).

Contributions and Limitations of Law and Policy Tools for Existential Risks

While our deconstruction of existential risks lead to fairly broad claims, it also provides a few concrete questions and insights. First and foremost, if existential risks can indeed be triggered by non-existential hazards, we need to broaden the scope of investigation in order to draw a more accurate roadmap of the existential risks field, one which can deal with questions of vulnerability and exposure explicitly.

Second, the type of perceived challenge channels the range of appropriate responses which can be developed. While existential hazards may appropriately be met by narrower forms of technical solutions and technologically-oriented mitigation strategies, our broader perspective of existential risks open up other toolboxes to confront existential risks. In particular, social vulnerability and human-driven (anthropogenic) exposure require improved governance and coordination for adaptation strategies. Thus, when we reconstruct existential hazards through the optics of the social systems' inability to withstand them they, per definition, become social phenomena. As noted, many existential risk scholars have recently recognized the importance of reaching out to, and incorporating, law and governance approaches, even where the origin of the existential hazard itself is technological. The critical role of such law and governance approaches should be even more self-evident where the problems in question—the origins of existential vulnerability and exposure—are themselves social, not technological.

This opens up a field for law and governance scholars to work more productively and on an equal footing with technical experts and philosophers. Moreover, this allows for a different set of research questions to be posed as to how we might reduce the vulnerabilities underlying the existential risks against humanity, and our collective exposure to hazards leading to existential outcomes. In doing so, our taxonomy has the potential to elevate relevant aspects of otherwise mundane considerations within politics, economics and society to the plane of existential risks. In garnering this attention, we hope that law and policy tools might be more productively incorporated and deployed as a means to building resilience and robustness. Here, central legal institutions as rights, responsibility and societal relations might in fact contribute substantially to reducing both our vulnerability towards, and exposure to, existential risks.

The obvious limitations of this approach reside in the observation that many contemporary existential hazards, vulnerabilities and exposures are anthropogenic. This raises the spectre of either 'iatrogenesis' ('[complications] caused by the healer'), where our attempts at treating a problem accidentally give rise to new, potentially worse ailments. Thus, in our attempt to curtail existential vulnerabilities and exposures, we may inadvertently generate new or different existential risks. Yet, the framing remains critical: the vantage points created in our proposed taxonomy encourages alternative ways of thinking about existential risks and provide different accommodation strategies.

Finally, the perspective provided by existential vulnerabilities might also foster solutions that will be of more general benefit to humanity as tangential effects of efforts taken to reducing our collective vulnerability and exposure to existential risks. While this appears to be of a lower order of concern at first flush, our taxonomy appears to bind existential risks together with phenomena occurring at different levels. In this sense, existential vulnerabilities and exposures may possess fractal characteristics (Gleick, 1997; Johnson, 2002), reflecting the complexity of their constitution. Support for this claim might reside in the scalability of hazards and vulnerabilities in particular: if pedestrian threats can cascade into existential outcomes, for example, then mundane measures might feedback to reinforce humanity against existential risks. Pushing this to its limits, it is

possible the seemingly oblique effects of improved governance undertaken to shore up existential vulnerabilities actually end up as one of the very sources of humanity's resilience and robustness against existential outcomes.

Concluding Thoughts

The lessons that we can draw from deconstructing existential risks into hazards, vulnerabilities and exposures can be divided into internal and external lessons for the field of existential risk research.

In terms of the lessons for existential risk research, our taxonomy suggests that we may presently reside in a situation of pervasive risk. In identifying the catalogue of existential hazards looming over humanity, and focussing attention to confronting these challenges, the perception is that the outcome of these efforts is a lowering of the overall probability of an actualised existential risk. If our efforts are not actually achieving this, however (because they do not address vulnerabilities or exposures, only direct hazards), we run the risk of achieving safety that is merely 'symbolic': we perceive that we are 'all clear'—that we have successfully steered humanity past 'existential outcomes'—when we are in fact all the more fragile. Defeating a global pandemic, or securing mankind from nuclear war, would be historic achievements; but they would be hollow ones if we were to succumb to social strife or ecosystem collapse decades later. By proposing alternative paths that lead to existential outcomes, our taxonomy can recalibrate the calculus and reduce the prospect of an existential outcome.

Our taxonomy also provides the groundwork for concrete strategies for meeting the existential challenges revealed by our deconstruction of existential risks. In essence, our taxonomy enables more productive cross-disciplinary cooperation amongst researchers from the existential risk community and various other disciplines, in assessing the dynamics that might lead towards catastrophic or 'existentially adverse' outcomes.

This step in itself seems to enhance resilience and robustness by fostering greater variety of policy and governance responses—responses which can move beyond mitigation alone, to extend to adaptation, and which can better anticipate the strengths and weaknesses of governance. Two key limitations latent within such approaches need to be acknowledged. First, that these new perspectives to confronting existential risks import ingrained societal and institutional problems manifest in lower orders of problems. Second, that the additional complexity introduced into the field of existential risks necessarily makes attempts at framing responses more difficult. The payoffs of such a trade-off are open for discussion.

Yet, our deconstruction of existential risks, and the taxonomy we develop to do so, may show promise as tools to help consolidate and expand the field of existential risk research and bring aligned disciplines to bear on the effort to reduce the overall probability of an existential outcome for mankind. But these are early tentative steps to building alternative vantage points from which to examine existential risks: our hope is that the alternative perspectives that these provide will allow researchers in broader fields to bring their expertise to identify trajectories that could lead to humanity's demise, and to devise strategies to obstruct those paths to existential outcomes.

Acknowledgements

The authors thank two anonymous reviewers for driving key elaborations and qualifications in the final version. In addition, Kristian Lauta would like to thank The Centre for the Study of Existential Risks as well.

Bibliography

Alexander, S. (2014). Meditations On Moloch. Retrieved from http://slatestarcodex.com/2014/07/30/meditations-on-moloch/

Amodei, D., & Clark, J. (2016). Faulty Reward Functions in the Wild. Retrieved 18 February 2017, from https://openai.com/blog/faulty-reward-functions/

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. *ArXiv:1606.06565 [Cs]*. Retrieved from http://arxiv.org/abs/1606.06565

Armstrong, S., Bostrom, N., & Shulman, C. (2013). *Racing to the precipice: a model of artificial intelligence development* (Technical Report No. 2013–1) (pp. 1–8). Future of Humanity Institute.

Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines*, *22*(4), 299–324. https://doi.org/10.1007/s11023-012-9282-2

Armstrong, S., & Sotala, K. (2012). How We're Predicting AI--Or Failing To. In J. Romportl, P. Ircing, E. Zackova, M. Polak, & R. Schuster (Eds.), *Beyond AI: Artificial Dreams* (pp. 52–75). Pilsen: University of West Bohemia. Retrieved from https://intelligence.org/files/PredictingAI.pdf

Armstrong, S., & Sotala, K. (2015). How we're predicting AI–or failing to. In *Beyond Artificial Intelligence* (pp. 11–29). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-09668-1_2

Asimov, I. (1981). *A Choice of Catastrophes: The Disasters That Threaten Our World*. New York: Ballantine Books.

Avin, S., Wintle, B. C., Weitzdörfer, J., Ó hÉigeartaigh, S. S., Sutherland, W. J., & Rees, M. J. (2018). Classifying Global Catastrophic Risks. *Futures*. https://doi.org/10.1016/j.futures.2018.02.001

Barrett, A. M., Baum, S. D., & Hostetler, K. R. (2013). Analyzing and reducing the risks of inadvertent nuclear war between the United States and Russia. *Science and Global Security*, *21*(2), 106–133.

Baum, S. (2015, June 5). Is stratospheric geoengineering worth the risk? Retrieved 18 January 2018, from https://thebulletin.org/stratospheric-geoengineering-worth-risk8396

Baum, S., & Barrett, A. (2017). *Towards an Integrated Assessment of Global Catastrophic Risk* (SSRN Scholarly Paper No. ID 3046816). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=3046816

Baum, S.D. (2015). Winter-safe Deterrence: The Risk of Nuclear Winter and Its Challenge to Deterrence. *Contemporary Security Policy*, *36*(1), 123–148. https://doi.org/10.1080/13523260.2015.1012346

Baum, S.D., Goertzel, B., & Goertzel, T. G. (2011). How long until human-level AI? Results from an Expert Assessment. *Technological Forecasting & Social Change*, *78*, 185–195.

Baum, S.D., Maher, J., & Haqq-Misra, J. (2013). Double catastrophe: Intermittent stratospheric geoengineering induced by societal collapse. *Environmentalist*, *33*(1), 168–180. https://doi.org/10.1007/s10669-012-9429-y

Baum, Seth D. (2016). On the promotion of safe and socially beneficial artificial intelligence. *AI & SOCIETY*. https://doi.org/10.1007/s00146-016-0677-0

Baum, Seth D., & Barrett, A. M. (2016). The most extreme risks: Global catastrophes. In V. Bier (Ed.), *The Gower Handbook of Extreme Risk* (p. http://sethbaum.com/ac/2018_Extreme.pdf). Farnham, UK: Gower.

Baum, Seth D., & Handoh, I. C. (2014). Integrating the planetary boundaries and global catastrophic risk paradigms. *Ecological Economics*, *107*, 13–21. https://doi.org/10.1016/j.ecolecon.2014.07.024

Beckstead, N. (2013). *On the overwhelming importance of shaping the far future*. Rutgers University - Graduate School - New Brunswick. Retrieved from https://rucore.libraries.rutgers.edu/rutgers-lib/40469/

Bermudez, J. L., & Pardo, M. S. (2015). Risk, Uncertainty, and Super-Risk. *Notre Dame Journal of Law, Ethics & Public Policy*, *29*, 471–496.

Borrie, J. (2014). A Limit to Safety: Risk, 'Normal Accidents', and Nuclear Weapons. In *ILPI-UNIDIR Vienna Conference Series*. Retrieved from http://www.isn.ethz.ch/Digital-Library/Publications/Detail/?ots591=0c54e3b3-1e9c-be1e-2c24-a6a8c7060233&lng=en&id=186094

Bostrom, N. (2002). Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*, *9*(1).

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, *22*(2), 71–85.

Bostrom, N. (2013). Existential Risk Prevention as Global Priority. *Global Policy*, *4*(1), 15–31.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bostrom, N., & Cirkovic, M. M. (2008). *Global Catastrophic Risks* (1 edition). Oxford; New York: Oxford University Press.

Bostrom, N., & Cirkovic, M. M. (2011). Introduction. In *Global catastrophic risks*. Oxford University Press.

Boulding, K. (1986). Confession of Roots. *International Studies Notes*, *12*.

Brooks, R. (2014, November 10). artificial intelligence is a tool, not a threat. Retrieved 30 April 2017, from http://www.rethinkrobotics.com/blog/artificial-intelligence-tool-threat/

Brundage, M. (2015). Modeling Progress in AI. *ArXiv:1512.05849 [Cs]*. Retrieved from http://arxiv.org/abs/1512.05849

Brundage, M. (2017, June 13). Guide to working in artificial intelligence policy and strategy. Retrieved 14 June 2017, from https://80000hours.org/articles/ai-policy-guide/

Burton, R. A. (2008). *On Being Certain: Believing You Are Right Even When You're Not*. New York: St Martin's Griffin.

Caplan, B. (2008). The Totalitarian Threat. In N. Bostrom & M. M. Cirkovic (Eds.), *Global Catastrophic Risks* (pp. 504–519). Oxford University Press.

Centeno, M. A., Nag, M., Patterson, T. S., Shaver, A., & Windawi, A. J. (2015). The Emergence of Global Systemic Risk. *Annual Review of Sociology*, *41*(1), 65–85. https://doi.org/10.1146/annurev-soc-073014-112317

Chalmers, D. J. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, *17*, 7–65.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *ArXiv:1706.03741 [Stat]*. Retrieved from http://arxiv.org/abs/1706.03741

Cutter, S. L., Emrich, C. T., Mitchell, J. T., Piegorsch, W. W., Smith, M. M., & Weber, L. (2014). *Hurricane Katrina and the Forgotten Coast of Mississippi*. Cambridge: Cambridge University Press.

Dawson, N. (2016, December 13). 2017 AI Safety Literature Review and Charity Comparison. Retrieved from http://effective-altruism.com/ea/14w/2017_ai_risk_literature_review_and_charity/

Dawson, N. (2017, December 20). 2018 AI Safety Literature Review and Charity Comparison. Retrieved from http://effective-altruism.com/ea/1iu/2018_ai_safety_literature_review_and_charity/

Denkenberger, D. C., & Blair Jr., R. W. (2018). Interventions that May Prevent or Mollify Supervolcanic Eruptions. *Futures*. https://doi.org/10.1016/j.futures.2018.01.002

Diamond, J. (2006). *Collapse: How Societies Choose to Fail or Survive*. London: Penguin.

Dietterich, T. G., & Horvitz, E. J. (2015). Rise of concerns about AI: reflections and directions. *Communications of the ACM*, *58*(10), 38–40. https://doi.org/10.1145/2770869

Farquhar, S., Cotton-Barratt, O., & Snyder-Beattie, A. (2017). Pricing Externalities to Balance Public Risks and Benefits of Research. *Health Security*, *15*(4), 401–408. https://doi.org/10.1089/hs.2016.0118

Farquhar, S., Halstead, J., Cotton-Barratt, O., Schubert, S., Belfield, H., & Snyder-Beattie, A. (2017). *Existential Risk: Diplomacy and Governance*. Global Priorities Project. Retrieved from https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf

Ferguson, N. (2011). *Civilization: The West and the Rest*. London: Penguin.

Geraci, R. (2010). *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality*. Oxford University Press.

Gladwell, M. (2001). *The Tipping Point: How Little Things Can Make a Big Difference*. London: Abacus.

Gleick, J. (1997). *Chaos: Making a New Science*. New York: Vintage.

Goertzel, B. (2015). Superintelligence: Fears, Promises and Potentials: Reflections on Bostrom's Superintelligence, Yudkowsky's From AI to Zombies,and Weaver and Veitas's "Open-Ended Intelligence". *Journal of Evolution & Technology*, *24*(2), 55–87.

Goertzel, B., & Pitt, J. (2014). Nine Ways to Bias Open-Source Artificial General Intelligence Toward Friendliness. In R. Blackford & D. Broderick (Eds.), *Intelligence Unbound* (pp. 61–89). John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118736302.ch4/summary

Good, I. J. (1964). Speculations Concerning the First Ultraintelligent Machine. In F. L. Alt & M. Rubinoff (Eds.), *Advances in Computers* (Vol. 6, pp. 31–88). New York: Academic Press.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). When Will AI Exceed Human Performance? Evidence from AI Experts. *ArXiv:1705.08807 [Cs]*. Retrieved from http://arxiv.org/abs/1705.08807

Haggstrom, O. (2016). *Here Be Dragons: Science, Technology and the Future of Humanity* (1 edition). Oxford, United Kingdom ; New York: Oxford University Press.

Hambling, D. (2016). *The Inescapable Net: Unmanned Systems in Anti-Submarine Warfare* (BASIC Parliamentary Briefings on Trident Renewal). British-American Security Information Council. Retrieved from http://www.basicint.org/sites/default/files/BASIC_Hambling_ASW_Feb2016_final_0.pdf

Haqq-Misra, J., Baum, S. D., & Denkenberger, D. C. (2015). Isolated refuges for surviving global catastrophes. *Futures, 72*, 45–56. https://doi.org/10.1016/j.futures.2015.03.009

Harari, Y. N. (2015). *Sapiens: A Brief History of Humankind*. New York: Harper Collins.

Helbing, D. (2013). Globally networked risks and how to respond. *Nature, 497*(7447), 51–59. https://doi.org/10.1038/nature12047

Holmes, J. R. (2016). Sea changes: The future of nuclear deterrence. *Bulletin of the Atomic Scientists, 72*(4), 228–233. https://doi.org/10.1080/00963402.2016.1194060

Jilk, D. J. (2017). Conceptual-Linguistic Superintelligence. *Informatica, 41*(4). Retrieved from http://www.informatica.si/index.php/informatica/article/view/1875

Johnson, S. (2002). *Emergence: The Connected Lives of Ants, Brains, Cities, and Software*. New York: Simon and Schuster.

Kahneman, D. (2012). *Thinking, Fast and Slow*. London: Penguin.

Kolbert, E. (2014). *The Sixth Extinction: An Unnatural History*. London: Bloomsbury.

Kuhlemann, K. (2018). 'Any size population will do?': The fallacy of aiming for stabilization of human numbers. *The Ecological Citizen*, *1*(2), 181–189.

Lieber, K. A., & Press, D. G. (2017). The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence. *International Security*, *41*(4), 9–49. https://doi.org/10.1162/ISEC_a_00273

Martin, J. (2006). *The Meaning of the 21st Century*. London: Eden Project Books.

Matheny, J. G. (2007). Reducing the Risk of Human Extinction. *Risk Analysis*, *27*(5), 1335–1344. https://doi.org/10.1111/j.1539-6924.2007.00960.x

Millett, P., & Snyder-Beattie, A. (2017). Existential Risk and Cost-Effective Biosecurity. *Health Security*, *15*(4), 373–383. https://doi.org/10.1089/hs.2017.0028

Motesharrei, S., Rivas, J., & Kalnay, E. (2014). Human and nature dynamics (HANDY): Modeling inequality and use of resources in the collapse or sustainability of societies. *Ecological Economics*, *101*, 90–102.

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In Müller, Vincent C. (Ed.), *Fundamental Issues of Artificial Intelligence*. Berlin: Synthese Library. Retrieved from http://www.nickbostrom.com/papers/survey.pdf

Ng, Y.-K. (1991). Should we be very cautious or extremely cautious on measures that may involve our destruction? On the finiteness of our expected welfare. *Social Choice and Welfare*, *8*(1), 79–88.

Ord, T., Hillerbrand, R., & Sandberg, A. (2010). Probing the improbable: Methodological challenges for risks with low probabilities and high stakes. *Journal of Risk Research*, *13.0*(2), 191–205. https://doi.org/10.1080/13669870903126267

Orseau, L., & Armstrong, S. (2016). Safely Interruptible Agents.

Pamlin, D., & Armstrong, S. (2015). *Global Challenges: 12 Risks that threaten human civilization*. Global Challenges Foundation. Retrieved from https://api.globalchallenges.org/static/wp-content/uploads/12-Risks-with-infinite-impact.pdf

Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Pearce, J. M., & Denkenberger, D. C. (2016). Cost-Effectiveness of Interventions for Alternate Food to Address Agricultural Catastrophes Globally. *International Journal of Disaster Risk Science*, *7*(3), 205–215. https://doi.org/10.1007/s13753-016-0097-2

Perrow, C. (2011). *Normal Accidents: Living with High Risk Technologies*. Princeton: Princeton University Press.

Perry, R. W. (2007). What is a Disaster? In H. Rodriguez, E. Quarantelli, & R. Dynes (Eds.), *Handbook of Disaster Research* (pp. 1–15). Springer.

Plebe, A., & Perconti, P. (2012). The Slowdown Hypothesis. In A. H. Eden, J. H. Moor, J. H. Søraker, & E. Steinhart (Eds.), *Singularity Hypotheses* (pp. 349–365). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-32560-1_17

Policy Subcommittee of the Strategic Advisory Group (SAG). (1995). *Essentials of Post-Cold War Deterrence*. United States Strategic Command. Retrieved from http://www.nukestrat.com/us/stratcom/SAGessentials.PDF

Posner, R. A. (2004). *Catastrophe: Risk And Response*. Oxford: Oxford University Press.

Posner, R. A. (2008). Public policy towards catastrophe. In N. Bostrom & M. M. Cirkovic (Eds.), *Global Catastrophic Risks* (pp. 164–183). Oxford University Press.

Rees, M. J. (2004). *Our Final Century: Will Civilisation Survive the Twenty-first Century?* London: Arrow.

Richards, E. P. (2011). The Hurricane Katrina Levee Breach Litigation: Getting the First Geoengineering Liability Case Right Essay. *University of Pennsylvania Law Review PENNumbra, 160*, 267–288.

Rumsfeld, D. (2002, June). *Press Conference by US Secretary of Defence, Donald Rumsfeld*. Brussels. Retrieved from http://www.nato.int/docu/speech/2002/s020606g.htm

Sagan, C. (1983). Nuclear War and Climatic Catastrophe: Some Policy Implications. *Foreign Affairs*, *62*(2), 257–292. https://doi.org/10.2307/20041818

Sagan, S. D. (1993). *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons.* Princeton: Princeton University Press.

Shulman, C. (2009). Arms Control and Intelligence Explosions. Presented at the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain. Retrieved from https://intelligence.org/files/ArmsControl.pdf

Smil, V. (2005). The Next 50 Years: Fatal Discontinuities. *Population and Development Review*, *31*(2), 201–236. https://doi.org/10.1111/j.1728-4457.2005.00063.x

Soares, N., & Fallenstein, B. (2014). *Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda* (Technical Report No. 2014–8). Machine Intelligence Research Institute. Retrieved from https://intelligence.org/files/TechnicalAgenda.pdf

Sotala, K. (2017). How feasible is the rapid development of artificial superintelligence? *Physica Scripta*, *92*(11). https://doi.org/10.1088/1402-4896/aa90e8

Sotala, Kaj. (2017, June 5). Cognitive Science/Psychology As a Neglected Approach to AI Safety - Effective Altruism Forum. Retrieved 5 June 2017, from http://effective-altruism.com/ea/1b3/cognitive_sciencepsychology_as_a_neglected/

Taylor, P. (2008). Catastrophes and Insurance. In N. Bostrom & M. M. Cirkovic (Eds.), *Global Catastrophic Risks* (pp. 164–183). Oxford University Press.

Tegmark, M., & Bostrom, N. (2005). How unlikely is a doomsday catastrophe? *ArXiv:Astro-Ph/0512204*. Retrieved from http://arxiv.org/abs/astro-ph/0512204

Torres, P. (2016). Agential Risks: A Comprehensive Introduction. *Journal of Evolution & Technology*, *26*(2). Retrieved from http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=15410099&AN=120541879&h=W6%2FI4Lf30f4AfT%2B3w%2BXJMuAx5hpDMRm8zWXXywbC73bGVB2jv0XMGgrrSyYNbZ7R%2Bhex4RkwH2kO3Ib9dd7azg%3D%3D&crl=c

Trask, A. (2017, June 5). Safe Crime Prediction: Homomorphic Encryption and Deep Learning for More Effective, Less Intrusive Digital Surveillance. Retrieved 8 June 2017, from https://iamtrask.github.io/2017/06/05/homomorphic-surveillance/

Wilson, G. (2013). Minimizing global catastrophic and existential risks from emerging technologies through international law. *Va. Envtl. LJ*, *31*, 307.

Wisner, B., Blaikie, P., Cannon, T., & Davis, I. (2004). *At Risk: Natural Hazards, People's Vulnerability and Disasters*. London: Routledge.

Wright, R. (2006). *A Short History of Progress*. Edinburgh: Canongate Books.

Yudkowsky, E. (2008a). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In N. Bostrom & M. M. Cirkovic (Eds.), *Global Catastrophic Risks* (pp. 308–345). New York: Oxford University Press.

Yudkowsky, E. (2008b). Cognitive biases potentially affecting judgment of global risks. *Global Catastrophic Risks*, *1*(86), 13.

Yudkowsky, E. (2013). Intelligence explosion microeconomics. *Machine Intelligence Research Institute, Accessed Online October*, *23*, 2015.

Yudkowsky, E. (2017). *Inadequate Equilibria: Where and How Civilizations Get Stuck*. Machine Intelligence Research Institute.

Zaitsev, A. (2006). Messaging to Extra-Terrestrial Intelligence. *ArXiv:Physics/0610031*. Retrieved from http://arxiv.org/abs/physics/0610031