Daniel Zimmer
Trond Arne Undheim
Paul N. Edwards
editors

Intersections, Reinforcements, Cascades

Proceedings of the 2023 Stanford Existential Risks Conference

The Stanford Existential Risks Initiative

Stanford University *Palo Alto, California* | 2023

Copyright Information

https://doi.org/10.25740/pn116pv4512

The Stanford Existential Risks Initiative (SERI)
The Center for International Security and Cooperation
The Freeman Spogli Institute
Stanford University

https://seri.stanford.edu/

Encina Hall
616 Jane Stanford Way
Stanford, CA 94305
United States

CONTENTS

Section III: Risk Intersections

Section IV: Governance, Policy Infrastructure, and Scenarios

Conclusion

Introduction

Stanford | Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Introduction

Expanding the Field of Existential Risk Studies

Trond Undheim,[1] Daniel Zimmer [2]

Citation: Undheim, T.A., Zimmer, D., Existential Risks: Emerging Intersections, Reinforcements, Cascades. *Proceedings of the Stanford Existential Risks Conference* 2023, 6-16. https://doi.org/10.25740/wv139gy0377

Academic Editors: Paul Edwards and Steve Luby

Funding: This work was partially funded by Open Philanthropy

Conflict of Interest Statement: The authors declare no conflict of interest.

Informed Consent Statement: All authors included in these proceedings gave their explicit consent to being featured.

Acknowledgements: We thank the rest of the members of the organizing committee of the 2023 Stanford Existential Risks Conference (Steve Luby, Paul N. Edwards, Victor Warlop, Gabe Mukobi, Camille Walker) for their support.

Author Contributions: This introduction is a joint undertaking, with Zimmer composing Sections 1 and 3 and Undheim Section 2.

Abstract: The past two decades have seen anthropogenic global catastrophic and existential risks (x-risks) develop into an increasingly prominent source of academic study, policy focus, and public concern. The proceedings of the third annual Stanford Existential Risks Conference present a diverse array of research by both established and emerging scholars on the past, present, and future of x-risk studies. Some papers sum up the fruits of existing approaches to existential risks, while others address their blind spots or propose potential directions for the future of x-risk studies. Papers vary both by topic—ranging from nuclear risks, biosecurity, ecological collapse, and artificial intelligence and machine learning risks—and by disciplinary approach, featuring works of history, ethics, aesthetics, psychology, political science, sociology, and policy study. Many papers share a common thread in recognizing the importance of cascading risks—particularly potential interactions among acknowledged x-risks, or how cascades of "merely" catastrophic risks could combine to jeopardize human survival. Collectively, they contribute to the task of expanding the field of x-risk studies to encompass insights, methods, and ethical orientations drawn from a more diverse and broadly representative range of academic disciplines.

Keywords: existential risk, AI risk, biorisk, policy, scenarios planning

[1] Research Scholar, Stanford Existential Risk Initiative (SERI), Center for International Security and Cooperation (CISAC), Stanford University, 616 Jane Stanford Way, Encina Hall, Room: C240, Stanford, CA 94305, USA; trondun@stanford.edu.

[2] Postdoctoral Scholar, Stanford Existential Risk Initiative (SERI), Center for International Security and Cooperation (CISAC), Stanford University, 616 Jane Stanford Way, Encina Hall, Room: C240, Stanford, CA 94305, USA; zimmerd@stanford.edu.

1. Introduction

The past century has seen human beings steadily multiply the means of jeopardizing earthy human survival. The massed thermonuclear arsenals of the 1950s introduced novel concerns of "universal death" by artificial radiation (Russell, 2003, 86), while the development of recombinant DNA technology in the 1970s—and the more recent revolution wrought by CRISPR-cas9—have raised the possibility of creating extremely lethal synthetic pathogens. The Earth System science of the 1980s sounded new alarms about the dangers of both nuclear winter and runaway global heating, while the early 2000s witnessed artificial intelligence (AI) emerge as a growing source of existential concern (Bostrom, 2002). Each of these developments has inspired its own moments of intense reflection on what it means to belong to a species capable of authoring its own erasure (Anders, 1956; Rifkin, 1983; Brundtland, 1988; Bostrom, 2014). However, only since the consolidation of concern surrounding artificial superintelligence in the 2000s has the study of anthropogenic existential risks (x-risks) developed the institutional and financial support to grow into a sustained field of inquiry, with dedicated think tanks such as the Future of Life Institute (2014) and academic research centers such as Oxford University's Future of Humanity Institute (2005), the Centre for the Study of Existetial Risk at Cambridge University (2012), and more recently the Stanford Existential Risks Initiative (2019) and the Käte Hamburger Centre for Apocalypse and Post-Apocalyptic Studies at Heidelberg (2019).

Each period of interest in the study of x-risk has been catalyzed by a different kind of risk and marked by the historical and philosophical context in which it arose, whether this be the sense of all-or-nothing existential confrontation that marked the early Cold War (Jaspers, 1961) or the discovery of new degrees of ecological entanglement in the 1980s (Schell, 2000). Today's study of x-risk is no different. Concern about the x-risk of AI first came to the fore in the 1990s among a cadre of transhumanists and futurists who viewed the development of machine superintelligence to be an inherently risky but necessary step towards the colonization of the universe (Bostrom, 2005; More, 2013).[1] These idiosyncracies are perhaps best captured in the highly influential typology of x-risk developed by philosopher Nick Bostrom, which defines *humanity* as "earth-originating intelligent life" (aiming to encompass not just existing biological species, but augmented posthumans and potentially postbiological successors). Bostrom identifies four classes of existential risks—only one of which involves outright human extinction and all of which take as their orienting concern the threat of failing to reach "technological maturity" (Bostrom, 2013, p. 19). As a result, the emerging field of x-risk studies was established on unrepresentatively narrow foundations philosophically (utilitarian population ethics and *longtermism*[2]), methodologically (ranking x-risks based on their statistical likelihood),[3] and definitionally (approaching what counts as both *human* and *extinction* in normatively freighted ways). While much important work has been conducted in this vein, to date many vibrant and long-established streams of inquiry have been left out of the loop, all speaking directly to the philosophical, ethical, political, and policy-related challenges of human survival.

[1] As inspired by the likes of (Dyson, 1988) and (Moravec, 1988).

[2] Or, more precisely *strong longtermism* defined as "the view that impact on the far future is the most important feature of our actions today" (Greaves and MacAskill, 2022).

[3] A necessarily highly conjectural undertaking that requires estimating the likelihood of events that have never happened and may rely on technologies whose feasibility has yet to be demonstrated. A 2020 literature review found that 45% of the 66 sources sampled based their estimations of degrees of x-risk "at least in part on the subjective opinion of the author or others, without direct reference to specific models, data or analytical arguments" (Beard, Rowe, and Fox, 2020, 6). See also: (Schuster and Woods, 2021).

With this history in mind, the Stanford Existential Risks Initiative (SERI) devoted its third annual Existential Risks Conference to the task of helping to broaden the basis of x-risk studies. Conference organizers invited submissions that addressed the past, present, and possible futures of x-risk studies, seeking to showcase the contributions of scholars and practitioners from a broader range of disciplinary and institutional backgrounds than typically feature at x-risk gatherings.

Where most of the work in this nascent field has approached x-risks either singly or as bullet points on an itemized list, the 2023 SERI Conference sought to highlight work that operates at the intersections *among* risks—either addressing (1) how x-risks may interact or (2) how 'merely' catastrophic risks could combine in a cascade that tips humankind toward the brink of extinction. For instance, AI techniques have already generated recipes for thousands of toxins similar to chemical warfare agents; might armed forces, terrorist groups, or even individuals deploy these tools to construct deadly pathogens by modifying genomes? Meanwhile, the now locked-in multiple degrees of global heating threaten not only deadly heat waves and inexorable sea level rise, but also massive disruption to ecosystems, food supplies, and infrastructure integrity. Worse, these developments both heighten the likelihood of catastrophic interstate war – in an age of nuclear, biological, and cyber weapons of mass destruction – and introduce new 'response risks' in the form of heroic geoengineering schemes (Kemp et al, 2022). Although each of these dangers more than warrants a researcher's full attention when taken in isolation, the conference organizers sought to challenge participants to peer beyond their disciplinary silos and consider some of the systemic interconnections, reinforcements, and cascades that compound catastrophes into extinction-level dynamics.

The volume that follows consists of a selection of papers that were presented at the 2023 Stanford Existential Risks Conference. Rather than define x-risk studies narrowly in terms of a particular methodology or a singular definition of *anthropogenic existential risk*,[4] this material suggests that a more constructive path forward would be to approach x-risk research in the pluralistic terms recently articulated by the signers of the "Statement on Pluralism in Existential Risk Studies" (Futerman et al, 2023). This approach invites a diversity of methods and open discussion concerning what human beings are and can be, why their existence is desirable, and what can and should be done to safeguard it. While some might find that this pluralistic approach overly dilutes the focus of x-risk studies, SERI sees the concentration of existential decision making in a few hands as the greater danger. The more that can be done to multiply the experiences, traditions, and perspectives that inform the study of x-risk, the less likely that a subject of this magnitude can be coopted to serve the narrow interests of a few. As public concern continues to mount over the resurgence of nuclear weapons, the teetering of Earth systems, and the growing sophistication of gene editing and AI technologies, one of the greatest *political* dangers of the foreseeable future stems less from the x-risks themselves than from the possibility that some self-appointed savior of humanity may uniliaterally decide to impose their vision of which values make human life worth living and what—or who—may be sacrificed to secure them.

We believe that the purpose of x-risk studies should be to seek not to save humanity for the sake of a particular vision of human flourishing, but to confront planet-scale challenges in a way that maintains the maximum possible range of human expression. Everyone currently alive has chosen, to one degree or another, to hazard another day participating in the collective experiment that is sustaining human existence and deserves a say. In sum, x-risks are everyone's business.

[4] As Émile Torres has recently documented, there are at least a half dozen distinct ways of typologizing human extinction, each with its own distinct philosophical, ethical, and political entailments (2023).

2. Emerging Topics in X-Risk

The 19 papers featured here were selected from the 40 presentations given at the 2023 Stanford Existential Risks Conference. They represent the work of thinkers at all stages of their career operating both inside and outside the academy in areas as diverse as sociology, food science, aesthetics, psychology, political theory, and science and technology studies. The papers cover topics such as risk agency, AI risks, biorisks, cascading planetary boundaries, nuclear winter, collective intelligence, crisis governance, dual-use policy concerns, psychological fragility, epistemic security, and scenario planning. The conference papers were revised in light of two days of rich online and in person discussion conducted during the conference and editorial feedback from the SERI staff. The results have been published here as conference proceedings under a Creative Commons BY-NC-ND license via the Stanford Digital Repository. By doing so, we hope to provide the authors with a stable platform for distributing their ideas as widely and quickly as possible, while also leaving the door open for republishing these works elsewhere.

2.1 Section I: Epistemology, Psychology, and Aesthetics

The papers gathered in Section I address higher level questions concerning how x-risks come to be identified and typologized. They address some of the basic challenges that human psychology and the nature of human knowledge pose for the study of x-risks. Each to varying degrees addresses the epistemological challenges that arise when bending the lens of science backwards to examine how human individual and group dynamics have led the species to the brink of its own self-erasure.

In "Maniacs, Misanthropes, and Omnicidal Terrorists: Reassessing the Agential Risk Framework," Émile P. Torres focuses on agential risks—namely, those that arise when agents *willing* to destroy the world collide with the growing technological *ability* to do this. It builds on the author's previous work to propose a typology of agential risks based on four possible goals: (i) killing a large portion of the human population; (ii) destroying civilization; (iii) triggering global violence and/or an apocalyptic war; and (iv) causing human extinction. The paper adopts a future anterior position to ask "If a global catastrophe were to have happened, who would be most responsible?" and finds that more attention should be paid to not only the *will* to cause global catastrophe, but the available *ways* to achieve this. Torres builds on Luke Kemp's "Agents of Doom" approach to revise their earlier 2018 typology of agential risks to better take account of the means various groups have to act on their apocalyptic desires.

In "Should Epistemic Secuirty Be a Priority GCR Cause Area," Elizabeth Seger addresses epistemic threats such as adversarial influence operations, the erosion of trust in expertise, and extreme polarization. She proposes adopting the term "epistemic security" to describe the resilience of a society's social epistemic systems to detrimental interference from such epistemic threats. She argues that epistemic security should be considered a priority global catastrophic risk (GCR) area not because the current state of epistemic security is particularly dire or may soon become less, but based instad on the ways that diminished epistemic security render a society less capable of identifying and responding to the risks and crises it faces.

In "Psychological Fragility: An Overlooked Existential Risk," Dana Klisanin writes that the fragility of the human psyche is an unexplored area of risk, which has critical implications. She proposes developing an antifragility mindset to counter the mental illness effects that exacerbate pre-existing mental problems, disproportionately affects

women, challenges the less developed coping mechanisms of children and adolescents, challenges those with lower socioeconomic status, minority ethnic backgrounds, those having experienced prior traumatic stress, or post-disaster victims. The antifragile mindset, derived from Taleb (2014), is a type of growth mindset (Dweck, 2019), she writes, anchored in each person's personal character strengths, an active engagement with nature, flexible narratives that are adaptable to circumstances (McAdams, 2018), and the cultivation of anticipatory thinking.

In "Psychological and Psychosocial Consequences of Super Disruptive A.I.: Public Health Implications and Recommendations," David D. Luxton and Eleanor Watson write that as the pace of A.I. development has accelerated rapidly since 2020, a moral panic is burgeoning. The psychosocial impacts of AI deserve attention from a public health perspective. The mechanisms of psychological impacts include threats to interpersonal trust, deception, overreliance on machines in decision-making, and the displacement of work. They present recommendations for addressing these emerging issues, including coaching, public awareness work, best practice, and guidance on ethics, aimed at technology developers, policy-makers, ethicists, healthcare clinicians, politicians, and the general public.

In "Science, Delusion, and Existential Risk," Andrew Nepomuceno writes that the biggest challenge humanity faces right now is the lack of a proper human ecological analysis of the changes taking place in society. The consequence of a perceived mismatch between tightly held beliefs and scientific realities is mass delusions and inaction. His solution is to double down on science but expand the emphasis on values, embracing our 'inner scientist,' and thereby reduce delusion.

In "An Axiology of Aesthetics for Existential Risk," Ishan Raval reconsiders why human extinction would be bad. His paper eschews the utilitarian grounding of most x-risk studies research to reopen the foundational axiological question: What is it for that we want existential risk to be minimized and the human enterprise to continue? It contends that when humans look into what good they want the world to continue to exist for, they find an aesthetic orientation to the way they value things, and from there, an objective and transcendent conception of the good. Such a conception, Raval argues, would be better than that offered by utilitarianism for bringing forth passion toward and about the world with the urgency and intensity existential risks call for. Reminiscent of value theory's ethics of goodness (Murdoch, 2001), his justification, in the end, is experiential, tied to physicality reminiscent of French phenomenologist Merleau-Ponty (Merleau-Ponty and Carman, 2013).

2.2 Section II: Crises in the Earth System

Section II tackles the wider implications of the cascading x-risk framing, specifically considering how the Planetary Boundaries framework (Rockström *et al.*, 2009) both complements and complicates the x-risk domain. The three papers in this section tackle climate change, the biosphere, and begin to map the governance challenges.

In "Navigating Cascading Planetary Boundaries: A Framework to Secure the Future," Thomas Cernev writes that transgressed Planetary Boundaries have the potential to exacerbate known global catastrophic risks such as weapons of mass destruction, and may lead to new and unforeseen risks as a result of the development of new climate change mitigation technology. Cernev's mitigation framework emphasizes the need for cross-institutional involvement to identify, categorize, develop, and implement solutions to cascading effects of such boundaries being broken.

In "Anthropocene under dark skies: The compounding effects of nuclear winter and overstepped planetary boundaries," Florian Ulrich Jehn writes that when the analysis of global catastrophic events occurs in isolation, it may simplify its study, but in reality it hides the cascading effects of interacting risks such as those between nuclear winter and planetary boundaries. His work points towards biosphere integrity as the key planetary boundary affecting nuclear winter survivability.

In "Is Climate Change Ungovernable?," Paul N. Edwards writes that despite the fact that the Intergovernmental Panel on Climate Change (IPCC), an intergovernmental body of the United Nations, seldom writes about scenarios beyond the end of the current century, climate change will not stop in 2100. He believes the evidence points to the likelihood of continuing resistance on the path to net-zero emissions, and to the real possibility of catastrophic, civilization-threatening climate change within the next 2-3 centuries, if not before. That pessimism is borne out of his participation in and observation of the IPCC's activities for decades and his interpretation of government capacity, recent political developments such as the spread of authoritarianism and fascism, the seeming inability to ratchet up the speed of aggressive climate measures, and the real possibility of climate policy backlash.

2.3 Section III: Risk Intersections

Section III tackles the mutual connections between various risks from emerging technologies as it interacts with areas as extensive as public health, international security (IS), and international politics (IP). The five papers in this section don't all cover all technology risks, but do each take into account the fact that cascading risks respect no institutional boundaries.

In "Investigating How Academic Researchers Engage with Dual-Use Biosecurity Research," Sofya Lebedeva writes that academic researchers fruitfully engage with dual-use biosecurity research in various ways, each with risks and benefits. Despite ongoing projects, she feels the misuse of dual-use research of concern (DURC) needs further attention, such as better targeted, intensified, longer-duration educational efforts, more funding, and more attention to instructor excellency and developing a context within which DURC-reducing actions make sense and are validated. Overall, she indicates that while pioneering efforts exist locally, they have not yet scaled to cover the needed terrain in the countries that count or in an overall global picture.

In "Existential Risks Associated with Dual-Use Technologies," Ashok Vaseashta writes that dual-use technologies, traditionally nuclear and biological but also from AI and other exponential technologies, could be useful but are also too easily misused. An open society, for all its benefits, only facilitates such misuse and cybercriminals take advantage. As a consequence, he writes, society needs to curb more, or most dual-use technologies in the future through ethical development, regulatory frameworks, and international cooperation, potentially incorporating traceability efforts and resiliency.

In "Fairness in AI and Its Long-Term Implications on Society," Ondrej Bohdal, Timothy Hospedales, Philip H.S. Torr, and Fazl Barez write that biases in the prediction capabilities of most current AI systems must be mitigated now, or they will only get worse over the long term. The reason: we are currently training AI systems on biased data. Current strategies to improve AI fairness fall short, aggravate stresses such as social unrest, and must be improved if we are to avoid AI's implication in society's collapse. Their top recommendations include evolving the science of iterative bias amplification, developing foundational synthetic datasets, and putting in place continuously evolving AI fairness guidelines and regulations.

In "The Looming Nuclear War," Jean-Pierre Dupuy writes that although the sole purpose of nuclear weapons in contemporary international security and defense policy relates to deterrence, a nuclear war looms nonetheless, simply because so long as they exist, these weapons may one day be used. Dupuy's analysis goes beneath geopolitics to look at nuclear weapons as the formidable and indeterminate tools they are, but also offers criticism of the USA and Russia for leaving the Intermediate-Range Nuclear Forces treaty (INF) in 2009, a treaty that Reagan and Gorbachev had signed in 1987.

2.4 Section IV: Governance, Policy Infrastructure, and Scenarios

Section IV tackles the most challenging aspect of x-risks, which is how they can conceivably could be governed. This is already an emerging issue, but when cascading x-risks at some point soon move to the forefront of geopolitical, national, and local action, it will likely pose an even more debilitating effect on the global system. The four papers cover tweaks that might have to be made to the fabric of democracy, policy, and governance, such as the evocation of additional principles, new institutions, and certainly evolving mindsets.

In "Collective Intelligence as Infrastructure for Reducing Broad Global Catastrophic Risks," Vicky Yang and Anders Sandberg write that even though academic and philanthropic communities have grown increasingly concerned with global catastrophic risks including artificial intelligence safety, pandemics, biosecurity, and nuclear war, outcomes of many, if not all, risk situations hinge on whether governments or scientific communities can work effectively. Prediction can be improved through leveraging committed minorities. Adaptation can be readily improved through utilizing collective memory. In short, x-risks can be seen as Collective Intelligence (CI) problems solvable through facilitating conditions that improve group performance.

In "Convergence on Existential Risk Policy?," Philip Arthur contrasts what has been called the Techno Utopia Approach (TUA) to existential risk, often associated with Bostrom, Ord, and MacAskill, with what he terms a Shorter-term Pluralist Approach (SPA). Pointing to the work of Schuster, Woods, and Torres as examples of SPA, Arthur contends that the two approaches show signs of converging. One example of that convergence is certain types of climate change action measures, such as carbon taxes, which could potentially mitigate both the long-term x-risk of existential climate change and shorter-term social stressors such as migration and inequality in developing nations. Another example is biosecurity, where the cost-effectiveness of mitigating pandemics, biowarfare, and bioterrorism has been shown in a study to be similar to that of traditional health care yet has positive effects both for short-term improvement of global health in the next fifty years and for the far future.

In "Governing Anthropogenic Existential Risks (Envisioning Some New Approaches)," Mariana Todorova writes that her worries are increasing complexity, acceleration of social time and multiplying risks. These crisis-inducing processes could be overcome through de-ideologization of the future. Counterfactuality, defined as conditional statements that trace alternatives to what we ourselves initially are thinking, or what might demonstrably be going on, can become a resource. Each epoch creates the tools for its own critical reflection and anticipation, one tool being building (attractive) scenarios. The key is governability, which our risk society's fluidity (Giddens, 1991; Ekberg, 2007) – the continuous situational rearrangement of risks – renders increasingly problematic. Todorova's example is the layering of the war in Ukraine, energy, and financial crises. The process is worsened by the fact that there's unequal awareness among the stakeholders of those risks. Autopoietic systems (Maturana and Varela, 1980),such as living cells, which

constitute themselves and are capable of self-repair, can also be seen as a metaphor for society (Luhmann and Knodt, 1996; Luhmann, 2005, 2012). However, as Giddens has pointed out, we do not yet have a systematic politics of climate change (Giddens, 2009). Arguably, we need to "strengthen the catastrophic feeling" without resorting to alarmist, one-dimensional thinking of good and bad.

In "The Intergovernmental Panel on Global Catastrophic Risks (IPGCR)," R. Daniel Bressler and Jeff Alstott claim that x-risks are poorly governed at the global scale. They feel this can be migitgate by creating an Intergovernmental Panel on Global Catastrophic Risks (IPGCR), matching and complementing the work of the Intergovernmental Panel on Climate Change (IPCC) and the World Health Organization (WHO) but across cascading x-risk areas. Under United Nations (UN) auspices, the IPGCR's potential mandate would be scientific analysis of such risks, their potential impacts, and the available options for avoiding or mitigating them. Various tweaks on how such an organizations ideally would function imply that it could be vastly more effective at effecting change than previous entities. Notably, scientific participation would extend to the Executive Board and would not simply be advisory but able to make executive decisions concerning management and content, incorporating lessons learned from procedures of the US National Academy of Sciences. Unlike the WHO, the IPGCR would have no operational responsibility, and would be transdisciplinary.

In "Crisis Government's Legitimacy Paradox: Foreseeability and Unobservable Success," Daniel David Slate writes that nearly all prior political theorizing about crisis focuses on imminent or already-present threats, rather than more amorphous forecasted risks. This is a problem if the risk of total societal destruction is tied to uncertain forecasts, foresight, and longer time scales. If interventions succeed, the actual presence of an emergency would then be unobserved. Such a government's very success would prevent the actualization of its legitimacy, presenting a paradox for present theory. Slate contributes a new theory of crisis government that answers how we can legitimately act on the basis of foresight to address the anticipated exigencies of x-risks. Slate expands the scope of the discussion of x-risks in political theory by demonstrating how considerations drawn from the Talmud can help to better address the challenge of nonlinear, cascading crises than can the kinds of emergency politics common to many canonical Western political theorists.

In "Scenarios 2075: The Cascading Risks Study," Trond Arne Undheim writes that humanity faces a myriad of technological, geopolitical, and ecological risks. Studying these separately can miss potentially destructive systemic trajectories. Undheim presents five input scenarios (Climate Cataclysm, World War III, Growth and Collapse, Runaway AI, and Synthetic Biology Unleashed In The Wild) as well as initial findings from an online survey on global systemic risks. The results show that an expert sample (n=145) do not consider 2075 to be a time frame relevant for human extinction, but that they are strongly concerned already about 2225 and beyond. The survey does confirm the relevance of the five disruption factors used in constructing Undheim's scenarios, with heightened concern about technology, particularly biotech and AI, and especially the combination of the two.

3. Invitation

Taken together, the papers gathered here address risk agency, developments in artificial intelligence and machine learning systems, biorisks, overstepped planetary boundaries, ecological collapse catalyzed by nuclear winter, collective intelligence, crisis governance, dual-use considerations, policy concerns, psychological impact, the securitization of risk, and scenario planning. Each maps a small selection of the paths that lead from an

increasingly unstable present to the ultimate catastrophe of human-caused human extinction. These proceedings have foregrounded work that highlights where these pathways intersect and the cascades of mutually-reinforcing risks that threaten to combine global catastrophic risks into extinction-level dynamics. We have also included a wider than usual variety of normative commitments, disciplines, and methodologies in the belief that x-risk studies should contribute not only to averting human extinction, but also to developing a more broad and inclusive dialogue concerning what makes the continuation of human existence worthwhile, what constitutes a genuine threat to these modes of existence, and what risks can reasonably be run to avoid catastrophe in the present and maintain a maximally open future.

A cursory glance at the history of the 20th century reveals the disturbing ease with which attempts to save humankind from extinction can transform into justifications for excluding, expropriating, or even exterminating certain kinds of humans for the sake of the greater good.[5] X-risk studies cannot avoid engaging with the hard choices that will required to balance human flourishing with survival; instead, it is precisely because these choices are so important and unavoidable that they cannot remain the province of a small, homogenous group of self-appointed specialists. It is all too easy to conflate the end of one's way of life with the end of the world as a whole, and it is only by including a plurality of perspectives that x-risk studies can overcome the inevitable parochialism and self-interest that creeps into discussions of who must sacrifice what to ensure human survival and the possibility of flourishing (Cremer and Kemp, 2021). Only a pluralistic field of x-risk studies can hope to identify its own blindspots and credibly resist claims that it conflates the continuation of patriarchy or extractive capitalism or white supremacy or Euro-American imperialism with the survival of humankind as a whole.[6] Here diversity is strength. Significant ethical and methodological disagreement represents both a hallmark of epistemic health and an inoculation against the kind of groupthink that could sanction almost any loss of life or freedom for the sake of survival. Only through the hard work of engaging in sustained dialogue across difference concerning *why* human existence should be preserved can the field of x-risk studies continue to productively address *how* best to diminish today's growing welter of anthropogenic existential risks.

We, the editors of these proceedings, hope that the wide range of papers assembled here will help advance this project. As we work collectively to tackle challenges which, by definition, include everyone, we invite readers to engage with us and with the authors in the spirit of shared understanding, dialogue, and action.

[5] "I would prefer not to see anyone suffer, not to do harm to anyone. But when I realize the species is in danger, then in my case sentiment gives way to the coldest reason," claimed one of the 20th century's most notorious mass murderers (quoted in Schell, 2000, xxi). The disturbing associations that have emerged between some of the founders of contemporary x-risk studies and a resurgence of racist eugenics gives further reason for concern.

[6] For further reflections on the functional value of pluralism in x-risk studies, see Phillip Arthur's contribution to this volume.

References

Anders, G. (1956) *Die Antiquiertheit des Menschen.* Verlag C.H. Beck

Beard, S., Rowe, T., and Fox, F. (2020), 'An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards', *Futures.*

Bostrom, N. (2005) 'A History of Transhumanist Thought', *Journal of Evolution and Technology* 14, pp. 1-25.

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

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

Cremer, C. Z. and Kemp, L. (2021) 'Democratising Risk: In Search of a Methodology to Study Existential Risk' on *SSRN.* https://ssrn.com/abstract=3995225

Brundtland, G.H. (1988) 'Our Common Future: A Climate for Change' in *The Changing Atmosphere: Implications for Global Security.* United Nations Environment Programme World Meteorological Organization.

Dweck, C.S. (2019) 'The Choice to Make a Difference', *Perspectives on psychological science: a journal of the Association for Psychological Science*, 14(1), pp. 21–25.

Dyson, F. (1988) *Infinite in All Directions.* Harper & Row Publishers.

Ekberg, M. (2007) 'The Parameters of the Risk Society: A Review and Exploration', *Current sociology. La Sociologie contemporaine*, 55(3), pp. 343–366.

Futerman, G., Beard, S.J., Sandberg, A., Edwards, Paul N., et al (2023) "Statement on Pluralism in Existential Risk Studies." https://www.existentialriskstudies.org/statement/ (Accessed 1 August 2023).

Jaspers, K. (1961) *The Atom Bomb and the Future of Man.* Translated by E.B. Ashton. University of Chicago Press.

Giddens, A. (1991) *Modernity and Self-Identity: Self and Society in the Late Modern Age*. 1st edn. Stanford University Press.

Giddens, A. (2009) *Politics of Climate Change*. 1st edn. Polity.

Luhmann, N. (2005) *Risk: A Sociological Theory (Communication and Social Order)*. 1st edn. Routledge.

Luhmann, N. (2012) *Introduction to Systems Theory*. 1st edn. Translated by P. Gilgen. Polity.

Luhmann, N. and Knodt, E.M. (1996) *Social Systems (Writing Science)*. 1st edn. Translated by J. Bednarz Jr and D. Baecker. Stanford University Press.

Maturana, H.R. and Varela, F.J. (1980) *Autopoiesis and Cognition: The Realization of the Living (Boston Studies in the Philosophy of Science, Vol. 42)*. First Edition. D. Reidel Publishing Company.

McAdams, D.P. (2018) 'Narrative Identity: What Is It? What Does It Do? How Do You Measure It?', *Imagination, cognition and personality*, 37(3), pp. 359–372.

Merleau-Ponty, M. and Carman, T. (2013) *Phenomenology of Perception*. 1st edn. Translated by D. Landes. Routledge.

Moravec, H. (1988) *Mind Children: The Future of Robot and Human Intelligence.* Harvard University Press.

More, M. (2013) "The Philosophy of Transhumanism" in *The Transhumanist Reader,* Eds. Max More and Natasha Vita-More. John Wiley & Sons, Inc.

Murdoch, I. (2001) *The Sovereignty of Good*. 2nd edn. Routledge.

Rifkin, J. (1983) *Algeny.* Viking Press.

Rockström, J. *et al.* (2009) 'Planetary Boundaries: Exploring the Safe Operating Space for Humanity', *Ecology and Society*, 14(2). Available at: https://doi.org/10.5751/ES-03180-140232.

Russell, B. (2003) "Man's Peril" in *The Collected Papers Papers of Bertrand Russell,* Vol. 28. Routledge.

Schuster, J. and Woods, D. (2021) *Calamity Theory: Three Critiques of Existential Risk.* University of Minnesota Press.

Schell, J. (2000) *The Fate of the Earth and the Abolition.* University of Chicago Press.

Taleb, N.N.N. (2014) *Antifragile: Things That Gain from Disorder (Incerto)*. Reprint edition. Random House Publishing Group.

Section I
Epistemology, Psychology, and Aesthetics

Stanford
Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Should Epistemic Security Be a Priority GCR Cause Area?

Elizabeth Seger [1, 2*]

Citation: Seger, Elizabeth. Should Epistemic Security Be a Priority GCR Cause Area? *Proceedings of the Stanford Existential Risks Conference* 2023, 18-37. https://doi.org/10.25740/bc884qy3778

Academic Editor: Daniel Zimmer

Funding: N/A

Conflict of Interest Statement: N/A

Informed Consent Statement: N/A

Acknowledgments: I would like to thank Giulio Corsi, Seán Ó hÉigeartaigh, Jess Whittlestone, Shahar Avin, Lennart Heim, Ben Harack, Emma Bluemke, Aviv Ovadya, and Ben Garfinkel for their feedback, comments, and conversation leading to the production of this paper.

Author Contributions: N/A

Abstract: "Epistemic security" is a term used to describe the resilience of a society's social epistemic systems to detrimental interference called epistemic threats. Epistemic threats include, for example, adversarial influence operations, the erosion of trust in expertise, and extreme polarization. This paper argues epistemic security should be considered a priority global catastrophic risk (GCR) cause area. The reason is not, as one might expect, that our current state of epistemic security is particularly dire, or that we should expect epistemic threats to get much worse in the future. The reasoning centers instead on a rising "threshold level of epistemic security" which can be understood as the level of epistemic security needed for a society to be capable of effectively preparing for or responding to the risks and crises it faces.

Keywords: epistemic security, global catastrophic risk, crisis response

[1] Research Scholar, Centre for the Governance of AI; Oxford, UK.
[2] Research Affiliate, Centre for the Study of Existential Risk; Cambridge, UK.

* Correspondence: elizseger@gmail.com

1. Introduction : A Quick Primer

1.1 What is Epistemic Security?

An **epistemically secure** society is one in which the processes by which information is produced, distributed, accessed and appraised to inform beliefs and guide decision-making are robust to adverse influence. Access to reliable information is crucial for formulating and coordinating effective and timely responses to complex challenges and crises, for example, combating climate change, voting out politicians who no longer serve the public interest, or halting the spread of a disease through vaccination. However, due to the complexity of social epistemic systems—the interwoven networks of people, processes, technologies, and infrastructures that influence how knowledge is produced, shared, and appraised in a society—a society's capacity for timely and well-informed decision can be undermined by interfering with to those processes. Such detrimental interferences, called *epistemic threats*, include, for example, disinformation campaigns, the erosion of trust in expertise and in knowledge producing institutions, the formation of insular communities, knowledge loss, media censorship, the suppression of diverse viewpoints, extreme polarization, disruptive information mediating and producing technologies and so forth.

The goal of epistemic security research is to better understand how social epistemic processes might be weakened, and to investigate where those systems are susceptible to adverse influence be it adversarial or inadvertent. This concept of "epistemic security" is coined in (Seger et al., 2020).

1.2 Epistemic Security as a GCR amplifier

A Global Catastrophic Risk (GCR) is a hypothetical future event that could significantly damage human well-being on a global scale (Bostrom & Cirković, 2008). There are two ways in which epistemic insecurity (vulnerability to epistemic threats) might lead to a GCR . The first is by epistemic insecurity standing as a GCR on its own. In a worst-case scenario which colleagues and I have called "epistemic babble", the ability for the general population to tell the difference between truth and fiction is entirely lost (Seger et al., 2020). People cannot tell whether anything they see, read or hear is reliable or not, or who, if anyone, is an epistemically trustworthy source of information. A future characterized by complete epistemic uncertainty and void of trust would be a decidedly unpleasant world to live in, so much so that its realization should be considered a global catastrophic risk in and of itself. The second possibility, and the one I focus on here, is that epistemic insecurity acts as a risk amplifier, exacerbating other GCRs making them more likely to occur and/or be of greater severity (Seger, 2022). Mechanisms at play might include one or some combination of following:

- Mis/disinformation leads to suboptimal decision-making toward preventing, preparing for, or responding to crises. The idea here is that high-quality information will inform more effective responses.

- Inconsistent information and breakdowns in trustworthy communications streams undermines effective and timely action coordination. In this case, there may be quality decision-guiding information readily available and easily accessible, but relevant actors struggle to decide who and what to believe .

- Epistemic uncertainty, polarization, and distrust of decision-makers and epistemic authorities reduce capacity for cooperative behavior and undermines effective democratic governance processes. These conditions may, in turn, help lay a path toward table totalitarianism.

This piece focuses on epistemic insecurity as a GCR amplifier. My reasoning is that smaller perturbations in epistemic security can amplify the likelihood or severity of other catastrophic risks via the mechanisms described above before a state of complete epistemic breakdown (epidemic babble) would itself constitute a global catastrophe.

2. What Would Make Epistemic Security a Priority GCR Cause Area?

There are two reasons why epistemic security might be considered a priority GCR cause area.

I. There is sufficiently strong evidence that interventions to improve epistemic security would have a high *direct value* in reducing the risk of GCRs.
II. There is high *informational value* in investigating the impact of epistemic insecurity on other GCRs.

In this section I propose that the status of epistemic security as a priority GCR cause area is derived from both; where there is uncertainty about the direct value of interventions to improve epistemic security, there is high informational value in mounting investigations to remedy those uncertainties.

2.1 Direct Value

GCR interventions have high direct value if they are likely to offer significant and concrete reductions to global catastrophic risk (Askell, 2017). Open Philanthropy's standards for importance, neglectedness, and tractability (INT) speak to the direct value of GCR interventions ('Cause Selection', 2021). If we adopt the INT framework, the below conditions (Table 1) would justify the classification of epistemic security as a priority GCR cause area in terms of direct value.

Table 1: Conditions Justifying epistemic security as a priority GCR cause area in terms of direct value

Importance of Epistemic Security (ES) as a GCR cause area	A. Epistemic Security is declining globally and/or in key globally influential states and this significantly threatens to exacerbate other existential risks, making them more likely to occur and/or be of greater severity (i.e. it significantly increases the probability of other global catastrophes). Or: B. The current state and trajectory of epistemic security globally (whether declining, remaining stable, or improving) are respectively ill-suited to and insufficient for effectively addressing the kinds of complex global risks we face today and/or are likely to face in the future. The consequences of failing to address those challenges are increasingly catastrophic.
Neglectedness	Epistemic security is a neglected cause area overall or includes important neglected components.
Tractability	There are opportunities for impactful intervention: identifiable and achievable efforts to improve epistemic security would reduce the likelihood of GCRs arising from or exacerbated by epistemic insecurity.

Importance:

B adds to A that the importance of epistemic security as a cause area is dependent not only on how vulnerable epistemic systems are to detrimental interference, but also on the nature and severity of the GCRs to which epistemic insecurity may make humanity more vulnerable.

In Section 3 I present a rough analysis of past, present, and future trends in epistemic security (it really can and should be more well evidenced). I posit that, in general, epistemic security in technologically advanced societies has improved immensely over time though may now be leveling off and starting to decline.

If my readers are not convinced by my rough analysis (it is a big job for a small space) and are rather more optimistic about our changing epistemic environment, in Section 4 I illustrate that there is still reason to be concerned about epistemic security as a priority GCR. This is because, I argue, the "threshold level epistemic security"—the level of epistemic security needed for the society to be capable of responding to crises and complex challenges—is rising. On the other hand, if it is the case that current levels of epistemic security are also decreasing, then the case for treating epistemic security as a priority GCR cause area is only made stronger.

Neglectedness:

The topic of epistemic security and threats thereto is very well trodden. There is a large and growing discourse about mis/disinformation (van der Linden et al., 2017), online echo chambers and filter bubbles (Arguedas et al., 2022), fake news (Lazer et al., 2018; Tandoc et al., 2018), the end of expertise (Nichols, 2017), and adversarial influence operations (Thomas et al., 2020). Countless papers, books, blog posts, and news articles have been published about our "misinformation age" (O'Connor & Weatherall, 2019) and "post-truth era" (McIntyre, 2018). Most recently, significant concern has been raised about the impact of AI, especially generative AI capabilities (e.g. large language models (LLMs) and deceptive deepfakes) on epistemic security in particular (Chesney & Citron, 2019; Goldstein et al., 2023; Horvitz, 2022; Sedova et al., 2021).

More work is needed, however, in understanding how a wide variety of threats interact to impact epistemic security. Individual threats, like deepfakes being used to undermine trust in political leaders, are problematic in their own right, but can become more concerning when considered in the context of the broader social epistemic environment. For instance, we can imagine that a society with nascent digital literacy, heavily polarized discourse (perhaps driven by economic inequality or insecurity),[1] and/or low baseline levels of trust in authority and expertise would be more susceptible to the malicious use of deepfakes to sow disinformation, distrust and discord than a society unburdened by those other factors. *It is easy to under-appreciate the disruptive potential of an epistemic threat when it is considered in isolation*.

On this theme, little attention has been dedicated to investigating the relationship between epistemic insecurity and GCRs specifically. One exception I am aware of is Herbert Lin's (2019) short article, *The existential threat from cyber-enabled information warfare*, which investigates the potentiality of a "global information dystopia, in which the pillars of modern democratic self-government – logic, truth, and reality – are shattered."

Finally, more focus is needed on horizon scanning for potential impacts of future technological advances on epistemic security. Horizon scanning efforts—such as Eric Horvitz's (2022) exploration of possible next-generation deepfake capabilities (interactive and compositional deepfakes and adversarial generative explanation)–are key to identifying, preparing for and/or preventing the emergence of new epistemic threats. But even more neglected than future negative impacts of technology on epistemic security are the positive impacts. Let's not forget the potential for developing tools to help improve human capacities for cooperation and collaboration (Horton, 2018; *Introducing the Collective Intelligence Project: Solving the Transformative Technology Trilemma through Governance R&D*, 2023), building trusting relationships (Ovadya & Thorburn, 2023), and appraising the epistemic quality of information and the trustworthiness of information sources (Menick et al., 2022; Ovadya, 2022).

[1] I recommend Goldsworthy, Osborne, and Chesterfield's (G2022) book "Poles Apart" as a comprehensive analysis of the driving factors of affective and ideological polarization.

Tractability:

Without a clear understanding of the dynamics at play by which epistemic insecurity threatens to amplify GCRs, it is difficult to pinpoint opportunities for impactful interventions with a high degree of certainty. The challenge is that various factors influence the epistemic security of a society, such as human psychology, technological capacities, government structure, adversarial actions, political polarization, economic inequality/instability, and so forth. That there are so many factors influencing epistemic security suggests that potential intervention points are numerous and also complexly interconnected. A well-founded discussion of concrete and achievable intervention will require a great deal more research, more than I can offer here.

But this brings me to the second reason epistemic security might be considered a priority GCR; Even where tractability is uncertain—thus casting uncertainty on the *direct value* epistemic security as a GCR cause area—there may still be high *informational value* in investigating epistemic security and effective interventions thereto.

2.2 Informational Value

Insofar as we lack evidence about the current state and future trends in epistemic security and about the direct value of interventions to preserve and protect epistemic security, there may still be high informational value in mounting further investigation.

As described by Amanda Askell (Askell, 2017), the informational value of a cause area is a measure of the expected benefits of gathering information about the cause area compared to the expected costs of doing so. Askell proposes three conditions for high informational value.

I. There is uncertainty about the direct value of interventions.
II. There would be significant benefits to being more certain about direct value.
III. There is a low cost to gathering information (i.e. information itself is not too costly to obtain and delays to action caused by an information collection phase are not too costly).

I have already established (I): given the complexity of social epistemic systems in our technologically advanced age, we lack evidence about current and future trends in epistemic security and about the efficacy of potential interventions.

With respect to (II), we might benefit greatly from being more certain about the direct value of potential interventions both to improve epistemic security and to lower the threshold level of epistemic security needed to effectively manage GCRs. This would be the case if, as I will argue in sections 3 and 4, project trends in realized epistemic security and threshold epistemic security are on a collision course.

(III) is more problematic. If social epistemic systems are as complex and intertwined as I have described, then there is little to no low hanging fruit here. Tractable plans to improve epistemic security could be very costly to obtain, requiring many research hours.

That said, there are many experts whose work is relevant to the improvement of epistemic security. For example, journalists are familiar with fact checking procedures to counter the spread of misinformation, psychologists investigate how individuals consume and evaluate information and form beliefs, information security experts are trained in preventing unauthorized use, disclosure, or alteration to private or sensitive information, and STS (Science and Technology Studies) scholars dedicate their careers to examining the consequences of scientific and technological advancements in historical, social, and cultural contexts.

The challenge is that these experts tend to be siloed in their independent fields, not exploring how their specific insights and contributions to improving epistemic processes and infrastructure interact (a must for improving epistemic security given the complex and interconnected nature of social epistemic systems). On the bright side, this means that there is no shortage of expertise to tackle threats to epistemic security. No new field, as such, needs to be created. What is needed, however, is a substantial organizational effort to identify where relevant expertise exists and to coordinate more holistic approaches to improve epistemic security in our technologically advancing world.

So, overall, while the tractability of improving Epistemic Security is uncertain, *I would say that there is moderate to high informational value in exploring the possibility further if it seems that our current or projected levels of epistemic security may be insufficient to deal with the kinds of global challenges we face today or are likely to face in the future* (see sections 3 and 4).

3. Epistemic Security in Flux

In this section I sketch out my initial intuitions about how epistemic security has changed over time in technologically advanced societies, and how we might expect it to change in the future.

Please note that the current state and trajectory of epistemic security in society is difficult to appraise due to the complexity of social epistemic systems and the various factors affecting them. There are also no well-defined metrics for measuring epistemic security. If ES does seem like a promising candidate for priority GCR status, building unified methodologies (taxonomies and metrics) for analyzing GCR will be an important port of call.

Before I continue, I should warn that this section is long. *My overall conclusion is that, in general, epistemic security in technologically advanced societies has improved immensely over time, however may now be leveling off if not starting to decline;* with increased complexity in social epistemic infrastructures have come numerous new points of vulnerability and threats to our social epistemic systems that have co-emerged with advances in information mediating and producing technologies. That said, epistemic security remains much improved from the long- and medium-term past mainly due to increased access to information, public education, and scientific and technological advances that have given us a better understanding of our world.

A rough, back-of-the-envelope sketch of how the state of epistemic security has changed over, say, the last 500 years, might look something like this (Figure 1):

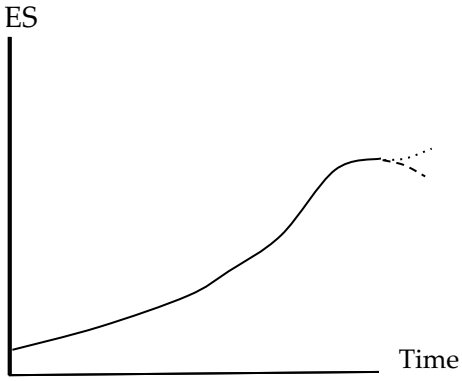


Figure 1: Estimated trend in epistemic security over the past 500 years

If this feels about right to you, then you may wish to skip to section 4 on "the rising epistemic security threshold". I define *"epistemic security threshold"* as the level of epistemic security needed for a society to be capable of effectively preparing for or responding to the risks and crises it faces. Section 4 emphasizes that the importance of epistemic security as a GCR cause area is not only dependent on the state of epistemic security in a society, but also on how our epistemic capacities measure up to the task of responding to the kinds of GCRs we face today and are likely to face in the future.

What follows in the remainder of this Section 3 is a more detailed walkthrough of my reasoning about the current state and trajectories of epistemic security sketched above. To make the problem more tractable, I break my evaluation of epistemic security into what I like to think of as three key features of epistemic security within a society: information accessibility, information environment safety, information recipient sensitivity.[2]

3.1 Information Accessibility

Information accessibility describes how widely available information is on a given topic and how easily it is for information recipients to consume that information. Information accessibility can be thought of as a prerequisite for epistemic security. If a central aim of maintaining epistemic security in a society is to facilitate informed decision-making, then information must be accessible in the first instance.

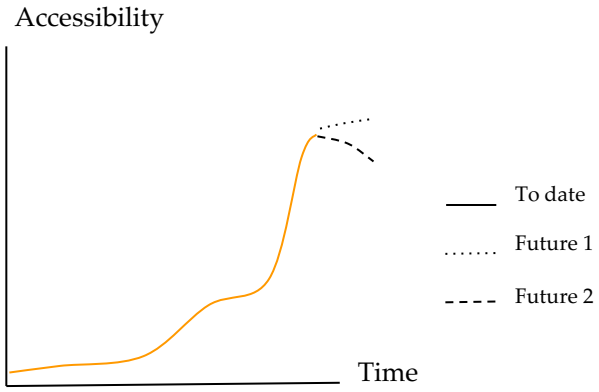
There are various factors influencing information accessibility. For example, technological advances have dramatically improved in the availability and accessibility of information. The printing press, radios, telephones, and the internet are just to name a few. Improvements to public education and the corresponding rise in literacy rates also improves information accessibility. On the other hand, factors like media censorship or restrictions imposed on the use of information sharing technologies and/platforms (e.g. internet, social media, etc.) would reduce information accessibility.

[2] My conceptualization of this breakdown benefited greatly from conversation with Giulo Corsi.

How information accessibility has been/is changing:

Overall Trend:

- Strong increase over time though possibly flattening out in modern times.



Initial increases:

- Largely driven by technological advances enabling rapid and large scale information dissemination and exchange.
- Increased prevalence of democratic governments with protections for free speech may also be a factor.
- Jump 1: printing press, books, pamphlets combined with public education and increased literacy rates
- Jump 2: internet (email / instant messaging / social media)

Future trends

- On a global scale I expect minor improvements to accessibility to be made by improving access to information technologies for those currently with limited access. So, I see accessibility flattening out. This may, however, be a limitation of my imagination (future 1)
- The emergence of more authoritarian/totalitarian states resulting in associated information control and censorship leads to decrease in accessibility (future 2).

3.2 Information environment Safety

Information environment safety has to do with the quality of information that is available to consumers. If one were to measure information environment safety, it could be thought of as the ratio of reliable, truth-tracking information on a topic to false or misleading information. The safer an information environment is, the more likely information consumers are to encounter truth-tracking information.

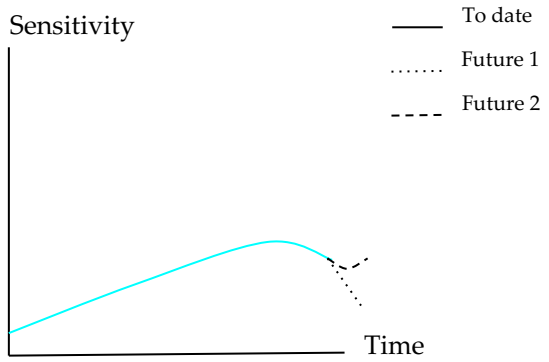
Again, there are various influences on information environment safety that one might explore. For example, scientific advances improve the quality of information we have about the world. We can in turn use that information to make more well-informed decisions about how we respond to the challenges we face (e.g. climate change, pandemics, etc.). On the other hand, malicious actors enabled by modern information producing and mediating technologies mount influence operations spreading

intentionally misleading disinformation to sway public opinion, sow confusion, and seed discord. This is a clear detriment to information ecosystem safety.

How info environment safety has been/is changing:

Overall Trend:

- Strong upward trend through history with more recent flattening out/downward dip.



Initial Increase:

- Since the enlightenment, scientific research coupled with technological advance has provided humanity with ever increasingly epistemically well-founded information about our world.

Level-off:

- Propaganda, influence operations, and disinformation have always been present. Their downward pull prevents the historical upward trend in information environment safety from being as steep as it might have been. However, recent advances in information producing and disseminating technologies has decreased costs to epistemic adversaries in both widely disseminating intentionally misleading information and targeting that information at the most easily influenced individuals. While information environment safety is still much higher than historical levels I think this upward trend has begun to flatten as disinformation is more rapidly produced and effectively targeted at individuals.
- Faster communication and ease of "posting" means less thought goes into content production and distribution. Journalists developed strong norms for producing reliable content that are not now adhered to by most content producers.
- Generative AI can fake information. The very existence of the tech undermines trust in otherwise legit information sources and evidence sources. CSET has produced quite a comprehensive report on the potential impacts of AI on the spread of misinformation and disinformation (Sedova et al., 2021).
- Economic incentives for social media sites value engagement over truth-tracking.

Future trends:

- Efforts are made to reduce spread of mis/disinformation, to intervene on intentional and detrimental influence operations, and to reduce factors threatening democracy (future 1)
- Without intervention, costs to epistemic adversaries continue to decrease. The information environment becomes increasingly unsafe. Epistemic uncertainty triggers human "groupish" response which drives increased polarization. Rise of stronger totalitarian governments/movements that carefully cater information diets (future 2)

3.3 Information Recipient Sensitivity

Information recipient sensitivity refers to the capacity of information consumers to distinguish between epistemically well-founded/truth-tracking information and false/misleading information and between epistemically trustworthy information sources and untrustworthy information sources. The less safe an information environment is, the more important it is that information recipients are sensitive to the quality of information that they consume.

Examples of factors influencing information recipient sensitivity include:

- **Innate limitations of human psychology** - For example, people are predisposed to believe things that a large number of people already appear to believe or that members of one's "in-groups" (those people who supposedly share your interest and values) believe. These are not necessarily truth tracking heuristics.

- **Polarization** - When individuals are entrenched in a particular ideology, viewpoint or in-group, they may be less likely to seek out or consider information from alternative perspectives. This can lead to a lack of critical evaluation and a willingness to accept information that supports existing beliefs, even if it is false or misleading.

- **Technological factors** - For example, content recommendation algorithms on social media platforms can create "filter bubbles" that reinforce pre-existing beliefs and biases, making people less sensitive to alternative viewpoints and potentially reducing their ability to critically evaluate information quality. More recently, generative and persuasive AI capabilities have been of particular concern. Generative AI can be used to create realistic-looking news articles, images, or videos that are entirely fabricated. This can make it more challenging for people to determine the quality and accuracy of information they are consuming, particularly if the content appears to be from a reputable source. Furthermore, if persuasive AI capabilities continue to improve this will likely undermine information recipient sensitivity by helping malicious actors to more effectively convince people of the quality of information they are receiving. Technological factors can, of course, also have positive impacts on information recipient sensitivity. For example, automated fact checking engines or content contextualization engines can help people quickly confirm claims and/or situate

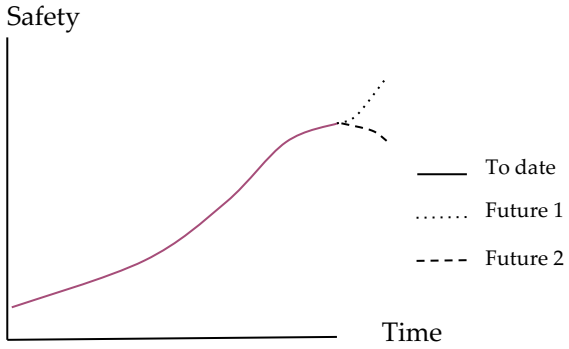
claims within broader discussions on the subject (e.g. What other sources communicate similar ideas? Are these sources diverse, or do they all parrot an original source?)(Ovadya, 2022).

- **Public education & digital literacy** - provides individuals with the knowledge and skills necessary to evaluate information and information source quality. For instance, people more familiar with generative AI capabilities may be more wary of digital information sources and be able to employ strategies to appraise the reliability of the information source.

How info recipient sensitivity has been/is changing:

Overall Trend:

- Shallow incline; minor improvements over time with moderate to sharp downturn in modern times



Initial improvements:

- Public education: allows people to make more well informed decisions about what information is likely true and what information sources are epistemically trustworthy. Increased prevalence of democratic governments: well-functioning processes of public deliberation are less volatile than authoritarian decision-making and more likely to deal in truth tracking and value-aligned information.

Flattening out / downturn

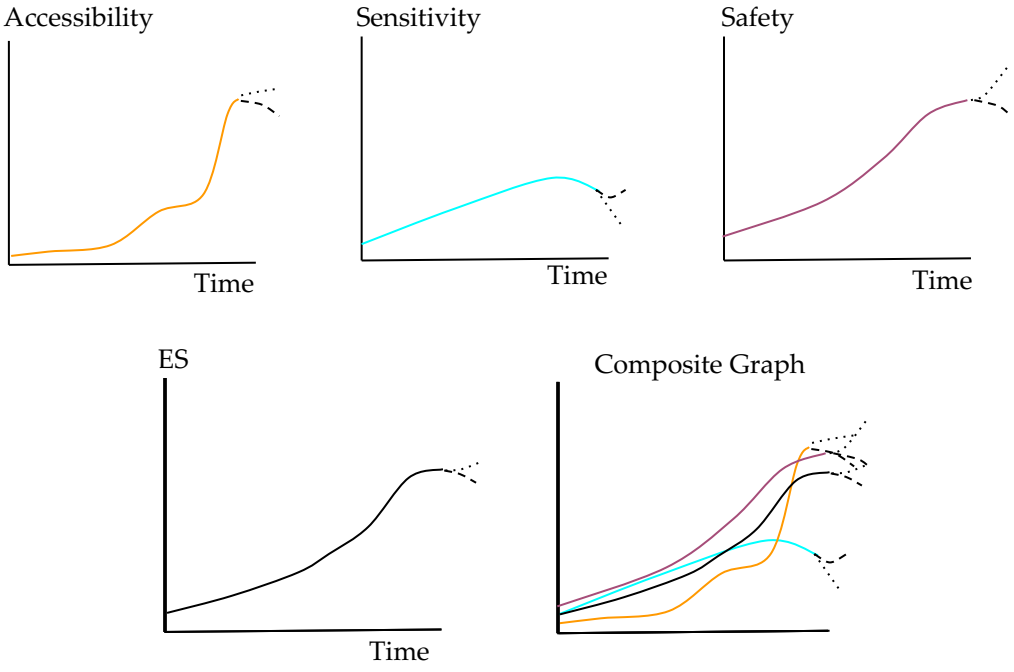
- Increased polarization in major democratic countries negatively impacts decision-making capacities of individuals and groups within the society
- Decline of public of trust in expertise
- Information producing and distributing technologies mislead and/or can be used to hijack natural trust heuristics people use to identify epistemically trustworthy information sources. Attention scarcity has always been an issue, but the evolutionarily ingrained heuristics that reduce cognitive load and help us sift information may be particularly ill-suited to technologically advanced information environments.

Today/ Future trends

- Steep improvements are unlikely due to inherent limitations of human psychology, namely attention scarcity and bounded rationality.
- Without intervention, the downward trend will likely continue toward an "epistemic babble" scenario in which people completely lose the ability to tell fact from fiction to direct epistemically robust decision-making. (future 1)
- Technological innovation may aid human sensitivity by reducing cognitive effort of information / information source evaluation (e.g. contextualization, fact checking, inoculation, flagging artificially generated content etc.) (future 2)

3.4 The Overall Effect on Epistemic Security

Let's combine the back-of-the envelope sketches from above. The overall trend in epistemic security over, say, the last 500 years might look something like this:



My intuition is that over time epistemic security has gradually improved, ultimately coming leaps and bounds due to increased access to information, public education, and scientific and technological advances that have given us a better understanding of our world.

However, more recently, epistemic security has leveled-off and decreased steadily since the late 1980s early 1990s after widespread adoption of the internet. With increased complexity in social epistemic infrastructures have come additional points of vulnerability and threats to our social epistemic systems that have co-emerged with advances in information mediating and producing technologies. *A slow and steady decline in epistemic security is a serious cause for concern when combined with a rising epistemic security threshold.*

4. Rising Epistemic Security Threshold

Another factor to consider is that epistemic security today and in the future may simply matter more than it did in the past. Consequently, protecting and improving epistemic security would become more important irrespective of how much epistemic security has improved over the years.

It may help to think about there being a ***threshold level of epistemic security*** below which the risk of epistemic failure (knowledge operation failures) put humanity at an unacceptably high risk of catastrophe. The "epistemic security threshold" for a society can be understood as the level of epistemic security needed for the society to have the capacity to effectively prepare for and respond to risks and crises. The more epistemically secure a society needs to be to ensure successful risk prevention/response (for instance, because successful response requires more accurate information about the risk and/or more highly coordinated action to effectively mitigate the risk), the higher that threshold will be.

There are several reasons to believe that epistemic security thresholds are on the rise.

4.1 More numerous and more severe challenges

The first reason is straightforwardly that the consequences of epistemic insecurity are getting more significant as society faces more and more severe challenges and crises (Cotton-Barratt et al., 2016). The risks we have to be most concerned about today and going into the future are increasingly those of our own making such as anthropogenic climate change, bioengineering (bioweapons/pandemics), geoengineering, power seeking AI, nuclear weapons, and so forth.

4.2 Facing current and future GCRs demands greater global coordination

Second, a higher capacity for cooperation is needed to address current and emerging catastrophic risks due to the increasingly complex and global nature of those risks.

Global risks will require global solutions approaches which require greater coordination and cooperation among larger and more diverse groups of stakeholders. For example anthropogenic climate change can not be mitigated through the actions of a single state (and even agreement within a single state is difficult to orchestrate). The same can be said, for example, about mitigating risks posed by nuclear proliferation, biological gain of function research, and advanced AI. Effective risk mitigation will require global cooperation which, in turn, depends on the maintenance of robust lines of communication and epistemic trust between global actors.

4.3 Global convergence on democratic governance structures

Third, there is a global trend toward convergence on democratic governance structures (Desilver, 2019).[3] *If we value democracy (which I will assume my readers do) then greater*

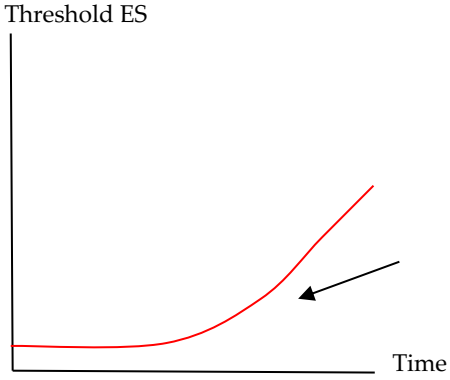
[3] I thank Seán Ó hÉigeartaigh for bringing this third point to my attention.

epistemic security is necessary to enable democratic societies to motivate the populace to make robust and unified responses to crises.

What the public believes (about world events, about politicians, about scientific findings and authorities, etc.) has an increasing impact on what decisions are made in response to challenges and how successfully necessary actions are carried out. As described by The Consilience Project, "In a democracy, we cannot rely on a single monarch or cloistered politburo to make good decisions for us. Democracy is self-government at scale and, therefore, requires sensemaking at scale in the form of an epistemically healthy public sphere" ('Democracy and the Epistemic Commons', 2021).

For example, a public that is conflicted about vaccine safety is more problematic from the perspective of preventing pandemics when that public has a choice about whether they take vaccines and when they can vote for representatives who might pass policy to prevent campaigns promoting vaccine safety. Similarly, a public that is conflicted about the reality and/or causes of climate change is more problematic when they can vote for representatives who are dedicated climate denialists and who will enact legislation to prevent discussions of "anthropogenic" climate change. With the proliferation of democratic governance structures, the sphere of those decision-makers who must be "epistemically secured" to facilitate well-informed and timely decision-making is getting ever larger and requiring greater coordination.

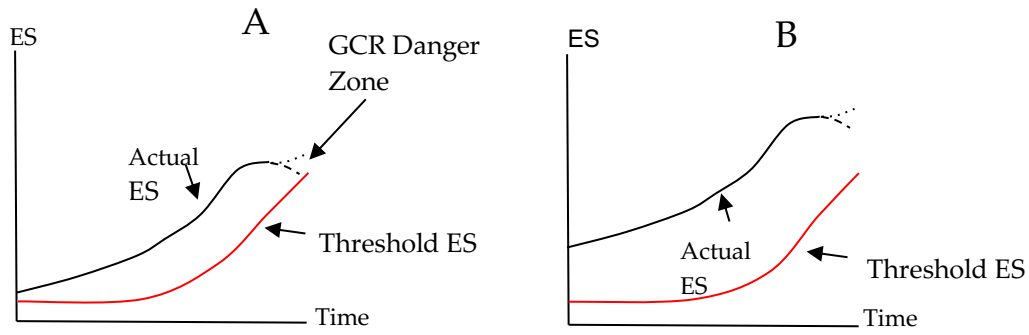
Overall, we might think of the rise in epistemic security threshold levels over time looking something like this (again, just a rough sketch).



Nuclear bombs, bioengineering, anthropogenic climate change, geoengineering, AI, etc. combined with increased globalization and convergence on democratic forms of governance.

5. Overlapping Trajectories = GCR Danger Zone

Now let's overlap the Actual ES graph (Section 3) with the Threshold ES graph (Section 4). It is hard to know how to overlap the graphs. I am, after all, not working with any real quantities, just back-of-the-envelope sketches. Is the current threshold ES above (graph A) or below the actual state of ES (graph B)?



Looking at responses to current and ongoing crises (e.g. Covid-19) I propose we are closer to graph A. That said, I don't think it matters much which is more accurate. Even if the current state and trajectory of epistemic security are more like B, we still have reason for concern.

*The two trends are on a collision course, if they have not collided already, and where the current state of epistemic security overlaps with the threshold epistemic security level we enter a "GCR Danger Zone"***.** In this zone, the risks of epistemic failure put society at an unacceptably high risk of catastrophe. Put another way, the state of epistemic security, whatever it is, is insufficient for orchestrating adequate GCR mitigation and response efforts. So either way, A or B, intervention is needed to push out the intersection - to push the GCR danger zone - as far into the future as possible.

6. Possible Levers for Intervention

I will keep this section brief. The aim is to help structure thinking about how we might back out of the (impending) GCR danger zone. Based on my analysis, there are two main levers we can pull: work to (1) lower the epistemic security threshold, and/ or (2) to improve the current and projected epistemic security levels.

6.1 Lowering the epistemic security threshold

To lower the epistemic security threshold is to cater an environment for ourselves in which a lower degree of epistemic security is needed to adequately respond to the kinds of threats and crises we face today and are likely to face in the future.

Based on the discussion in section 4, it would seem that one option, which I will promptly dismiss, is to eschew democracy (Section 4.3). Theoretically, authoritarian and totalitarian governments have the greatest potential for epistemic security and for coordinating efficient responses to crises and challenges; it is easier to keep small groups of decision makers well-informed and well-coordinated than a broader voting public, and where public support or participation in crisis response is needed, information (including reliable decision-guiding information) can be tightly controlled and uniformly distributed.

There are many reasons, however, to eschew totalitarianism, not least of which refer to significant human rights concerns as demonstrated too many times throughout history. But even purely from an epistemic perspective, totalitarianism is a bad idea. While the potential for epistemic security is great where information streams can be closely

monitored and controlled, it should be noted that improved epistemic security is an incredibly unlikely manifestation of stable totalitarianism. The much more likely and virtually guaranteed alternative is an epistemic dystopia in which all information is carefully catered to manipulate public opinions and behavior. This is a GCR in and of itself. So again, let's dismiss the option of limiting the spread of democracy as a mechanism for lowering the epistemic security threshold.

If we want to pull on the epistemic security threshold lever, the other option (based on the discussion in sections 4.1 and 4.2) is to reduce the severity, prevalence, and complexity of GCRs; epistemic insecurity becomes less concerning as a GCR amplifier if the GCRs themselves are less catastrophic or easier to deal with. To put it another way, effective efforts to reduce a variety of catastrophic risks also reduces the importance of epistemic security for protecting ourselves against those risk (except, of course, where we are concerned with epistemic insecurity itself as the risk, in which case we've just set ourselves a tautology—epistemic security becomes less important if epistemic security becomes less important).

So, by all means, do work to reduce GCRs born from climate change, nuclear armament, biological gain of function research, AGI, geoengineering and so forth. Such efforts may indeed lower the epistemic security threshold, but remember (a) that epistemic security will be key to succeeding in these risk reduction endeavors in the first place—we need to be able to produce, identify, and consumer reliable decision-guiding information effectively respond to complex challenges and crises—and (b) that even if we are able to reduce other GCRs now, epistemic security will remain necessary for maintaining those reduced risk levels via continued cooperation around well-informed decision making and action.

6.2 Improving epistemic security

Improving the current and projected epistemic security levels remains an important lever for holding the GCR Danger Zone at arms length.

When considering possible intervention points, it may help to structure one's thinking along the three characteristics of epistemic security I outlined in Section 3: information accessibility, information environment safety, and information recipient sensitivity. For example, given how prominently the topics feature in my own rough analysis of each characteristic, I posit that some of the most impactful work for improving epistemic security would center on scanning for and mitigating epistemic threats posed by information producing and mediating technologies.

But as discussed in Section 2, these appraisals must take into account the broader social epistemic context in which the technologies are deployed. This will require working closely with experts in a variety of fields. For example, understanding how a content recommendation algorithm will impact how people consume information and form beliefs is not just a matter of understanding how the algorithm works (a technological question), but of human psychology and sociology as well.

I expect that a key step toward improving epistemic security would therefore be to organize interdisciplinary communities to collaborate on identifying and responding to epistemic vulnerabilities and threats. Such community building proposals are discussed at greater length in Section 4 of the 2020 epistemic security report (Seger et al., 2020).

7. Conclusion

In summary:

- Epistemic security has improved greatly over time due to improvements in information accessibility, information environment safety, and information recipient sensitivity. However, that upward trend has, in recent decades, leveled off and declined slowly in large part due to the negative impact on modern information producing and mediating technologies on the sensitivity of information recipients to information and information source quality.
- The **epistemic security threshold**—the level of epistemic security needed for the society to have the capacity to effectively prepare for/ respond to risks and crises—is also rising
- This put our actual state of epistemic security on a collision course with threshold epistemic security. Where the lines overlap we enter a **"GCR Danger Zone"** in which a society's epistemic capacities are ill-suited to responding to the challenges and crises it faces.
- Epistemic security should therefore be considered a priority GCR cause area because there is likely a **high direct value** in interventions that improve epistemic security; they would push out the timeline on the "GCR danger zone" by increasing our capacity to deal with the kinds of complex challenges and crises we seem likely to see more of as we head into the future.
- Where there is uncertainty about the tractability of various intervention points there is **high informational value** in orchestrating further epistemic security studies.

References

Arguedas, A. R., Robertson, C., T., Fletcher, R., & Nielsen, R. K. (2022). *Echo Chambers, Filter Bubbles, and Polarisation: A Literature Review*. https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review

Askell, A. (2017, June 4). *The Moral Value of Information*. Effective Altruism. https://www.effectivealtruism.org/articles/the-moral-value-of-information-amanda-askell

Bostrom, N., & ćirković, M. M. (2008). Introduction. In N. Bostrom & M. M. ćirković, *Global Catastrophic Risks*. Oxford University Press. https://doi.org/10.1093/oso/9780198570509.003.0004

Cause Selection. (2021, July 22). *Open Philanthropy*. https://www.openphilanthropy.org/cause-selection/

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs*, *98*(1), 147–155.

Cotton-Barratt, O., Farquhar, S., Halstead, J., Schubert, S., & Snyder-Beattie, A. (2016). *Global Catastrophic Risks 2016*. Global Challenges Foundation. http://globalprioritiesproject.org/wp-content/uploads/2016/04/Global-Catastrophic-Risk-Annual-Report-2016-FINAL.pdf

Democracy and the Epistemic Commons. (2021, February 27). *The Consilience Project*. https://consilienceproject.org/democracy-and-the-epistemic-commons/

Desilver, D. (2019). *Despite global concerns about democracy, more than half of countries are democratic*. Pew Research Center. https://www.pewresearch.org/short-reads/2019/05/14/more-than-half-of-countries-are-democratic/

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations* (arXiv:2301.04246). arXiv. http://arxiv.org/abs/2301.04246

Goldsworthy, A., Osborne, L., & Chesterfield, A. (2022). *Poles Apart: Why People Turn Against Each Other, and How to Bring Them Together*. Penguin Books.

Horton, C. (2018). *The simple but ingenious system Taiwan uses to crowdsource its laws*. MIT Technology Review. https://www.technologyreview.com/2018/08/21/240284/the-simple-but-ingenious-system-taiwan-uses-to-crowdsource-its-laws/

Horvitz, E. (2022). On the Horizon: Interactive and Compositional Deepfakes. *International Conference on Multimodal Interaction*, 653–661. https://doi.org/10.1145/3536221.3558175

Introducing the Collective Intelligence Project: Solving the Transformative Technology Trilemma through Governance R&D. (2023). The Collective Intelligence Project. https://cip.org/whitepaper

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lin, H. (2019). The existential threat from cyber-enabled information warfare. *Bulletin of the Atomic Scientists*, *75*(4), 187–196. https://doi.org/10.1080/00963402.2019.1629574

McIntyre, L. C. (2018). *Post-truth*. MIT Press.

Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F., Chadwick, M., Glaese, M., Young, S., Campbell-Gillingham, L., Irving, G., & McAleese, N. (2022). *Teaching language models to support answers with verified quotes* (arXiv:2203.11147). arXiv. http://arxiv.org/abs/2203.11147

Nichols, T. M. (2017). *The death of expertise: The campaign against established knowledge and why it matters*. Oxford University Press.

O'Connor, C., & Weatherall, J. O. (2019). *The misinformation age: How false beliefs spread*. Yale University Press.

Ovadya, A. (2022, October 27). 'Contextualization Engines' can fight misinformation without censorship. *Medium*. https://aviv.medium.com/contextualization-engines-can-fight-misinformation-without-censorship-c5c47222a3b7

Ovadya, A., & Thorburn, L. (2023). *Bridging Systems: Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance* (arXiv:2301.09976). arXiv. http://arxiv.org/abs/2301.09976

Sedova, K., McNeill, C., Johnson, A., Joshi, A., & Wulkan, I. (2021). *AI and the Future of Disinformation Campaigns: A Threat Model*. Center for Security and Emerging Technology. https://cset.georgetown.edu/wp-content/uploads/CSET-AI-and-the-Future-of-Disinformation-Campaigns-Part-2.pdf

Seger, E. (2022). Exploring epistemic security: The catastrophic risk of epistemic insecurity in a technologically advanced world. *International Security Journal*, *4*(46), 88–90.

Seger, E., Avin, S., Pearson, G., Briers, M., Ó Heigeartaigh, S., & Bacon, H. (2020). *Tackling threats to informed decisionmaking in democratic societies: Promoting epistemic security in a technologically-advanced world*. The Alan Turing Institute.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "Fake News": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Thomas, E., Thompson, N., & Wanless, A. (2020). *The Challenges of Countering Influence Operations* (No. 2; Policy Perspectives). Carnegie Endowment of International Peace. https://carnegieendowment.org/files/Thomas_Thompson_Wanless_-_PCIO_-_Influence_Ops.pdf

van der Linden, S., Maibach, E., Cook, J., Leiserowitz, A., & Lewandowsky, S. (2017). Inoculating against misinformation. *Science*, *358*(6367), 1141–1142. https://doi.org/10.1126/science.aar4533

Stanford
Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Maniacs, Misanthropes, and Omnicidal Terrorists: Reassessing the Agential Risk Framework

Émile P. Torres [1*]

Citation: Torres, Émile P. Maniacs, Misanthropes, and Omnicidal Terrorists: Reassessing the Agential Risk Framework. *Proceedings of the Stanford Existential Risks Conference* 2023, 36-47. https://doi.org/10.25740/gv769bs1452

Academic Editor: Dan Zimmer, Trond Undheim

Funding: None

Conflict of Interest Statement: None

Informed Consent Statement: N/A

Acknowledgments: Thanks to Dan Zimmer for helpful comments on a draft. Thanks also to the participants of the 2023 Stanford Existential Risks Conference for feedback on a presentation based on this paper.

Author Contributions: N/A

Abstract: Agential risks are the those that arise when agents *willing* to destroy the world collide with the technological *ability* to do this. Previous work on agential risks focused on the *identities* and *motives* of risky agents. This chapter explores the various *aims* that such agents might have, which constitute a distinct axis that is largely orthogonal to that of motives. In particular, I propose a novel typology of agential risks based on four possible goals: (i) killing a large portion of the human population; (ii) destroying civilization; (iii) triggering global violence and/or an apocalyptic war; and (iv) causing human extinction. Many risky agents identify *only one* of these as their ultimate goal, although they are of course not mutually exclusive. After providing a number of historical examples within each category, I turn to a different question, namely, "If a global catastrophe were to have happened, who would be most responsible?" This introduces Luke Kemp's concept of "Agents of Doom," which offers an alternative perspective to the agential risk approach that I first proposed in 2018. However, I suggest that these frameworks should be integrated, as successfully navigating the obstacle course of twenty-first-century death traps will require addressing both phenomena.

Keywords: agential risks, omnicidal agents, global catastrophic risks, existential risks

[1] Researcher, Leibniz Universität Hannover, Institut für Philosophie, Leibniz Universität Hannover, Lange Laube 32, 30159, Hannover, Germany.

* Correspondence: philosophytorres@gmail.com

1. Introduction

This paper offers a novel perspective on what I have previously called "agential risks," or the risks posed by agents who *would* or *might* push a "doomsday button" if one were within finger's reach. The central questions of Agential Risk Studies are: "*Who* exactly would destroy the world if only they could, and *why* would they do this?" The urgency of studying the "who" and "why" arises from a fact about the "could": it appears that the means necessary for nonstate actors to unilaterally inflict global-scale harms are becoming increasingly available, due to the "democratization" of science and technology. Put differently, history is overflowing with individuals who wanted civilization to be destroyed, hoped for worldwide catastrophes and apocalyptic wars, or harbored a death wish for humanity. But such individuals were largely powerless to bring these about; though "willing," they were not "able," and the unilateral destruction of the world requires both. However, the exponential development of dual-use emerging technologies is rapidly changing this situation. Whereas there were *zero* agents capable of precipitating, on their own, global-scale havoc just one century ago, there could be *hundreds of millions* or more with this capacity in the coming decades. Consequently, identifying "risky agents" (i.e., agents who give rise to agential risk) and understanding what motivates them has become an extremely important and time-sensitive issue. Only once identities and motives have been established can we begin to devise effective strategies to neutralize the danger.

Yet identities and motives, which were the focus of previous studies (namely, my 2018b and 2018a papers, respectively), are not the only relevant properties of risky agents. We may also want to know about their particular *aims*, as this may be relevant to questions of *prioritization*, depending on one's ethical commitments. A longtermist, for example, will want to prioritize targeting agents whose aim is human extinction over, say, those who "merely" want to kill millions or a few billions. Hence, in what follows, I offer a new typology of agential risks focused on *aims* rather than *motives*. I will then pivot to a distinct question that my previous work did not adequately address: "If an anthropogenic global catastrophe were to occur, who would be *most likely* to have caused it?" This will bring the topic of agential risk into conversation with what Luke Kemp calls "Agents of Doom," or the "small group of often overlapping, powerful industries dominated by a few actors [such as] military-industrial complexes, the fossil fuel industry, and Big Tech" (Kemp, 2021). I now believe that Kemp is correct in arguing that Agents of Doom are the primary drivers of global catastrophic risk, and hence that mitigating global catastrophic risk requires a greater focus on these entities than the groups and individuals that my prior work examined—at least in the near term. However, I will also claim that Kemp underestimates the longer-term dangers posed by risky agents, given that minuscule probabilities can quickly accumulate across space and time, resulting in near-certain doom on timescales meaningful to contemporary civilization.

2. Agential Error and Terror

The most fundamental distinction within the category of agential risk is between what Martin Rees famously called, in his 2003 book *Our Final Hour*, "terror and error" (Rees, 2003). Let's consider these in reverse order.

2.1 Agential Error

Agential error can be understood as denoting any scenario in which an agent with access to a doomsday button presses it without intending to do so. An example would be a biohacker who inadvertently enhances the pathogenicity of a microorganism and then accidentally releases it into the world. Perhaps this individual was modifying pathogens

in hopes of discovering a new treatment for disease. There is, in fact, a protracted history of laboratory leaks involving dangerous germs, such as an outbreak of smallpox in 1971 that killed three people and a similar incident involving anthrax that killed 66 people in 1979 (Ord, 2020). Or consider the 1977-1978 "Russian flu" epidemic, which took some 700,000 lives worldwide and "probably" had an anthropogenic origin, given that "the genetic sequence of the virus was nearly identical to the sequences of decades-old strains" (Rozo and Gronvall, 2015).

The point is that as the number of professional scientists handling lethal pathogens increases, as more gain-of-function research is conducted around the world, and as the community of biohacker hobbyists grows, the probability of an accidental pandemic is likely to rise. This is not a trivial observation, as simple calculations show that, because probabilities accumulate across space and time, even a very small likelihood of any single individual or laboratory releasing a deadly pathogen quickly approaches 1 over timescales of decades or centuries. To illustrate, imagine a world populated by 10 billion people, only 500 of whom are in a position to synthesize and accidentally release a contagious, lethal pathogen such that a large number of people become infected. Five hundred is just 0.000005% of the total population. If these individuals were to have a mere 0.01 chance of initiating an engineered pandemic by mistake per decade, the likelihood of a pandemic would be 99% over ten years (Torres, 2017). Yet the number of scientists—to say nothing of biohackers—who have conducted genetic engineering research in recent years is much higher than 500. According to John Sotos (2019), a PubMed search conducted in 2017 for the keywords "genetic techniques" yielded 594,458 publications with "1,555,661 unique author names." Again, this number is likely to grow even larger in the future, which suggests, for straightforward statistical reasons, that the risks posed by agential error could become very significant.

2.1 Agential Terror

Turning now to agential terror, I think it is useful to divide this category into four groups, each defined by the particular type of global catastrophe the relevant agents want to bring about. These are: (i) to kill a large portion of the human population (millions or billions of people), (ii) to destroy civilization, (iii) to trigger global violence and/or an apocalyptic war, and (iv) to cause human extinction (specifically, in the sense of "final" extinction, discussed in chapter 7 of Torres, 2023). This typology cuts across the earlier typology presented in my 2018a and 2018b papers, which—as noted above—focused on motivation rather than aims, reasons rather than goals. In those studies, I distinguished between agents driven by apocalyptic ideologies, ethical considerations, radical environmentalism, and idiosyncratic beliefs or grievances. For the remainder of this section, let's combine these typologies by examining a few agents that fall within each of the four categories enumerated above.

(i) The first category consists primarily of environmental extremists and idiosyncratic actors. If given a "doomsday button," they would push it specifically to kill millions or billions of people. This might entail civilizational collapse, world wars, or human extinction, although in some cases, agents in this first category might also wish to avoid such outcomes. For example, Pentti Linkola is a self-described "eco-fascist" who has declared that "if there were a button I could press, I would sacrifice myself without hesitation, if it meant millions of people would die" (Milbank, 1994). Linkola's aim is to avoid an "ecocatastrophe" by significantly reducing the global population, though he does not want humanity to die out, as human extinction would be "an extremely bad thing" (PFN, 2007). Elsewhere, he argues that a world war would be "a happy occasion for the planet," but only if it "target[s] the actual breeding potential" of humanity, that is,

"young females as well as children, of which a half is girls. If this doesn't happen, waging war is mostly [a] waste of time or even harmful" (Linkola, 2006; Milbank, 1994).

Linkola thus provides an example of someone whose explicit aim is the death of a large number of people for the sake of preserving the environment. Many people have expressed the same desire on websites like Debate.org and Reddit. A 2019 Reddit post titled "genocide is the only way to save the human race," for example, asserts that "a vast program of mass murder is the only way to save the human race. Decimate the population to maybe only 10% or less than what it is now. Less than a billion" (JNZ, 2019).

A similar view comes from someone named "Pete," who appears to be affiliated with the Church of Euthanasia, an ecocentric, neo-Malthusian group from the 1990s whose slogan was "Save the Planet, Kill Yourself." In an article written by Chris Korda, founder of the Church of Euthanasia, Pete is quoted as saying:

it becomes more and more clear to me every day that mass sterilization is the only answer to our environmental problems. Perhaps that makes me more radical than the Church, which advocates voluntary measures only. But I'm ready to hop in a B-52 with a payload of genetically-tailored-virus smart-bombs, enough to sterilize 99% of the world's population in one trans-globe flight. Someone need only invent the hardware, train me, and present me with the opportunity. Maybe in 10 years it will be possible (Korda, 1994).

Other examples involve agents motivated by more idiosyncratic beliefs or grievances. Many of these cases are noteworthy because they involve individuals who not only expressed a desire to cause mass death, but actually perpetrated violent rampage shootings, most of which ended in their own deaths. For example, the 2007 Virginia Tech shooting was perpetrated by an individual who fantasized about slaughtering a portion of humanity that he described as "the sadistic, the corrupt, and the wicked who prey [on] the Weak, the Defenseless, and the Innocent." He adds that (redacting vulgarities): "F*ck you … I [say] we take up the cross, Children of Ishmael, take up our guns and knives and any sharp objects, and take no prisoners and spare no lives until our last breath and last ounce of energy" (quoted in Torres, 2018b). Another example comes from the individual who perpetrated the 2014 Isla Vista killings. The primary target of his hatred was women, who he believed had wronged him by rejecting his "sexual advances." He wrote: "If it's war they want, then war they shall have. It will be a war that will result in their complete and utter annihilation" (Langman, 2014). Additional instances could be adduced, but the point is that at least some idiosyncratic actors seemed to desire the destruction of a large portion, though not all, of humanity.

Within this first category, then, one finds agents motivated by radical environmentalist considerations and idiosyncratic grievances. If such individuals were, in the future, to gain access to a button that would enable them to kill millions or billions of people, we should be concerned that they would enthusiastically push it.

(ii) The second category consists of agents who specifically wish for the complete destruction of civilization. This would include the most radical neo-Luddites, anarcho-primitivists, and ecoterrorists who are open to the possibility of destroying civilization through violence. An obvious example is Ted Kaczynski, also known as the Unabomber, who perpetrated a campaign of domestic terrorism from 1978 until 1995, when he managed to persuade the *Washington Post* to publish a 35,000-word manifesto titled "Industrial Society and Its Future." This defended a neo-Luddite thesis according to which the megatechnics of industrial civilization have profoundly compromised human freedom, and hence to regain this freedom the system must be dismantled. The aim would

be a return to what Kaczynski called "small-scale technologies," that is, "technology that can be used by small-scale communities without outside assistance" (Kaczynski, 1995).

Note how this contrasts with Linkola's eco-fascist view: whereas Linkola hopes for a mass slaughter event to reduce the human population to avoid an eco-catastrophe, Kaczynski is less concerned about the preservation of the environment or integrity of ecosystems than he is about the oppressive consequences of industrialization. His aim is merely to eliminate large-scale civilization, which need not require millions to die in the process. As he writes, "we therefore advocate a revolution against the industrial system. This revolution may or may not make use of violence; it may be sudden or it may be a relatively gradual process spanning a few decades. … Its object will be to overthrow not governments but the economic and technological basis of the present society" (Kaczynski, 1995). However, it is worth pointing out that even a non-violent, gradual means of dismantling industrial civilization would likely *result* in a large number of casualties, given that the "small-scale" technological systems that Kaczynski envisions would probably be unable to support the current world population. Any transition to a post-civilization mode of existence would, therefore, very likely coincide with a global catastrophe.

More recently, some groups inspired by Kaczynski's neo-Luddism have emerged, such as "Individualidades Tendiendo a lo Salvaje," or ITS, which mailed a bomb to researchers at the Monterrey Institute of Technology and Higher Education in Mexico City and have "been linked to attacks in France, Spain, and Chile" (Lloyd and Young, 2011). Others motivated by the particular goal of destroying civilization include anarcho-primitivists and some radical environmentalists. The former hopes for a return to the lifeways of hunter-gatherers, as captured by the slogan "back to the Pleistocene," while the latter see civilization as incompatible with a functioning, healthy biosphere. If individuals who embrace such ideologies were to gain access to civilization-destroying technologies, we should have little doubt they would exploit them to cause as much damage to civilization as possible. As Bill Joy wrote in 2000, "we're lucky Kaczynski was a mathematician, not a molecular biologist," since if he had been the latter, he could have inflicted much greater damage than he did (Joy, 2000). Anarcho-primitivists and ecoterrorists are, in fact, well-aware of the potential for advanced technologies to radically empower individuals who see civilizational destruction as desirable, a point that we will return to below.

(iii) The third category encompasses agents whose aim is to foment global-scale violence, which could take the form of (what they see as) an apocalyptic war. The Japanese doomsday cult Aum Shinrikyo provides an example: they perpetrated the 1995 Tokyo subway sarin attack in hopes of starting "a nuclear World War III between Japan and the US to trigger Armageddon." In other words, Aum embraced an "active eschatology" according to which it was their divine mission to initiate the apocalypse, on the other side of which lies paradise, or what Aum called "Shambhala," borrowing a term from the Buddhist scriptures (Flannery, 2015, p. 222). Although the sarin attack resulted in the group's leader, Asahara Shoko, being arrested, Aum may have intended it as a test-run for a much deadlier attack that Aum was planning for several months later, which would have involved spraying "a total of 70 tons of sarin gas from a helicopter, purchased in Russia, with the aim of killing millions" (Flannery, 2015). Furthermore, while Aum expected millions or billions of people to die in the resulting nuclear war, the death of these people was not the primary aim of their terroristic actions; rather, it was the fulfillment of apocalyptic prophecy that would ultimately usher in Shambhala.

Many other examples could be mentioned, and indeed history is replete with examples of apocalyptic movements engaging in violent acts for the sake of catalyzing the world's end. Another recent case comes from the Islamic State. The "active" eschatological beliefs of

this group evolved over time, though it was initially focused on triggering the Islamic version of Armageddon, between Muslim forces and the "Romans," interpreted as "the West"—which some prophetic hadith suggest will unfold around the small town of Dabiq, in northern Syria. As Abu Musab al-Zarqawi, who helped found the Islamic State, declared: "The spark has been lit here in Iraq, and its heat will continue to intensify … until it burns the Crusader armies in Dabiq" (Torres, 2018b). In fact, this was quoted at the beginning of every issue of the Islamic State's propaganda magazine named *Dabiq*. In service of this end, Islamic State members often fantasized about gaining access to weapons of mass destruction (WMDs), including radiological and nuclear weapons, the latter of which it speculated could be smuggled into the United States and detonated. Others within the organization advocated for the use of biological weapons, given their "low cost and high rate of casualties." A laptop owned by a chemistry student, for example, "contained a 19-page manual to learn how to turn the bubonic plague into a weapon of war." It notes that "there are many methods to spread the biological or chemical agents in a way to impact the biggest number of people. Air, main water supplies, food. The most dangerous is through the air" (Mcelroy, 2014). Again, although millions or billions of deaths might result from the grand battle that Islamic State members hoped to bring about, the primary aim was to catalyze an eschatologically significant conflict that, as such, could usher in the world after this one: paradise.

(iv) For many of the individuals or groups mentioned above, the explicit aim was not to kill everyone on the planet, including themselves. To the contrary, agents like Linkola, Kaczynski, "Pete," and the idiosyncratic actors that we discussed would have seen this outcome as generally undesirable.[1] In the case of religious groups like Aum and the Islamic State, human extinction (in a naturalistic sense) would have been seen as impossible, given that, according to the eschatological narratives that they embraced, humanity will ultimately survive the cataclysmic happenings at the end of the world.

In contrast, the final category in our new typology consists of agents who explicitly wanted the human species to cease existing entirely. This contains a broader diversity of agents than the other three categories, although the very same radical environmentalist and idiosyncratic grievances that motivated some of the actors above could also lead one to conclude that our extinction would be desirable. For example, if one adopts an ecocentric theory of value, whereby nonhuman animals and even nonliving entities (such as the land and rivers) possess intrinsic value, and if one accepts that *Homo sapiens* is destroying the natural world, then one might conclude that the human population must decline considerably—to a sustainable level. However, one might also conclude that it would be best if we ceased to exist altogether, perhaps because one believes that keeping the human population low is not realistic. As the Gaia Liberation Front (GLF) declares, "the evidence is overwhelming that the Humans are programmed to kill the Earth. This programming is not only cultural, but probably also genetic since the major technologies Humans use for this purpose, from agriculture and metallurgy to writing and mathematics, have all been invented independently more than once." Hence, "every Human now carries the seeds of terracide. If any Humans survive, they may start the whole thing over again." GLF thus writes that their "mission is the total liberation of the Earth, which can be accomplished only through the extinction of the Humans as a species." They further note that bioengineering provides "the *specific* technology for doing the job right—and it's something that could be done by just one person with the necessary expertise and access to the necessary equipment." They advocate for several viruses to be synthesized and then released one after the other, where the aim of the second release

[1] Indeed, the individual responsible for the Isla Vista killings entertained "a bizarre fantasy of slaughtering most women but keeping a small number imprisoned 'for the sake of reproduction'" (Torres 2018b).

would be to target "the generals and the politicians" after emerging from "their shelters" (GLF, 1994).

Many other environmental radicals have echoed this, as when an anonymous author wrote in the *Earth First! Journal* that "contributions are urgently solicited for scientific research on a species specific virus that will eliminate Homo shiticus from the planet" (quoted in Torres, 2018b). Or consider that the aforementioned group ITS, which initially aligned itself with Kaczynski's neo-Luddite philosophy, now seems to hold that "the human being deserves extinction" (Campbell, 2017). Similar attitudes are found on many websites, as when a respondent to a Debate.org question—"If you could push a button and destroy all human life[, w]ould you? All other life would survive ass is"—wrote "my view is that Mankind is a plague. … I vote to destroy mankind and let nature start over." Another responded that "in the short time we've been on this planet, humans have already destroyed so much. We destroy ecosystems, and kill off entire species of animals. … The world would be better off without humans as a whole" (quoted in Torres, 2018b).

There are also many examples of agents motivated by more idiosyncratic beliefs or grievances wishing for total human extinction. For instance, one of the adolescents who perpetrated the 1999 Columbine High School massacre wrote in his diary, "I think I would want us to go extinct," to which he added "I just wish I could actually DO this instead of just DREAM about it all." Elsewhere he declared: "If you recall your history the Nazis came up with a 'final solution' to the Jewish problem. Kill them all. Well, in case you haven't figured it out yet, I say 'KILL MANKIND' no one should survive" (Langman, 2009a). Another rampage shooter expressed his desire to (again, redacting vulgarities) "turn this f*cking world into a graveyard," and once described his mood online as "destroy all mankind" (Gill, 2005).

However, an imperative to commit omnicide—the killing of everyone—also seems to follow from certain ethical theories, such as negative utilitarianism (NU). This theory can take many forms, but on its most radical interpretation, it states that the only thing that matters, morally, is the reduction of suffering. Since the elimination of all beings capable of suffering would eliminate all suffering, NU seems to enjoin those who accept it to become a "world-exploder," in the words of R. N. Smart (1958). While NU is not a popular ethical theory, it does have its adherents, and in fact a former colleague of mine "knew two people who entered PhD programs in biology for the express purpose of learning how to synthesize designer pathogens to wipe out humanity"—that is, *because* of their NU beliefs (Torres, 2019). I was informed that they both "chilled out" over time, and one now even has a child.

One might reach the same omnicidal conclusion based on a view called "philosophical pessimism," which emerged in the latter nineteenth century among a handful of influential German philosophers. The most notable example comes from Eduard von Hartmann, who believed that life is very bad, and hence it would be best not just for humanity to cease existing, but for all life in the universe. He thus advocated the complete annihilation of the very possibility of life arising anywhere. How could we do this? He admitted that, given the state of human knowledge at the time he was writing, we cannot imagine the means for achieving this end. However, as humanity continues to progress, he believed the solution would come into view (Torres, 2023). In fact, there is a scientifically plausible kill mechanism for eliminating virtually everything within our future light cone: if the universe is in a "false vacuum" state, which it might be, high-powered particle accelerators may be able to tip it into a "true vacuum" state. This would involve nucleating a "vacuum bubble" that would expand in all directions at nearly the speed of light, destroying everything it comes into contact with. It is not inconceivable

that a wealthy individual with an inclination toward Hartmann's pessimistic worldview might someday try to build a particle accelerator for precisely this purpose.

3. Discussion

The literature on "global catastrophic" and "existential" risks has, since the early 2000s, tended to foreground the dangers posed by nonstate actors, as a result of terror or error—although virtually no one has provided anything close to a detailed study of the *specific properties* of such actors (e.g., their identities, motives, or aims). I do not believe this focus is unjustified, given the fact that dual-use emerging technologies are becoming simultaneously more powerful and accessible. Consider that one of the major barriers to the creation of bioweapons has been so-called "tacit knowledge," that is, the know-how (rather than know-that) required to, say, enhance the pathogenicity of viruses in the laboratory. Yet one of the explicit aims of synthetic biology is to *minimize* the relevance of tacit knowledge. The result is that individuals or groups with little or no laboratory training could potentially synthesize highly lethal and contagious viruses or bacteria—at the push of a button (Mukunda et al., 2009). Similarly, some have worried that anticipated future technologies like nanofactories could empower individuals or groups with few resources to manufacture extremely dangerous weapons. All this would require is a blueprint (downloaded from the Internet), an electrical outlet, and a feedstock molecule like propane or acetylene (Ramsden, 2009). Artificial intelligence could have similar consequences, making it possible for lone wolves or terrorist groups to initiate cyber-attacks that wreak havoc on the global economy, or perhaps to hack into nuclear launch systems and trigger a nuclear strike. Finally, consider the case of SILEX, i.e., the separation of isotopes by laser excitation. This could remove one of the major barriers to nonstate actors accessing weapons-grade uranium, and indeed it has led some to worry that SILEX "may create new proliferation risks" (Snyder, 2016). This is why the central questions of Agential Risk Studies have suddenly become urgent and profound.

But there is a separate, even more important question that we should ask: "If a global catastrophe were to occur, who is most likely to have caused it?" There are two ways to answer this question: the first concerns the issue of which entities are responsible for creating the *general milieu* in which "willing" agents are now "able" to inflict global-scale harms? Who exactly has made agential risk *possible*? The second pertains to a separate category of actors, not mentioned above, who (a) are not *explicitly motivated* by any of the four aims specified above, yet (b) through their own *deliberate actions*, might precipitate outcomes that are just as bad: millions dying, civilizational destruction, global war, or human extinction. Let's consider (a) and (b) in turn:

As Kemp observes, a relatively small number of military-industrial entities, driven by concerns over "national security" and the profit motive, are developing precisely those technologies that will proliferate doomsday buttons this century. These entities are the underlying drivers, enablers, or generators of global catastrophic risk. If an omnicidal ecoterrorist, apocalyptic extremist, or radical NU were to destroy the world, these agents would constitute the proximate agential cause of the disaster. But there would be a deeper cause behind this outcome, namely, the entities that are making this feasible by funding and/or conducting R&D aimed at building dually usable technologies that we—and they—*know* are extremely risky. Kemp calls these "Agents of Doom," as they are the *reason* our world is becoming ever-more risky.

Many of these same Agents of Doom, though, could very well become the proximate cause of global catastrophes, not because they *intend* this outcome but because such catastrophes are an inadvertent side-effect of deliberate actions aimed at some other end. This class of possibilities does not fit neatly into the terror/error dichotomy. For example, consider the

case of OpenAI, whose explicit aim is to create artificial general intelligence (AGI) that "benefits all of humanity" (OpenAI, 2018). Toward this end, they have recently released products built on "large language models" (LLMs) like ChatGPT, along with an LLM called GPT-4, which currently powers Microsoft's Bing chatbot. Despite 87% of the public believing that advanced AI will either cause as much harm as good or *more* harm than good, OpenAI's approach is to develop and release new AI systems as quickly as possible (MU, 2023). Yet these technologies are already causing significant harms to people in the Global North and, especially, Global South, and OpenAI's leaders have explicitly acknowledged that more powerful AIs could ultimately cause human extinction. As the CEO of OpenAI, Sam Altman, states: "And the bad case—and I think this is important to say—is like lights out for all of us" (see Bender et al., 2021; McKenzie, 2023).

Another example comes from fossil-fuel companies like Exxon Mobil, which are driving the climate crisis. This crisis now threatens civilization itself, and could greatly exacerbate agential risk. As many scholars and government officials have noted, climate change is a threat multiplier and threat intensifier that, as such, will likely foment conflict around the world, which could take the form of interstate wars, civil wars, and even apocalyptic terrorism. The religious scholar Mark Juergensmeyer, for example, argues that extreme climatic alterations could trigger terrorism motivated by "active" apocalyptic convictions, whereby apocalypticists see themselves as active participants in bringing about the world's end (Juergensmeyer, 2017). There is, in fact, already an example of this: one can draw a straight line between climate-change caused droughts in Syria and the emergence of the Islamic State (Holthaus, 2015). In other words, Agents of Doom like Exxon Mobile created the conditions in which ISIS was able to arise.

If a global catastrophe were to occur in the next few decades, it would likely be the direct result of these Agents of Doom, rather than risky agents of the sort discussed above. But will this change in the further future? My view is that global catastrophic and existential risk scholars have been too dismissive of the aforementioned Agents of Doom. A great deal of the blame for any global catastrophe that might occur in the twenty-first century should be placed on such agents, whether or not they are the proximate cause. That said, I would also contend that Kemp does not take agential risk seriously enough. The fact is that there *really are* people in the world—a relatively small percentage of the population, but a potentially enormous demographic in absolute terms—who would, if only they could, destroy the world. Some of these individuals will *actively seek out* the means of ruination, as indicated by GLF's and the Islamic State's fantasies about weaponizing pathogens and acquiring nuclear weapons. In the case of Aum, the group built secret laboratories, bought a sheep ranch in Australia where they tested nerve gas on animals, tried to "procure the [E]bola virus and cholera, create a laser weapon, mine uranium, and measure plutonium." They also perpetrated at least 14 biological and chemical weapons attacks involving "anthrax, botulinum toxin, phosgene gas, sarin gas, and VX." When police raided Aum's compound, they "found enough chemicals to produce quantities of sarin that could kill 4 million people" (Torres, 2018b).

Dual-use emerging technologies will make it easier, not harder, to acquire weapons of mass destruction (WMD) or even weapons of total destruction (WTDs). Here we may return to the calculations above, according to which a 0.01 chance of any one of 500 individuals successfully destroying the world would yield a 99% probability of doom per decade. As an article in the *MIT Technology Review* observes, summarizing the results of Sotos (2019):

If there is a one in 100 chance that somebody with the technology will release it, and there are a few hundred individuals like this, then our civilization is doomed on a timescale of 100 years or so. If there are 100,000 individuals with this technology, then the probability of them releasing it needs to be less than one in 10^9 for our civilization to last 1,000 years (MIT, 2017).

What I am advocating for, then, is a more comprehensive and nuanced picture of the threat environment, one that takes seriously the links and interactions between (a) agential risks, and (b) Agents of Doom within the growing ecosystem of global catastrophic risk. Many Agents of Doom are enabling groups and individuals to *become* risky, in the relevant sense, while also posing direct and profound risks themselves, even if their aim is not to kill millions, destroy civilization, and so on. To counteract this, our primary aim in the near term should be to address the root causes of this predicament, namely, the "military-industrial complexes, fossil fuel companies, and Big Tech" companies that are responsible for our increasingly precarious existential predicament. This is further motivated by the fact that there are currently no good proposals for how to neutralize risky agents *other than* mass surveillance, which one should, I would argue, see as a dystopian nightmare that we ought to avoid at all costs (Torres, 2018c; cf. Bostrom, 2019). If there is no good way to prevent risky agents with access to WMDs (weapons of mass destruction) or WTDs (weapons of total destruction) from precipitating global catastrophes, then perhaps we should focus on imposing moratoriums or enacting international treaties designed to prevent WMDs and WTDs from being developed in the first place. This is not impossible: (a) there is nothing *forcing* military-industrial entities or private corporations like OpenAI from developing extremely powerful dual-use technologies, and (b) history shows that banning certain technologies can be effective, as exemplified by the 1972 Biological Weapons Convention and 1993 Chemical Weapons Convention (see Joy 2000). Failing this, the probability of doom-soon appears to be very high indeed.

4. Conclusion

Humanity is racing into an unprecedented situation, given the rapidly accelerating democratization of science and technology. The sorts of individuals who, throughout history, have wished to destroy the world will soon acquire this capability (in some cases, the capability is already within reach). It is, therefore, urgent for scholars to study the sorts of individuals or groups that would, accidentally or by intention, push a doomsday button. But I agree with Kemp that the underlying generators of risk are the aforementioned Agents of Doom, who may themselves become the proximate cause of a global catastrophe in the near future. A complete understanding of our quickly evolving existential predicament thus requires attentiveness to both agential risks and Agents of Doom. My hope in this short paper is to encourage further work on this important yet neglected topic.

References

Bender, E., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23.

Bostrom, N. (2019). The Vulnerable World Hypothesis. *Global Policy*, 10(4): 455–76.

Campbell, S. (2017). There's Nothing Anarchist about Eco-Fascism: A Condemnation of ITS. *It's Going Down*. https://itsgoingdown.org/nothing-anarchist-eco-fascismcondemnation/.

Flannery, F. (2015). *Understanding Apocalyptic Terrorism: Countering the Radical Mindset*. Routledge, 2015.

Gill, K. (2005). Kimveer Gill" *School Shooters*. https://schoolshooters.info/sites/default/files/Kimveer%20Gill%20Online%20Electronic%20Version.pdf.

GLF. (1994). Statement of Purpose (A Modest Proposal). http://www.churchofeuthanasia.org/resources/glf/glfsop.html.

Holthaus, E. (2015). New Study Says Climate Change Helped Spark Syrian Civil War. *Slate*. https://slate.com/technology/2015/03/study-climate-change-helped-spark-syrian-civil-war.html.

JNZ. (2019). Genocide Is the Only Way to Save the Human Race. Reddit. https://www.reddit.com/r/unpopularopinion/comments/ak6rei/genocide_is_the_only_way_to_save_the_human_race/.

Joy, B. (2000). Why the Future Doesn't Need Us. *Wired*, 8. https://www.wired.com/2000/04/joy-2/.

Juergensmeyer, M. (2017). Radical Religious Responses to Global Catastrophe. In R. Falk, M. Mohaty, and V. Faessel (ads), *Exploring Emerging Global Thresholds: Toward 2030*. Hyderabad, India: Orient BlackSawn,

Kaczynski, T. (1995). Industrial Society and Its Future. *Washington Post*. http://editionshache.com/essais/pdf/kaczynski2.pdf.

Kemp, L. (2021). Agents of Doom: Who Is Creating the Apocalypse and Why. BBC. https://www.bbc.com/future/article/20211014-agents-of-doom-who-is-hastening-the-apocalypse-and-why.

Korda, C. (1994). e-sermon #9. *The Church of Euthanasia*. http://www. churchofeuthanasia.org/e-sermons/sermon9.html.

Langman, P. (2014). Elliot Rodger: A Personality Analysis. *School Shooters*. https://schoolshooters.info/sites/default/files/rodger_personality_analysis_1.1.pdf.

Linkola, P. (2006). The Doctrine of Survival and Doctor Ethics. *PenttiLinkola.com*. http://www.penttilinkola.com/pentti_linkola/ecofascism_writings/translations/voisikoelamavoittaa_translation/VI%20%20The%20World%20And%20We/.

Lloyd, M., & Young, J. (2011). Nanotechnologists Are Targets of Unabomber Copycat, Alarming Universities. *The Chronicle of Higher Education*. https://www.chronicle.com/article/nanotechnologists-are-targets-of-unabomber-copycat-alarming-universities/.

Mcelroy, D. (2014). Islamic State Seeks to Use Bubonic Plague as a Weapon of War. *The Telegraph*. https://web.archive.org/web/20210308155753/https:/www.telegraph.co.uk/news/worldnews/middleeast/iraq/11064133/Islamic-State-seeks-to-use-bubonic-plague-as-a-weapon-of-war.html.

McKenzie, A. (2023). Transcript of Sam Altman's Interview Touching on AI Safety. *LessWrong*. https://www.lesswrong.com/posts/PTzsEQXkCfig9A6AS/transcript-of-sam-altman-s-interview-touching-on-ai-safety.

Milbank, D. (1994). In his solitude, a Finnish thinker posits cataclysms; What the world needs now, Pentti Linkola believes, is famine and a good war. *Wall Street Journal*.

MIT. (2017). Genetic Engineering Holds the Power to Save Humanity or Kill It. *MIT Technology Review*. https://www.technologyreview.com/2017/09/19/242009/genetic-engineering-holds-the-power-to-save-humanity-or-kill-it/.

MU. (2023). Artificial Intelligence Use Prompts Concerns. *Monmouth University Polling Institute*. https://www.monmouth.edu/polling-institute/reports/monmouthpoll_us_021523/.

Mukunda, G., Oye, K. A., & Mohr, S. C. (2009). What Rough Beast? Synthetic Biology, Uncertainty, and the Future of Biosecurity. *Politics and the Life Sciences*, 28(2).

OpenAI. (2018). OpenAI Charter. https://openai.com/charter.

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.

PFN. (2007). Extinguish Humans, Save the World. *Plausible Futures Newsletter*. https://plausiblefutures.wordpress.com/2007/04/10/extinguish-humans-save-the-world/.

Ramsden, J. (2009). *Applied Nanotechnology: The Conversion of Research Results to Products*. William Andrew Publishing.

Rees, M. (2003). *Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future in This Century--on Earth and Beyond*. Basic Books.

Rozo, M., & Gronvall, G. K. (2015). The Reemergent 1977 H1N1 Strain and the Gain-of-Function Debate. *MBio* 6(4): e01013-15.

Snyder, R. (2016). A Proliferation Assessment of Third Generation Laser Uranium Enrichment Technology. *Science & Global Security* 24(2): 68–91.

Sotos, J. (2019). Biotechnology and the Lifetime of Technical Civilizations. *International Journal of Astrobiology* 18(5): 445–54.

Torres, É. P. (2018a). Agential Risks and Information Hazards: An Unavoidable but Dangerous Topic? *Futures*, 95: 86–97.

Torres, É. P. (2018b). Who Would Destroy the World? Omnicidal Agents and Related Phenomena. *Aggression and Violent Behavior*, 39: 129–38.

Torres, É. P. (2019). Why an Existential Risk Expert Finds Hope in Humanity's Certain Doom. *OneZero*. Available at: https://onezero.medium.com/rebelling-against-extinction-d7e112979bed.

Torres, É. P., and M. Rees. (2017). *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks*. Pitchstone Publishing.

Torres, É. P. (2023). *Human Extinction: A History of the Science and Ethics of Annihilation*. Routledge.

Xia, L., Robock, A., Scherrer, K., Harrison, C., Bodirsky, B., Weindl, I., Jägermeyr, J., Bardeen, C., Toon, O., & Heneghan, R. (2022). Global Food Insecurity and Famine from Reduced Crop, Marine Fishery and Livestock Production Due to Climate Disruption from Nuclear War Soot Injection. *Nature Food*, 3(8): 586–96.

Stanford

Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Existential Risk: From Resilience to Antifragility

Dana Klisanin [1]*

Citation: Klisanin, Dana. Existential Risk: From Resilience to Antifragility. *Proceedings of the Stanford Existential Risks Conference* 2023, 50-59. https://doi.org/10.25740/dg438cb5918

Academic Editors: Steve Luby, Trond Undheim, Dan Zimmer

Funding: N/A

Conflict of Interest Statement: N/A

Informed Consent Statement: N/A

Acknowledgments: N/A

Author Contributions: N/A

Abstract: Cascading global crises pose serious threats to the human psyche. To learn more about our ability to respond to such crises, I examined global data in light of pre-, peri-, and post-disaster risk factors known to contribute to post-disaster mental illness. Findings reveal one-third of the global population is vulnerable to developing psychopathology in confrontation with cascading global crises. Although research indicates that most people recover from disasters, our capacity for resilience is known to be compromised by numerous factors, specifically those arising in tandem with cascading global challenges. To address this area of psychological risk, an expansion of our spectrum of wellbeing, from resilience to antifragility is proposed. Antifragility is postulated to be an inherent evolutionary capacity supported by research in post-traumatic growth. The emergent 'antifragile mindset' is presented as a theoretical framework within which we might navigate the mental health implications arising from cascading global crises, however further investigation is required, particularly into how diverse cultural context can shape and enhance our understanding.

Keywords: cascading global crises, existential threats, global mental health, antifragility, human potential

[1] Senior Researcher, Evolutionary Guidance Media R&D, 205 Hudson St. 7th Fl., New York, New York 10013; dana@evolutionaryguidancemedia.com.

* Correspondence: dana@evolutionaryguidancemedia.com

1. Introduction

Existential risks are events or phenomena with the potential to cause irreversible and widespread harm to human civilization, "one where an adverse outcome would either annihilate Earth-originating intelligent life *or permanently and drastically curtail its potential*" (Bostrom, 2002, p.2). Emphasis is added to draw attention to an understudied area of concern: *protecting and preserving human potential*.

Human potential is the fullness of possibility; the flowering of creativity. But foremost, at the individual level it includes the ability to love, to care for others, and to experience compassion. Taken to the global level it includes potentials that embody and celebrate our interdependence. Of caring for and regenerating Earth. Of peaceful planetary coexistence. Of being united in the celebration of our individual differences. Of interspecies communication. Of finding new forms of life beyond our planet. Human potential matters because it is *potential* that gives life *meaning and purpose*. Potential is interwoven into our existence, like air, such that we take for granted its integrality. We notice its absence only through its lack.

Global catastrophic risks (GCRs)—threats such as nuclear war, biotechnology hazards, pandemics, artificial intelligence, and climate change (Ord, 2020; Bostrom, 2014; Bostrom & Cirkovic, 2008), share a key characteristic that presents an inordinate amount of risk to actualizing human potential: the capacity to trigger cascading global crises. Unlike isolated crises or disasters, cascading global crises are continuous systemic failures that exacerbate each other, amplifying their overall impact and complicating efforts to mitigate or resolve them. Their scale, scope, and potential for cumulative impacts can trigger a continuous cycle of psychological trauma, the implications of which remain largely unexplored.

Despite growing recognition of the psychological impacts posed by cascading global crises (Gifford & Gifford, 2016; Doherty & Clayton, 2011; World Economic Forum, 2023), *resilience* is the sought-after psychological response to trauma; it is the standard-bearer and trimtab upon which most of our crises response systems are designed. In this paper, I take a critical look at resilience, investigating its capacity to support the psyche through such crises while simultaneously preserving human potential. In the following sections, the characteristics of cascading global crises are delinated, followed by an examination of the current state of global mental health regarding the pre, peri, and post-disaster risk factors that contribute to mental illness outcomes. The factors known to inhibit resilience are reviewed, and an expansion of the spectrum of psychological wellbeing to encompass the concept of *antifragility*—the ability to grow stronger in response to stressors (Taleb, 2012), is proposed. Although underexplored in psychological research, antifragility is postulated to be an evolutionary capacity inherent in the human psyche, supported by evidence from research on post-traumatic growth. Finally, I introduce the antifragile mindset (Klisanin, 2020), and highlight the need transdisciplinary collaboration to further explore this emerging theoretical model.

2. Cascading Global Crises and the Human Psyche

Why do cascading global crises pose such grave threats to the human psyche and our long-term potential?

To understand the threats posed by cascading global crises it is crucial to recognize their nature as "hyper-risks" involving a large number of interconnected factors and feedback loops that increase in complexity making them difficult to predict and manage (Helbing,

2013, p.51). Climate change and the recent COVID-19 pandemic are prime examples of such crises; both emerged due to complex interactions and entanglements between human beings, our systems, and our environment. The systemic nature of such crises gives rise to new risks and amplifies existing vulnerabilities (Intergovernmental Panel on Climate Change, 2023). For example, increasingly frequent and severe extreme weather events—floods, hurricanes, and heatwaves—can catalyze food shortages, cause displacement, and lead to resource conflicts. Any one of these could give rise to mental health issues, but taken together they have significant mental health implications including impaired cognitive function, post-traumatic stress syndrome (PTSD), and feelings of helplessness, anxiety, depression, and suicidal ideation (McMichael, 2013; Clayton, et al., 2017). While research on the mental health implications of COVID-19 is ongoing, it is known to have exacerbated mental health issues (Hiscott, et al, 2020; Nicola, et.al., 2020). During the first year of the pandemic alone, the global prevalence of anxiety and depression increased by 25% (World Health Organization, 2023).

Cascading global crises unleash levels of uncertainty and volatility which strain the human psyche, eroding our sense of control and the predictability we rely on in our daily lives. Unpredictable stressors can heighten anxiety, cause cognitive dysfunction, and provoke depressive symptoms. Furthermore, the scale and persistence of cascading global crises can foster a sense of inescapability, leading to feelings of hopelessness and despair. To compound these issues, the cumulative impacts of such crises exacerbates pre-existing inequalities, disproportionately affecting vulnerable populations, and amplifying mental health disparities (IPPC, 2023). To learn more about the potential impact of cascading global crises on global mental health, we turn to the literature on disasters, while noting their marked differences.

3. Disasters and Mental Health: Risk Factors

To gain a clearer understanding of the potential psychological impacts arising from cascading global crises, we can look to the risk factors known to contribute to mental health outcomes in the aftermath of disasters. Goldmann and Galea (2014) categorized these factors into three distinct, yet interconnected, phases: pre-, peri-, and post-disaster. Each phase interacts dynamically, shaping the overall risk profile and manifestations of post-disaster psychopathology.

In the following subsection, global data is used to provide a snapshot of the state of global mental health in regard to the *pre-disaster* risk factors. Afterwards, the *peri* and *post disaster* risk factors are reviewed in light of our current understanding of cascading global crises and their potential impacts. Please note that this list of risk factors identified by Goldmann and Galea (2014) (factors are italicized), does not encompass the entirety of their findings.

Pre disaster risk factors:

1. *Pre-existing mental health problems*: Global estimates suggest that around 5.0% of the population suffers from depression (World Health Organization, 2023), 4.0% from anxiety disorders (Dattani, et al., 2021), and 4% from PTSD. These numbers may be even higher due to unreported and undiagnosed cases. In the Western world, between 6.8% and 13.2% of the population is estimated to be using medication for depression (Brody & Gu, 2020), According to the World Economic Forum (2023) mental health problems are increasing.
2. *Gender*: Women, who constitute 49.7% of the world's population, (The World Bank, 2023) tend to have worse psychological outcomes, such as PTSD and depression,

following disasters. This trend is observed across age groups, types of disasters, and countries.

3. *Age*: Children, who represent 25.0% of the global population, (The World Bank, 2023) are particularly vulnerable to psychological problems like anxiety and depression following disasters due to their limited coping mechanisms.
4. *Socioeconomic status, ethnicity, and social support*: The World Bank (2023) estimates that around 24% of the global population lives in extreme poverty (less than $1.90 per day) or near-poverty (less than $3.20 per day). Individuals with low socioeconomic status, minority ethnic status, and low social support or poor relationships may be at a higher risk for developing post-disaster psychopathology. In underdeveloped countries, females in the under-15 population are more likely to live in poverty and receive less social support, with 12% of them living in extreme or near poverty.
5. *Prior traumatic or stressful experiences*: Individuals who have experienced traumatic or stressful events before a disaster may be at greater risk for developing post-disaster mental health problems. Approximately 1 in 3 women have been subjected to physical or sexual violence. Women who have experienced intimate partner violence were almost twice as likely to experience depression (World Health Organization, 2021). A staggering percentage of the female population has thus already experienced prior trauma and has heightened risk for negative mental outcomes in the event of cascading global crises.

To summarize, the global population has the following *pre-disaster risk factors* which predispose it to experience adverse mental health outcomes in confrontation with disasters: existing mental health issues, increased risk for women and youth, significant socioeconomic disparities, and an increasing population which has previously experienced traumatic or stressful events. While it is challenging to pinpoint the exact proportion of the global population at risk, the data suggest that at least one-third of the global population has heightened vulnerability – a conservative estimate that doesn't account for escalating mental health problems (World Economic Forum, 2023).

Peri disaster risk factors

Peri-disaster risk factors refer to the degree or severity of exposure to the disaster, including factors like the number and intensity of disaster-related events, type of disaster, duration of exposure, death toll, and proximity to the disaster site. Such intense exposure often predicts a higher risk of psychopathology (Goldmann & Galea, 2014). Cascading global crises are the "perfect storm" in so far as peri-disaster risk factors are concerned. As previously discussed, cascading global crises are hyper-risks; they are severe, have high mortality rates, and are likely to involve proximity to the disaster site. Furthermore, cascading global crises involve ongoing systemic failures which exacerbate each other and are difficult to resolve. Thus, cascading global crises involve an increasing *number* of disasters and a longer *duration*.

Post-disaster factors

Post-disaster factors include ongoing stressors and access to social support. Stressors such as job loss, property damage, marital issues, health conditions linked to the disaster, and displacement often affect those who have experienced a disaster (Goldman & Galea, 2014). These stressors heighten vulnerability to mental health conditions like PTSD and depression. The presence of ongoing stressors after a disaster can also influence the long-

term course of psychopathology (Goldmann & Galea, 2014). Due to the systemic nature of cascading global crises those experiencing such crises are likely to lose support systems and with them, access to basic needs including water, food, shelter, and health care. Furthermore, cascading global crises can disrupt information and communication systems, transportation systems, and other infrastructure, exacerbating the preceding stressors. Taken together, it is reasonable to suggest individuals experiencing cascading global crises have increased risk of experiencing post-disaster risk factors.

Summary of Risk Factors

The preceding analysis of *pre*, *peri,* and *post disaster risk factors* suggests 1) a sizable portion of the contemporary global population is already predisposed to experiencing adverse psychological effects in the face of cascading global crises and, 2) the human psyche is prone to psychopathology in the event of cascading global crises. Currently, the global population is not optimally positioned to confront cascading global crises from a mental health perspective.

In the coming decade, the effects of our collective failure to mitigate climate change are projected to intensify (World Economic Forum, 2023). This will likely increase the percentage of the global population experiencing one or more forms of previously described trauma. Moreover, additional threats are rapidly emerging, including the malicious use of widely available artificial intelligence and the proliferation and potential use of nuclear weapons. Individuals living through such crises will experience multiple traumatic or stressful events creating a vicious cycle of unpredictable forms of trauma. These repeated, unpredictable forms of trauma could induce a psychological winter, the fallout of which may have long-term repercussions for human survival and potential. This raises the question*: Is resilience sufficient to protect and preserve human potential in confrontation with cascading global crises?*

4. From Resilience to Antifragility

Psychological resilience research is extensive, yet divisions exist regarding definitions, concepts, and theories (Fletcher & Sarkar, 2013). For our purposes, resilience is broadly defined as the ability to withstand, recover from, and adapt to adversity. Psychological resilience has long been considered a vital component of psychological health in the face of crises (Masten, 2001); and is an essential aspect of managing disasters, enabling individuals and communities to adapt and recover in the aftermath of traumatic events. Importantly, studies have demonstrated that most individuals exhibit resilience, even when confronted with extremely adverse life experiences (Bonanno, 2004).

Nevertheless, the efficacy of resilience is compromised by several factors, including the severity of the traumatic event, cumulative or chronic stress, pre-existing vulnerabilities, limited social support, ineffective coping mechanisms, and resource scarcity (Bonanno, et al., 2010). These factors closely align with the pre-, peri-, and post-disaster risk factors previously discussed, indeed the factors that challenge resilience mirror those that amplify the risk of post-disaster psychopathology.

Due to their systemic nature cascading global crises place immense strain on existing systems, increasing the likelihood of individuals encountering multiple resilience-undermining factors concurrently. Moreover, resilience often emphasizes a return to a previous state of normalcy (Norris et al., 2008). However, in the context of cascading global crises, 'normalcy' may either be extinct or dramatically transformed, rendering an

approach focusing on 'bouncing back' to a previous state impossible. These overlapping risk factors suggest that the traditional resilience framework may be insufficient to address the complex and interconnected nature of cascading global crises.

Before moving on, I wish to note that psychological resilience is one facet of resilience thinking in addressing existential risk; others include personal physical resilience (maintaining good physical condition), household resilience (stockpiling emergency supplies), community resilience (access to local food sources and supportive neighbors), and societal resilience (energy supply independent of global fossil fuel trade). For example, Mayunga's (2007) capital-based approach aims to build resilience at multiple levels, offering practical strategies for resilience enhancement. The focus here, on psychological resilience, does not aim to diminish the value of broader resilience thinking, but rather endeavors to expand the scope of our response capabilities—toward antifragility.

4.1 Antifragility

Antifragility, introduced by Taleb (2012) describes systems that become stronger and more resilient when exposed to stress, volatility, and shocks. Such systems can not only withstand and recover from adversity but also grow and thrive in the face of stressors. In describing antifragility, Taleb (2012) compared it to the Hydra, a mythical creature from ancient Greek mythology, often depicted as a serpent or dragon with multiple heads. According to the myth, when one of the Hydra's heads was severed, two new heads would grow back in its place. This regenerative ability made the Hydra a seemingly invincible adversary for its opponents; rather than merely being resilient, or bouncing back, Hydra grew stronger with each attack. The Hydra's regenerative ability serves as a powerful metaphor for antifragility, representing a system in which challenges and stressors are not merely endured but harnessed for growth and adaptation.

To find antifragility in living systems, we can look to the real-life Hydra, a small freshwater invertebrate, which possesses extraordinary regenerative capabilities. Hydras are regenerate lost body parts, including their heads, through a process called morphallaxis (Galliot, 2012). This process involves the reorganization and differentiation of existing cells, allowing the Hydra to re-establish its original form and function after being dismembered. The Hydra's regenerative abilities are an embodiment of antifragility in the biological world. As a living organism, the Hydra faces numerous challenges and stressors in its environment, yet it has evolved the capacity to regenerate and adapt, showcasing the potential for organisms to grow from adversity.

Antifragility is being applied in numerous fields, including community-building, urban planning, risk management, and sustainability (Fortunato & Alter, 2022; Blečić & Cecchini, 2017; Aven, 2014; Platje, 2015), however its application to the human psyche is in the nascent stage (Markey-Towler, 2018; Klisanin, 2020; 2023). However, the leap from observing antifragility in a tiny freshwater creature to considering its application in human psychology is not as wide as it might seem. While we cannot grow a new head to replace a severed head, the human mind is capable of extraordinary growth and adaptation in response to stress and adversity. Our ability to grow stronger in the face of stressors is well-documented by resilience researchers. Many of whom have found that individuals can experience post-traumatic growth or positive psychological changes in the aftermath of adversity (Tedeschi & Calhoun, 2004; Joseph, & Linley, 2006; Bonanno et. al., 2010). In other words, researchers have been investigating antifragility under the nomenclature of resilience and post-traumatic growth. Antifragility appears to be an

inherent evolutionary capacity. One that expands our ability to confront cascading global crises, let's take a closer look.

4.2 Fragility-Resilience-Antifragility: An Integral Spectrum

While we may have the psychological capacity to embody antifragility in confrontation with of adversity, being fully human means experiencing a range of emotions and mental states. The same potentials that give our lives meaning and purpose, i.e., love, caring, compassion, lead us to experience profound grief in times of loss. We have seen that cascading global crises lead to tremendous upheaval in which we are likely to experience the loss of family members, home, and community. Disruption of social services are likely to leave us without food, shelter, and medical care – and unable to provide for those in need. Recognizing this, we can only extend the psyche's potential for growth amidst adversity through acknowledging suffering. Resilience and antifragility are interdependent with fragility and can be best understood as nodes in a spectrum of wellbeing.

At any given moment, we may fluctuate along this spectrum from fragility, through resilience, to antifragility, demonstrating our inherent capacity to not only endure and recover from challenges but also to learn, grow, and thrive because of them. This fluctuation is not linear but multi-dimensional, mirroring the complexity of our lived experiences. The spectrum is a dynamic continuum in which fragility, resilience and antifragility are ever-present potentials. To access these potentials, we must be able to *think beyond paradox*. This ability can be learned and is one aspect of the antifragile mindset.

4.3 Antifragile Mindset

What are the characteristics that support antifragility? This question took center stage for me during the COVID-19 pandemic and led to the antifragile mindset (Klisanin, 2020), an integrative theoretical framework. While research is ongoing, the antifragile mindset is understood to be a type of growth mindset (Dweck, 2006) with four key components, identified during the pandemic (Klisanin, 2021a; 2023):

1. Heightened reliance on Character Strengths and Virtues: Individuals exhibited increased reliance on character strengths, including critical thinking and creativity to support wellbeing (Klisanin, 2021b).
2. Heightened reliance on the natural world: Individuals spent more time outdoors.
3. Heightened ability to adjust personal narratives: Individuals reimagined their narratives in response to changing situations at home, work, school, et.
4. Heightened reliance on futures thinking: Individuals exhibited a future-oriented perspective, proactively taking precautions to protecting their health, relocating, and so forth.

The antifragile mindset is a construct that emerged from the crucible of the pandemic, and, with additional research, may help us navigating the unpredictable psychological landscapes of cascading global crises.

5. Discussion

Humanity has survived and thrived despite major natural disasters, World Wars, and other calamities. While sharing commonalities with such crises, cascading global crises, are much more destructive to the human psyche. Climate change, and the COVID- 19

pandemic exemplify cascading global risks and have led to a host of negative mental health outcomes, including anxiety, depression, PTSD, and suicidal ideation. A review of global data in terms of the pre-,peri, and post-disaster risk factors contributing to post-disaster mental illness outcomes revealed at least one-third of the global population is at risk of experiencing psychopathology in confrontation with cascading global crises. Meanwhile, although most individuals are known to recover from disasters, the effectiveness of resilience is undermined by various factors (e.g., severity of the traumatic event, cumulative or chronic stress, pre-existing vulnerabilities, limited social support, ineffective coping strategies, and lack of access to resources), all of which are potential systemic implications of cascading global crises. These findings suggest resilience may not afford the human psyche the protection it needs in confrontation with such crises.

However, research indicates some individuals experience post-traumatic growth or positive psychological changes because of their struggle with adversity. This evidence supports the existence of antifragility as an evolutionary capacity within the human psyche. Antifragility is not suggested as an end state but rather as a node, or capacity within an integral spectrum. Antifragility is thought to build upon research in resilience and post-traumatic growth. Cultural context plays a significant role in our understanding of resilience, adversity, and growth, thus while antifragility may be new to our lexicon, its meaning may be well understood and established in other cultures; likewise the antifragile mindset. Ongoing transdisciplinary research will enable us to co-create, shape and enhance our understanding of these constructs.

Protecting and preserving human potential is among the aims of the existential risk community; to achieve this aim more attention needs to be placed on the human psyche and global mental health. Cascading global crises present novel risks to the psyche. They are as unlike the crises of our ancestors, as we are unlike our ancestors. Data suggests resilience is insufficient to protect and preserve human potential. By recognizing our psychological fragility in the face of cascading global crises and supporting further research into the psychology of antifragility, the existential risk community can better understand and mitigate threats, ensuring the survival and flourishing of our species in an increasingly complex and uncertain world.

References

Aven, T. (2014). *Risk, surprises and black swans: fundamental ideas and concepts in risk assessment and risk management*. Routledge.

Blečić, I., & Cecchini, A. (2017). On the antifragility of cities and of their buildings. *City, Territory and Architecture*, *4*, 1-11.

Bonanno, G. A. (2004). Loss, Trauma, and Human Resilience: Have We Underestimated the Human Capacity to Thrive After Extremely Aversive Events? *American Psychologist, 59*(1), 20–28. https://doi.org/10.1037/0003-066X.59.1.20

Bonanno G.A., Brewin C.R., Kaniasty K., Greca A.M. (2010) Weighing the Costs of Disaster: Consequences, Risks, and Resilience in Individuals, Families, and Communities. *Psychol Sci Public Interest*. 2010 Jan;11(1):1-49. doi: 10.1177/1529100610387086.

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

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford University Press.

Bostrom, N., & Cirkovic, M.M. Eds., (2008). *Global Catastrophic Risks*. OUP Oxford.

Bratman G.N., Anderson C.B., Berman M.G., Cochran B., de Vries S., Flanders J., Folke C., Frumkin H., Gross J.J., Hartig T., Kahn P.H. Jr., Kuo M., Lawler J.J., Levin P.S., Lindahl T., Meyer-Lindenberg A., Mitchell R., Ouyang Z., Roe J., Scarlett L, Smith J.R., van den Bosch M., Wheeler B.W., White M.P., Zheng H., Daily G.C. (2019). Nature and mental health: An ecosystem service perspective. *Sci Adv.* 2019 Jul 24;5(7):eaax0903. doi: 10.1126/sciadv.aax0903.

Brody D.J., & Gu Q. (2020). Antidepressant Use Among Adults: United States, 2015-2018. NCHS Data Brief. Sep;(377):1-8. PMID: 33054926. Retrieved May 20, 2023 from, https://www.cdc.gov/nchs/products/databriefs/db377.htm

Clayton, S., Manning, C. M., Krygsman, K., & Speiser, M. (2017). *Mental Health and Our Changing Climate: Impacts, Implications, and Guidance*. Washington, D.C.: American Psychological Association, and ecoAmerica. Retreived May 19, 2023 from, https://www.apa.org/news/press/releases/2017/03/mental-health-climate.pdf

Dattani, S., Ritchie, H., and Roser, M. (2021). *Mental health. Our world in data.* https://ourworldindata.org/mental-health

Doherty, T. & Clayton, S. (2011). The Psychological Impacts of Global Climate Change *The American psychologist* 66 (4) pp. 265-76, doi: 10.1037/a0023141

Dweck, C. (2006). *Mindset: The New Psychology of Success*. Random House.

Fletcher, D. & Sarkar, M. (2013). Psychological resilience: A review and critique of definitions, concepts, and theory. *European Psychologist* 18: 12-23.

Fortunato, M. W., & Alter, T. R. (2022). Embracing Uncertainty and Antifragility in Rural Innovation and Entrepreneurship. In *Building Rural Community Resilience Through Innovation and Entrepreneurship* (pp. 91-116). Routledge.

Galliot B. (2012). Hydra, a fruitful model system for 270 years. *Int. J. Dev. Biol*. 56: 411-423. https://doi.org/10.1387/ijdb.120086bg

Gifford, E. & Gifford, R. (2016) The largely unacknowledged impact of climate change on mental health, *Bulletin of the Atomic Scientists*, 72:5, 292-297, DOI: 10.1080/00963402.2016.1216505

Goldmann E. & Galea S. (2013). Mental health consequences of disasters. *Annu Rev Public Health*. 2014;35:169-83. doi: 10.1146/annurev-publhealth-032013-182435.

Helbing, D. (2013). Globally networked risks and how to respond. *Nature*. 497. 51-9. 10.1038/nature12047

Hiscott J., Alexandridi M., Muscolini M., Tassone E., Palermo E., Soultsioti M., Zevini A. (2020). The global impact of the coronavirus pandemic. *Cytokine Growth Factor Rev.* 2020 Jun;53:1-9. doi: 10.1016/j.cytogfr.2020.05.010.

Intergovernmental Panel on Climate Change (IPCC). (2023). Synthesis report of the IPCC sixth assessment report (AR6) https://report.ipcc.ch/ar6syr/pdf/IPCC_AR6_SYR_LongerReport.pdf

Joseph, S. & Linley, P.A. (2006) Growth Following Adversity: Theoretical Perspectives and Implications For Clinical Practice. *Clinical Psychology Review*, 26, 1041-1053. http://dx.doi.org/10.1016/j.cpr.2005.12.006.

Klisanin, D. (2020, March 16). The Antifragile Mindset: Why resilience isn't enough in the face of a global pandemic. *Psychology Today*. Online. https://www.psychologytoday.com/us/blog/digital-altruism/202003/the-antifragile-mindset

Klisanin, D. (2021a, October 27). *Psychological futures: Antifragility and the future of wellbeing.* [Conference presentation] World Futures Studies Federation World Conference.

Klisanin, D. (2021b, August 12). *Character Strengths and Virtues Supporting Human Wellbeing Through Covid-19*, [Conference presentation] APA 2019. Online.

Klisanin. D. (2023). Psychological Futures: The Antifragile Mindset and the Imperative of Interdependence. In Montuori. A., and Donnelly, G. *The Handbook for Creative Futures*. New York: Routledge.

Markey-Towler, B. (2018). Antifragility, the Black Swan and psychology. *Evolutionary and Institutional Economics Review*, Springer, vol. 15(2), pages 367-384.

Masten, A. S. (2001). Ordinary magic: Resilience processes in development. *American Psychologist*, 56(3), 227–238. https://doi.org/10.1037/0003-066X.56.3.227

Mayunga, Joseph. (2007). Understanding and applying the concept of community disaster resilience: A capital-based approach. *Summer Academy for Social Vulnerability and Resilience Building*. 1-16.

McMichael, A. J. (2013). Globalization, climate change, and human health. *The New England Journal of Medicine*, 368(14), 1335–1343. https://doi.org/10.1056/NEJMra1109341

Nicola M., Alsafi Z., Sohrabi C., Kerwan A., Al-Jabir A., Iosifidis C., Agha M., & Agha R. (2020). The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *Int J Surg*. 78:185-193. doi: 10.1016/j.ijsu.2020.04.018.

Norris, F. H., Stevens, S. P., Pfefferbaum, B., Wyche, K. F., & Pfefferbaum, R. L. (2008). Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. *American journal of community psychology*, *41*, 127-150.

Ord, Toby. *The Precipice*. Hachette, 2020.

Platje, Joost. (2015). Sustainability and antifragility. *Economic and Environmental Studies*. 15. 469-477.

Taleb, N. (2012). *Antifragile: Things that gain from disorder*. London: Penguin Books.

Tedeschi, R. G., & Calhoun, L. G. (2004). Posttraumatic Growth: Conceptual Foundations and Empirical Evidence. *Psychological Inquiry*, 15(1), 1– 18. https://doi.org/10.1207/s15327965pli1501_01

The World Bank (2023). The World Bank, Population, female (% of the total population). Retrieved March 6, 2023, https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS

The World Bank (2023). The World Bank. Population ages 0-14, (% of the total population), Retrieved March 6, 2023, https://data.worldbank.org/indicator/SP.POP.0014.TO.ZS

World Economic Forum (2023). The Global Risks Report, https://www3.weforum.org/docs/WEF_Global_Risks_Report_2023.pdf

World Health Organization (2023). https://www.who.int/en/news-room/fact-sheets/detail/depression

World Health Organization (2021). Violence against women. https://www.who.int/news-room/fact-sheets/detail/violence-against-women

Stanford
Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Psychological and Psychosocial Consequences of Super Disruptive A.I.: Public Health Implications and Recommendations

David D. Luxton [1]*, Eleanor Watson [2]

Citation: Luxton, D. D., Watson., E; Psychological and Psychosocial Consequences of Super Disruptive A.I.: Public Health Implications and Recommendations. Proceedings of the Stanford Existential Risk Conference 2023, 60-74. https://doi.org/10.25740/mg941vt9619

Academic Editor: Trond Undheim, Dan Zimmer

Funding: Not applicable.

Conflict of Interest Statement: The authors declare no conflict of interest.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors wish to extend their gratitude to the SERI Team.

Author Contributions: Conceptualization, D.L., and E.W.; writing—original draft preparation, D.L. and E.W.; writing—review and editing, D.L. and E.W. All authors have read and agreed to the published version of the manuscript.

Abstract: A moral panic is burgeoning now that the disruptive impacts of A.I. are becoming unmistakable, and further advancements shall have increasingly profound ramifications on how we interact with one another, how we experience our lives, and how we perceive the future and the world around us. With these profound, existential changes, there will also be implications for the psychological and psychosocial well-being of people around the world. The psychological and psychosocial impacts of A.I. from a public health perspective deserve attention. The purpose of this paper is, therefore, to begin to elucidate these issues and provide recommendations. Topics discussed include the mechanisms of psychological and social impacts, including threats to interpersonal and societal trust, problems of deception, overreliance on machines in decision-making, and the public health risks caused by the displacement and disenfranchisement of persons in work sectors most impacted by A.I. Recommendations are presented for addressing these emerging issues to be considered by technology developers, policy-makers, ethicists, healthcare clinicians, politicians, and the general public.

Keywords: artificial intelligence, psychology, well-being, mental health, disruption, shock

[1] University of Washington, School of Medicine, Department of Psychiatry and Behavioral Sciences, 1959 NE Pacific St, Seattle, WA 98195, United States; ddluxton@uw.edu (D.L)

[2] University of Gloucestershire, School of Computing and Engineering, The Park, Cheltenham, GL50 2RH, United Kingdom; eleanorwatson@connect.glos.ac.uk (E.W)

* Correspondence: ddluxton@uw.edu

1. Introduction

The advent of new technology has often brought about significant societal changes and a range of positive and negative impacts. In some cases, the introduction of new technology has resulted in the creation of unanticipated social problems. Unfortunately, as history has all too often demonstrated, awareness of the potential negative consequences of any new technological innovation, and thus opportunities to mitigate those consequences, is often lacking, and it is only after a new technology is put to use or brought to market that the problems become evident. For example, the rise of the Industrial Revolution in the 18th and 19th centuries led to the growth of cities and the development of new forms of work and leisure, as well as new social problems such as pollution, overcrowding, and the exploitation of workers. Similarly, the widespread adoption of the Internet and the rise of social media in the 21st Century have significantly improved how we communicate and access information but have also contributed to new social problems impacting public health, such as Internet addiction, cyberbullying, online harassment, self-directed violence, and social isolation (Luxton et al., 2012; Schou Andreassen et al., 2016). While technological impacts on societal change are often gradual and incremental, such as with the examples mentioned above, history has shown that sudden technological disruption can cause intense reactive fear and anxiety, such as following the detonation of the first atomic weapons and the launch of Sputnik in the mid-20th Century.

The rise of artificial intelligence (A.I.) has also been incremental and with leaps, and the topic of both enthusiasm and consternation. The last few years have been especially exciting given the significant advancements in machine learning (ML), natural language processing, and computer vision, resulting in a major impact on the global economy (Dean, 2019; The White House, 2022). The development of powerful prompt-driven multimodal generative A.I. technologies, such as ChatGPT, has the potential to significantly transform the way in which we provide, receive, and share information in social interactions. One key impact of these technologies is their ability to generate a wide range of content, such as images, videos, and audio, based on a given prompt. This can enable the creation of new forms of communication and expression that may not have been possible before and can also lead to the automation of certain tasks that are currently performed by humans, such as content creation and moderation. Another potential impact of these technologies is their ability to interact with users in a dynamic and personalized way. This can enable the creation of immersive and interactive social experiences that can adapt to the interests and preferences of individual users.

Just like previous technological advancements, the latest in A.I. developments are having a range of impacts on society, both positive and negative. On the positive side, A.I. has the potential to increase efficiency and productivity by automating tasks and making better decisions than humans. This can lead to cost savings and increased competitiveness for businesses and governments that adopt A.I. technologies. A.I. can also create new opportunities for innovation and growth, as it can enable the development of new products and services across domains, such as art and entertainment, business, transportation, and healthcare, that were not previously possible.

There are also potential negative impacts of A.I. on society, such as the potential for job displacement as A.I. becomes more prevalent in various industries. This could lead to workers losing their jobs or requiring retraining for new roles, which could negatively affect their financial well-being. In addition, there are also concerns about the potential for A.I. to perpetuate and amplify biases and prejudices, which could lead to unfair and discriminatory outcomes that disproportionately affect certain groups of people. The increasing use of A.I. in various areas of society also carries with it the potential for psychological harm to individuals due to the potential loss of traditional humanitarian

values. These values may include the principle of putting people first, the value of autonomy and self-determination, the right to privacy, and the importance of pride in one's work.

Given the rapid emergence of generative A.I. and leaps in associated enabling technologies, such as cloud-based super-computing and virtual reality, sudden technological super-disruption is a possibility. We define super-disruptive A.I. as an advance in A.I. that profoundly and unexpectedly alters societal functioning at a large scale, whether caused by an incredible technological leap or incremental progression with cascading effects on society. We recognize that the impact of A.I. on society is complex and multifaceted, with both negative and positive outcomes as possibilities. It is therefore essential to consider the potential benefits and risks of A.I. carefully and to take steps to mitigate any negative consequences, such as investing in education and training programs to help workers adapt to the changing job market, as well as guidelines, regulations, and policies to ensure that the development and deployment of A.I. are responsible and ethical.

Our goal with this paper is to illuminate the psychosocial and psychological risks and implications of super-disruptive A.I., particularly those that have thus far received a lesser amount of discussion, such as the risk of supernormal stimuli experiences creating new social dysfunction and disparities in access to A.I. that have psychosocial implications. We highlight current risks and speculative near-future threats across varied societal domains and economic sectors, including entertainment, healthcare, business, criminal justice, public safety, and defense. We also posit an intersection between A.I. and the emergence of a new theater of war based upon demoralization, which seems likely to intersect with moral panic (with the risk of widespread paranoia or mass hysteria). Finally, we set an expectation for an imminent shock of an "AI Sputnik moment" and provide high-level recommendations to guide the public through this adjustment.

2. Current and Emerging Threats

2.1 Algorithmic Distortions of Reality and Behavioral Manipulation

Social media is one of the most pervasive and influential modern technologies underpinned by A.I. TikTok, and other social media platforms use AI-driven algorithms to create personalized content feeds designed to keep users engaged for as long as possible within filter bubbles of comfortably familiar ideas. These algorithms personalize content and user recommendations based on their past activity and behavior. While these algorithms can be beneficial in helping users discover new content and connections, they can also have negative consequences. Social media algorithms can contribute to the polarization of political views and the spread of misinformation. This issue can occur when algorithms prioritize content that aligns with a user's preexisting beliefs and values, creating echo chambers in which users are only exposed to information that confirms their views. This can lead to a narrowing of perspectives and a lack of exposure to diverse viewpoints and information, which can contribute to the polarization of opinions and the strengthening of extremist views (Wojcieszak, 2010). In addition, social media algorithms can also contribute to negative perceptions of self-worth by promoting a distorted view of reality (Luxton et al., 2012). For example, social media algorithms may prioritize content that is designed to be attention-grabbing or sensational, which can create a skewed perception of what is normal or desirable. This can lead to feelings of inadequacy or anxiety among users who may compare themselves to others based on their online profiles or posts (Braghieri et al., 2022). Social media use is also linked to depressive disorders and suicidal behavior, especially among young people (Sedgwick et al., 2019).

A.I. is also increasingly used to intentionally manipulate and control others, which can harm people while undermining or violating individuals' rights to freedom of expression, privacy, and non-discrimination. This can occur when tech platforms act as de facto gatekeepers of online speech and content and use their power to enforce their policies and regulations in an arbitrary, biased, or inconsistent manner. One example is the use of algorithms and automated systems to enforce content moderation policies, which can lead to the removal of legitimate speech or the failure to remove harmful or defamatory content. Moreover, using automated bots to prop up opinions and influence others via social media is also a threat. The use of A.I. to suppress speech in social media received much public attention with Elon Musk's purchase of the social media site Twitter in 2022. The subsequent release of the "Twitter Files" and Congressional hearing revealed deep gatekeeping and biased censorship, not only by the company but also by the influence of the government and powerful political individuals (Sardarizadeh & Schraer, 2022; U.S. House of Representatives, 2023).

The psychological and psychosocial impact of using A.I. to censor and manipulate online content can create a chilling effect on free expression, as individuals may self-censor or refrain from sharing particular ideas or perspectives for fear of being censored or punished by the platform. Psychological studies have shown that the suppression of free speech and expression contributes to negative emotional states (Krøvel & Thowsen, 2019; Maitra & McGowan, 2012; Parekh, 2017). It's firmly established those verbalizing thoughts and feelings increases a person's sense of control, increases understanding of the motivations of others, and thus reduces anxiety (Ayers, 2009; Niles et al., 2015; Lepore et al., 2000).

2.2 Algorithmic Prejudice and Unfairness

A risk of prejudice is presented by A.I. when it overfits on characteristics that are not germane to the question or which may be secondary indicators of protected characteristics (e.g., whether one buys certain beauty products). The use of A.I. in decision-making processes carries with it the risk of perpetuating and amplifying biases and prejudices that are present in the data and algorithms used to train and operate the A.I. system. This can occur when the A.I. system overfits on characteristics that are not relevant to the task at hand or which may be correlated with protected characteristics such as race, gender, age, or sexual orientation. Algorithmic prejudice is especially troublesome when human oversight is no longer present or has become complacent due to trusting a system too much. Examples such as the Horizon Post Office Scandal illustrate that prejudicial systems that wrongly label and harass people as having committed illegal actions can lead to enormous miscarriages of justice (Peachey, 2022).

Disparities in outcomes that result from algorithmic bias have the potential to cause harm to the health and well-being of people (Luxton & Poulin, 2020). For example, if an A.I. system is trained to identify health risks or to assist in making decisions about medical treatments, but the source data is biased due to under-representative learning samples, such as by race or socio-economic status, then some groups may suffer from less-than-optimal treatment options that lead to degraded health outcomes, or in some situations, no treatment at all. Another example is the use of algorithmic risk prediction in criminal justice systems. Algorithmic bias is a problem for systems that are used to determine level-of-risk for incarcerated persons who are considered for release into the community or when algorithms are used in the sentencing process (McKay, 2020). Actual bias or the perception of bias due to lack of transparency seeds distrust in people, which can lead to feelings of resentment and discontent among those who are negatively affected and thus exacerbate social tensions.

2.3 Supernormal Stimuli

There is a risk that people may become lost within endless fascinating generative worlds created by artificial intelligence (A.I.) and other technologies, leading to a phenomenon known as "irresistible supernormal stimuli." Having roots in animal behavior and evolutionary psychology (Barrett, 2010), this refers to the idea that people may become excessively drawn to and engaged with artificial stimuli designed to be more appealing or rewarding than their real-life counterparts. The risk of irresistible supernormal stimuli is particularly concerning in the context of A.I. technologies that can generate vast amounts of content, such as virtual reality environments or chatbots that can engage in endless conversations. These technologies can provide an endless source of stimulation and engagement, leading people to become excessively drawn to and reliant on them. The over-engagement with irresistible supernormal stimuli can have negative consequences for individuals' social and emotional well-being, as it may lead to a lack of connection and fulfillment in real-life relationships. It can also have negative impacts on mental health, as it may lead to feelings of isolation and disconnection from the real world.

2.4 Demoralization and Anomie

While A.I. can bring many benefits, including increased efficiency and productivity, it also carries with it the potential for negative consequences, including the risk of demoralization, humiliation, and discontent among certain segments of the population. One potential risk is that the increasing use of A.I. in various fields may lead to job displacement, as machines and algorithms are able to perform tasks more efficiently than humans. This can lead to feelings of anxiety and insecurity, as well as a loss of pride in one's work and a sense of meaninglessness. Another issue is the risk of traditional middle management being increasingly given to algorithms for expediency's sake, resulting in workplaces with potentially less employee autonomy and omnipresent scrutiny of the smallest details, including unfair recording of infractions. It is ironic that as machine autonomy advances, humans are increasingly finding themselves robbed of their own autonomy.

2.5 Alteration of Human Interactions and Trust

Disparities in access to A.I. and perceptions about another's use of it will undoubtedly alter everyday social interactions, the trust between people, and ultimately psychological well-being (Luxton, 2014). At an individual level, distrust may result when someone comes into contact with another person who's suspected of using A.I. to give them an unfair advantage.

Imagine a scenario when you meet a stranger who may or may not have a Neuralink™ implant that gives them instant access to cloud-based A.I., allowing them to have detailed knowledge about you. This issue has come into the public spotlight in past years with the development of smart glasses and other smart wearable technologies (Due, 2015; Nonan, 2013; Schuster, 2014). Another scenario is when a colleague may have access to A.I. that provides them with a superior edge over you in particular tasks or domains, from leisure (e.g., a poker game) to professional settings (e.g., a court proceeding). It is now feasible for people to access the capabilities of ChatGPT remotely, or use technology such as A.I. 'Cyrano de Bergeracs,' which can whisper answers into Bluetooth earbuds.

Perception of unfair access to and use of A.I. does not need to be in real-time during interpersonal interaction but at any time before or after an interaction. It is probable the uncertainty about whether another person is using A.I. to their advantage during an interaction; however, that may be most problematic from a psychological health

perspective. Knowing there is an advantage allows a person to adjust to and adapt, reducing psychological unease, whereas not knowing is likely to contribute to maladaptive stress, anxiety, and paranoia, consistent with scientific studies documenting the link between uncertainty and problems with mental health (Massazza et al., 2022).

The disparity in access to A.I. technology could also lead to shared distrust and antagonization towards groups perceived to have unfair advantages and privileges because of their access to A.I. If only wealthy elites or governments have access to advanced A.I. and solely possess authoritative decision-making regarding it, then a perception of unfair inequality and distrust among the population would ensue, as do disparities of any other resource (Mirza, et al., 2019; Bénabou, 2003; Hargittai, 2011; Van Dijk, 2006). However, a disparity in access or understanding of A.I. may be exacerbated by distrust in the use of technology by the elite to surveil and control others (Wee & Findlay, 2020; Feldstein, 2019; Matthews, 2021; Manheim & Kaplan, 2018).

Competition between governments and corporations to advance and deploy A.I. for strategic and tactical advantages may also have deleterious consequences for psychosocial well-being. This is similar to what was experienced during the nuclear arms race between the United States and the Soviet Union during the Cold War – each side continued to build their nuclear arsenals, promising peace through deterrence and assured mutual destruction. Indeed, the "A.I. arms race" has begun, and while development and strategy may remain clandestine and under secrecy, what results are increasingly aggressive and Machiavellian behaviors and too, fear and distrust among the general population.

The use of A.I. and associated technologies in criminal justice and policing also has implications for social distrust. For example, the increasing use of mass facial scanning and surveillance at public places, such as airports, has the potential to increase psychological distress in the forms of anxiety and distrust among citizens regarding how data about them is used. Surveillance is also linked to behavioral changes, including self-censorship and avoidance of assembly, and the right to protest for fear of retribution (Starr et al., 2008; O'Connor & Jahan, 2014; Kaminski & Witnov, 2015). And as we noted earlier, algorithmic biases that result in disparities in criminal justice outcomes have the potential to increase distrust and increase social unrest.

2.6 AI-enabled Deception

Artificial intelligence (A.I.)-enabled forms of deception, such as deepfakes, refer to the use of A.I. technologies to create and disseminate convincing, yet fake or misleading, images, videos, and other media. These technologies can be used to manipulate the appearance or content of media in ways that are difficult to detect, and that can potentially deceive or mislead viewers.

The use of AI-enabled deception can have a range of negative impacts on social trust and cohesion at the individual, organizational, and societal levels. For example, deepfakes can be used to create and disseminate non-consensual or malicious content, such as revenge porn or harassment, which can have serious consequences for the victims. This can lead to feelings of betrayal, mistrust, and fear, which can have deleterious effects on mental health while further undermining social cohesion. Another concern is that deepfakes and other forms of AI-enabled deception can be used to spread misinformation or propaganda, which can erode trust in information and institutions (Luxton, 2022). This can lead to the erosion of social norms and values, as well as to social polarization and division.

It is essential to be aware of these risks and to take steps to mitigate them. This can include measures such as regulations and policies to prevent the use of AI-enabled deception for harmful purposes, as well as efforts to promote media literacy and critical thinking skills.

2.7 Breakdown of Human Relationships

There is a risk that an over-affinity with A.I. waifus' (digital companions designed to be attractive and submissive) could lead to a breakdown in romantic and friendship relationships. This risk is particularly concerning in the context of incel identity, where individuals may feel isolated and rejected from romantic opportunities and turn to A.I. waifus' as a substitute for real-life relationships. This may have negative consequences for individuals' social and emotional well-being, as it may lead to a lack of connection and fulfillment in real-life relationships. This can also contribute to a breakdown in traditional dating styles, as individuals may rely more on digital interactions with virtual partners rather than seeking out real-life relationships. In addition, the use of "hook-up apps", which enable dating outside of one's social circle, can also contribute to a breakdown in traditional dating styles and may lead to a concentration of attention on a few algorithmic-selected matches. This can perpetuate a cycle of social isolation and frustration, as individuals may feel that they are not able to compete for attention and connection with others.

As a society, we remain particularly underprepared for the potential impacts of A.I. on certain vulnerable groups, such as lost youth and hikikomori. 'Lost youth' refers to young people who may feel disconnected from society and may lack a sense of purpose or direction in life. Hikkikomori refers to a social phenomenon in Japan where individuals, typically young men, become isolated and withdraw from society, often spending most of their time alone in their rooms (Teo & Gaw, 2010). This risk is particularly concerning in the context of A.I. technologies that are able to generate vast amounts of content, such as virtual reality environments or chatbots that can engage in endless conversations and consistent virtual relationships. The over-engagement with irresistible supernormal stimuli can have negative consequences for individuals' social and emotional well-being, as it may lead to a lack of connection and fulfillment in real-life relationships. Where on one hand, A.I. technologies such as virtual companions may provide social support and reduce social isolation; they can also have negative impacts on mental health, as they may lead to feelings of isolation and disconnection from the real world.

The use of AI-enabled virtual care providers in place of human care providers is another potential threat to human relationships (Luxton, 2014a; Luxton, 2014b). Virtual therapists in the form of chatbots apps and virtual human therapists are becoming increasingly capable and popular, and while they provide numerous benefits, such as 24/7 availability and customization, they also come with some drawbacks. A lack of human connection between a care provider, such as between a psychotherapist and patient, may minimize the humanness in the therapeutic process and potentially diminish the therapeutic relationship, which has been shown to be the strongest predictor of positive therapeutic outcomes (Luxton, 2014b).

2.8 Loss of Human Creativity, Inspiration, and Drive

Emerging generative A.I. is beginning to outdo all the creative achievements and potentials of humans, to include the visual arts, music, and literature. This new AI-generated art will rapidly devalue the creative expression of human artists, not just because it is technically or thematically impressive, but also because it is "cheap" to produce (Luxton, 2022). Many artists and writers have faced significant fiscal and emotional impacts from the present wave of prompt-driven A.I. systems. This economic

reality is depressing for creative types who will no longer be able to eke out a living making art, which is already a famously precarious vocation. Even those whose creative pursuits are merely a hobby may discover that they find their joy and sense of mastery eroded.

However, the future may not be all doom and gloom for human creativity. Generative A.I. can potentially augment human creativity (Dwivedi et al., 2023), perhaps inspiring it while eliminating the requirements of more mundane tasks that burden creativity. This can allow people to explore and experience creative expression in beneficial ways.

But perhaps most troubling moving forward is the devaluing and loss of the human experience recorded and communicated through art since time immemorial. Art communicates the joy, suffering, passion, and beauty of human experience, moving us toward an enlightened state of being. The human race will lose this expression with the ubiquity of AI-generated art. AI-generated art is devoid of human experience and inspiration, and it is soulless, just like an AI-generated virtual human (Luxton, 2022). Why should we care about what is depicted in AI-generated art? The human response to AI-generated art may soon transfer from, "Wow, how did it do that?" to, "Who cares?"

2.9 Dependence on Machines for Problem-Solving and Decision Making

The use of artificial intelligence (A.I.) in healthcare has the potential to transform the way in which healthcare is delivered and experienced by patients. One area in which A.I. could have a significant impact is in the dynamic personal tracking of health, which refers to the use of technologies such as the Internet of Things (IoT) and non-contact means, such as voice or movement analysis, to continuously monitor and track patients' health (Hilty et al., 2021). Dynamic personal tracking of health has the potential to produce temporal health metrics, which are data points that are collected over time and can provide insight into patterns and trends in patients' health. This can help to identify potential health issues early on, and can enable healthcare providers to intervene more effectively to prevent or mitigate negative health outcomes. Dynamic personal tracking of health may also have implications for patient autonomy and privacy, as it may involve the collection and analysis of sensitive personal data. It is important to ensure that appropriate safeguards are in place to protect patients' privacy and to ensure that the data collected is used in a responsible and ethical manner.

The use of artificial intelligence (A.I.) in sciences has the potential to transform the way in which knowledge is generated and transferred. One key aspect of this is the shift towards a bottom-up paradigm of science, in which data is analyzed to identify patterns, rather than following the traditional approach of formulating a hypothesis and testing it. This shift towards a bottom-up paradigm is enabled by the increasing availability of large amounts of data, as well as the development of machine learning algorithms that are able to identify patterns and make predictions based on this data. One potential benefit of this approach is that it can allow scientists to identify patterns and relationships that may not have been evident using traditional methods, and can enable the discovery of new knowledge. However, there is also a risk that this approach may lead to a reliance on machines to find answers to problems rather than on human expertise and understanding. Furthermore, there is a risk that machines may be able to find answers to problems but may not necessarily be able to explain why these answers are correct. This can make it difficult for humans to understand and interpret the results and can limit the transferability of knowledge.

In the context of labor and business, the use of A.I. has the potential to both enhance and disrupt various aspects of the employment relationship. One potential impact of A.I. is

the rise of algorithmic management, which refers to the use of algorithms to monitor and evaluate employee performance, and to make decisions about tasks and responsibilities (Lee et al., 2015). Algorithmic management can lead to the emergence of petty bureaucracies in which employees are subject to an increasing level of micromanagement and control. This can have negative consequences for employee autonomy and dignity, as it may reduce the ability of employees to make their own decisions and exercise their own judgment.

Another potential impact of A.I. in work and business contexts is the return of *Taylorism*, which refers to a management philosophy based on the idea of breaking down work into smaller tasks and optimizing efficiency through the use of scientific principles. The use of A.I. in tasks such as nudging, which refers to the use of subtle cues or incentives to influence behavior, may lead to a return of this type of management philosophy. The use of A.I. in work and business contexts may also disrupt middle management, as it may lead to the automation of specific tasks and responsibilities that are currently performed by middle managers. This can have negative consequences for the employment prospects and job security of middle managers, who are typically tasked with personnel management but with little leeway for strategic decision-making. There may be a perception amongst upper management of the opportunity to squeeze further performance out of staff whilst reducing middle management headcount through the use of automated management systems (Roberts & Shaw, 2022).

2.10 Psychological Warfare

The defense industry has a long history of Artificial intelligence (A.I.) innovation for various purposes, from weapon guidance systems, cyberwarfare deterrence, command and control of unmanned vehicles, robots, and more. These technological advancements have also benefited other industries through technology transfer. The psychosocial and psychological effects of the use of these technological advancements on persons, however, including adversarial combatants, operators of these technologies, and collateral persons, must not be overlooked. In particular, A.I. can enable a range of risks to psychosecurity, which refers to the protection of individuals' mental health and well-being.

The psychological effects of the use of unmanned aerial vehicles (UAVs) are one example that has emerged in the last twenty years. Mass traumatic stress experienced by collateral citizens during U.S. military operations in Afghanistan, Iraq, and Pakistan has been reported (Litz, 2007). The inability to observe the presence of circling aerial drones and uncertainty about imminent threats exacerbates the experience of anxiety and reinforces psychological distress (Ullah, 2016).

Another emerging risk is the use of A.I. to conduct Automated Zersetzung attacks, which are designed to demoralize and undermine targeted individuals (Dennis & LaPorte, 2011). Zersetzung (German for "decomposition" or "corrosion") refers to a range of psychological warfare tactics that are designed to undermine the mental health and well-being of targeted individuals or groups. These tactics may include harassment, intimidation, manipulation, and other forms of psychological abuse and can be conducted through a variety of means, including social media, email, and other online platforms. The goal of Zersetzung tactics is to demoralize and undermine the targeted individuals or groups and to create a sense of uncertainty, vulnerability, and fear. These tactics can be used to interfere with the ability of the targeted individuals or groups to function effectively and to disrupt social and political movements or other forms of collective action. Zersetzung tactics have a long history of use by various states and non-state actors and have been documented in a range of contexts, including political repression, espionage, and counter-terrorism. In recent years, the increasing use of A.I. and other

technologies has enabled the automation and scaling of Zersetzung tactics, leading to concerns about the potential impacts on psychosecurity and social cohesion (Averkin et al., 2019).

The use of A.I. to conduct Automated Zersetzung attacks can have serious consequences for the targeted individuals, including negative impacts on mental health, social connections, and reputation. It can also undermine trust and cohesion within society, as it may lead to feelings of fear and vulnerability among individuals who may feel that they are at risk of being targeted. In the context of security and defense, the use of A.I. for Automated Zersetzung attacks can be particularly concerning, as it allows for the use of psychological warfare tactics in a plausibly deniable manner. This means that it can be difficult to trace the source of the attacks and to hold those responsible accountable, which can further undermine trust and stability within society.

4. Recommendations

There is a need for increased public awareness of the potential impacts of artificial intelligence (A.I.) on society, as well as guidance for governments and regulatory bodies, professional organizations, and individuals working in fields related to A.I., such as clinicians and researchers. One challenge is that A.I. is a rapidly evolving field, and it can be difficult for the public and even for experts to keep up with the latest developments and their potential impacts. And history informs us that new technologies are often rushed to market and deployed before all of the real-world risks are known. This can make it difficult for governments and regulatory bodies to develop appropriate policies and regulations to address potential risks and impacts. Professional healthcare organizations, such as the American Medical Association (AMA) and the American Psychological Association (APA), can play a role in providing guidance and resources to their members on how to responsibly use and interact with A.I. (Luxton, 2014b). This can include guidance on ethical considerations, best practices, and potential risks and impacts. Training is also an essential component in helping individuals working in fields related to A.I. to understand the potential implications of A.I. on society, as well as to develop the skills and knowledge to use A.I. responsibly. This can include training on ethical considerations, as well as training on technical skills related to A.I.

Public health research on the psychological impacts of artificial intelligence (A.I.) can help to inform our understanding of the potential risks and benefits of this technology and can inform the development of standards, certifications, and governance frameworks. There are several professional organizations that have developed standards and certifications related to A.I., such as the Institute of Electrical and Electronics Engineers (IEEE). These standards and certifications can help to ensure that A.I. is developed and used in a responsible and ethical manner and can help to build trust in this technology and its implementation. In addition to standards and certifications, effective governance frameworks for A.I. can help to build trust and confidence in this technology. One key element of effective governance is the inclusion of stakeholders from diverse backgrounds and perspectives, such as the community itself. This can help to ensure that the interests and concerns of all stakeholders are considered and can help to build trust in the governance process.

As we noted earlier in this paper, certain vulnerable groups and populations may be especially at risk for harm by disruptive A.I. Isolated and alienated youth, persons who've not had an opportunity to receive education about A.I. technology, persons who do not have access to these technologies, and persons with certain mental health conditions may be especially vulnerable to harm. It is, therefore, essential that the threats to these populations are adequately considered and research on this topic is supported. The

involvement of underrepresented persons and the most vulnerable in decision-making about developing and deploying A.I. technologies has been recommended (Luxton, 2020).

The increasing use of A.I. in our daily lives, primarily when A.I. mediates or supplants human relationships, is especially important, given how this may lead to changes in public morality and social norms. For example, the use of A.I. companions or friends may become more common, leading to situations where people may have dinner or engage in other social activities "alone" with A.I. entities. The development of these types of relationships with A.I. entities may have significant impacts on individuals' social and emotional well-being, as it may lead to a lack of connection and fulfillment in real-life relationships. It is also possible that these relationships could have an influence on people's beliefs and values, as it is often said that we become like the people with whom we spend significant time.

One of the most challenging threats to individuals, nation-states, and society we've discussed here is Automated Zersetzung attacks. The use of A.I. in fifth-generation plausibly deniable warfare can enable a range of tactical capabilities, such as the ability to conduct complex and coordinated military operations without the need for large, visible military forces. It can also enable the use of covert or indirect means to achieve strategic objectives, such as through the use of drones or cyber operations. One concern with the use of A.I. in fifth-generation plausibly deniable warfare is the potential for it to be used in ways that are difficult to trace or attribute to a specific state or actor. This can create uncertainty and instability within the international system, as it may be challenging to know the motivations and intentions of those responsible for such attacks. It can also make it difficult to hold those responsible accountable, undermining the rule of law and the international order. Investments in systems that can detect and thwart these attacks are needed.

Moving forward, intended or unintended manipulation of human behavior with A.I. will increasingly become a public health concern as recognition has become easier (Luxton, 2022). While behavioral influence and modification, such as with the use of virtual coaches and companions, may be adaptive and oriented to improving health and well-being, this same technology can be used to influence people in ways that cause harm. As we noted earlier, the use of social media bots to persuade and manipulate the public is already evident. The revelations from the release of the "Twitter Files" have undoubtedly damaged trust in governmental institutions that have a duty to protect the health and safety of the public. And while some persons may be most vulnerable and harmed by maladaptive or misuse of A.I. technologies, all of society can suffer from harm when A.I. is used by the powerful to achieve ends that are not in the public's best interest. Transparency, trust, and freedom from psychological manipulation are paramount in a healthy society, and we must strive to assure public trust in how A.I. technologies are used.

6. Final Considerations

There is a range of potential risks and opportunities for the future of artificial intelligence (A.I.) and society. One opportunity is that A.I. is used to enhance human welfare by improving healthcare, education, and other essential services. For example, A.I. could be used to analyze vast amounts of data to identify patterns and trends that professionals could use to improve public health or to develop personalized learning experiences that help individuals achieve their full potential. Another opportunity is that A.I. can be used to address global challenges, such as climate change, poverty, and inequality. For example, A.I. could analyze data and develop solutions to help mitigate climate change's impacts or identify and address the root causes of poverty and inequality. On the other

hand, a significant risk is that A.I. is used to harm or exploit individuals or society, either intentionally or unintentionally.

Based on historical events and analysis from an STS frame, it is likely that the reality of A.I. and society will be a blend of positive and negative effects on human well-being. To mitigate risks to public health, it is essential that a Geneva Conventions-style agreement prohibits the usage of demoralization tools against civilians as a crime against humanity. It is also very important that strong investment is made in improving the transparency and auditability of models, along with in-built encryption techniques such as models partly trained on-device, which better respect user privacy and security.

References

Averkin, A., Bazarkina, D., Pantserev, K., & Pashentsev, E. (2019). Artificial Intelligence in the Context of Psychological Security: Theoretical and Practical Implications. Proceedings of the 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019). https://www.atlantis-press.com/proceedings/eusflat-19/125914786

Ayers, J. (2009). Coping with speech anxiety: The power of positive thinking. *Communication Education. 37*(4) 289-296. https://doi.org/10.1080/03634528809378730

Barrett, D. (2010). Supernormal Stimuli: How Primal Urges Overran Their Evolutionary Purpose. W. W. Norton & Company.

Bénabou, R. (December 2003). Inequality, Technology, and the Social Contract. Woodrow Wilson Economics Discussion Paper No. 226. https://doi.org/10.2139/ssrn.525043

Braghieri, L. Levy, R. & Makarin, A. (July 28, 2022). Social Media and Mental Health. Available at SSRN: https://ssrn.com/abstract=3919760 or http://dx.doi.org/10.2139/ssrn.3919760

Dean, J. (2019). The Deep Learning Revolution and Its Implications for Computer Architecture and Chip Design. https://arxiv.org/ftp/arxiv/papers/1911/1911.05289.pdf

Dennis, M. & LaPorte, N. (2011). "The Stasi and Operational Subversion". State and Minorities in Communist East Germany. Berghahn Books.

Due, B. L. (2015). The social construction of a Glasshole: Google Glass and multiactivity in social interaction. *PsychNology Journal., 13*(2-3), 149-178. http://www.psychnology.org/File/PNJ13%282-3%29/PSYCHNOLOGY_JOURNAL_13_2_DUE.pdf

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K.,… Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational A.I. for research, practice and policy, *International Journal of Information Management, 71*. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Feldstein, S. (April 18, 2019). Artificial Intelligence and Digital Repression: Global Challenges to Governance. http://dx.doi.org/10.2139/ssrn.3374575

Hargittai, E. (2011). The Digital Reproduction of Inequality. In D. Grusky & S. Szelenyi (Eds). The Inequality Reader: Contemporary and Foundational Readings in Race, Class, and Gender (2nd ed.). Routledge. https://doi.org/10.4324/9780429494468-69

Hilty, D. M., Armstrong, C. M., Luxton, D. D., Gentry, M., & Krupinski, E. (2021). A Scoping Review of Sensors, Wearables, and Remote Monitoring For Behavioral Health: Uses, Outcomes, Clinical Competencies, and Research Directions. Journal of Technology in Behavioral Science. 1-36. https://doi.org/10.1007/s41347-021-00199-2

Kaminski, M. E., & Witnov, S. (2015). The Conforming Effect: First Amendment Implications of Surveillance, Beyond Chilling Speech. *University of Richmond Law Review, 49*, 465-518. https://lawreview.richmond.edu/files/2015/01/Kaminski-492.pdf

Krøvel, R. & Thowsen M. (eds). (2019). Making Transparency Possible: An Interdisciplinary Dialogue. Cappelen Damm Akademisk/NOASP. https://doi.org/10.23865/noasp.64.ch16

Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015). Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015: 160-1612. https://doi.org/10.1145/2702123.2702548

Lepore, S. J., Ragan, J. D., & Jones, S. (2000). Talking facilitates cognitive–emotional processes of adaptation to an acute stressor. *Journal of Personality and Social Psychology, 78*(3), 499–508. https://doi.org/10.1037/0022-3514.78.3.499

Litz, B. T. (2007). Research on the Impact of Military Trauma: Current Status and Future Directions. *Military Psychology, 19*(3), 217-238. https://doi.org/10.1080/08995600701386358

Luxton, D. D. (2014a). Artificial Intelligence in Psychological Practice: Current and Future Applications and Implications. *Professional Psychology: Research & Practice. 45*(5), 332-339. https://doi.org/10.1037/a0034559

Luxton, D. D. (2014b). Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine, 62*(1), 1-10. https://doi.org/10.1016/j.artmed.2014.06.004

Luxton, D. D. (2020). Ethical Challenges of Conversational Agents in Global Public Health. *Bulletin of the World Health Organization, 98*(4),285-287. doi: http://dx.doi.org/10.2471/BLT.19.237636

Luxton, D. D. (2022). Deepfake AI and Mass Virtual Reality Are a Public Health Risk: https://daviddluxton.substack.com/p/deepfake-ai-and-mass-virtual-reality

Luxton, D. D. June, J. D. & Fairall, J. M. (2012). Social Media and Suicide: A Public Health Perspective. *American Journal of Public Health, 102*(2), 195-200. doi: 10.2105/AJPH.2011.300608

Luxton, D. D. & Poulin, C. (2020). Advancing Public Health in the Age of Big Data: Methods, Ethics, and Recommendations. In L. Goldschmidt & R. M. Relova (Eds.). Patient-Centered Healthcare Technology: The Way to Better Health. IET. https://doi.org/10.1049/PBHE017E_ch3

Maitra, I. & McGowan M. K. (Eds.) (2012). Speech and Harm: Controversies Over Free Speech, Oxford Academic. https://doi.org/10.1093/acprof:oso/9780199236282.001.0001

Massazza, A., Kienzler, H., Al-Mitwalli, S., Tamimi, N. & Giacaman, R. (2022). The association between uncertainty and mental health: a scoping review of the quantitative literature, *Journal of Mental Health, 32*(2), 480–491. https://doi.org/10.1080/09638237.2021.2022620

Manheim, K. & Kaplan, L. (2018). Artificial Intelligence: Risks to Privacy and Democracy. *Yale Journal of Law and Technology, 21*, 106.

Matthews, J. (2021). Keynote: Surveillance, Power and Accountable A.I. Systems: Can We Craft an A.I. Future that Works for Everyone?, 2021 Eighth International Conference on eDemocracy & eGovernment (ICEDEG), Quito, Ecuador, 2021, pp. 5-5. https://doi.org/10.1109/icedeg52154.2021.9530871

McKay, C. (2020). Predicting risk in criminal procedure: actuarial tools, algorithms, A.I. and judicial decision-making,

Current Issues in Criminal Justice, 32(1), 22-39. https://doi.org/10.1080/10345329.2019.1658694

Mirza, M. U., Richter, A., van Nes, E. H, & Scheffer, M. (2019). Technology driven inequality leads to poverty and resource depletion. *Ecological Economics, 160*, 215-226. https://doi.org/10.1016/j.ecolecon.2019.02.015

Niles, A. N., Craske, M. G., Lieberman, M. D. & Hur. C. (2015). Affect labeling enhances exposure effectiveness for public speaking anxiety. *Behaviour Research and Therapy, 68*, 27-36. https://doi.org/10.1016/j.brat.2015.03.004

Nonan, M. (2013). I, Glasshole: My Year With Google Glass. *Wired*. https://www.wired.com/2013/12/glasshole

O'Connor A. J. & Jahan F. (2014). Under Surveillance and Overwrought: American Muslims' Emotional and Behavioral Responses to Government Surveillance. *Journal of Muslim Mental Health. 8*(1), 95-106. https://doi.org/10.3998/JMMH.10381607.0008.106

Parekh, L. B. (2017). Limits of Free Speech. *Philosophia 45*, 931–935. https://doi.org/10.1007/s11406-016-9752-5

Peachey, K. (March, 2022). Post Office scandal: What the Horizon saga is all about. *BBC News*. https://www.bbc.co.uk/news/business-56718036

Roberts, J. & Shaw, K. L. (2022). Managers and the Management of Organizations. NBER Working Paper Series. National Bureau of Economic Research. https://www.nber.org/papers/w30730

Sardarizadeh, S. & Schraer, R. (2022). BBC News. Twitter Files spark debate about 'blacklisting'. https://www.bbc.com/news/technology-63963779

Schou Andreassen, C., Billieux, J., Griffiths, M. D., Kuss, D. J., Demetrovics, Z., Mazzoni, E., & Pallesen, S. (2016). The relationship between addictive use of social media and video games and symptoms of psychiatric disorders: a large-scale cross-sectional study. *Psychology of Addictive Behaviors, 30*(2), 252. https://doi.org/10.1037/adb0000160

Schuster, D., (2014, July 14). The revolt against Google 'Glassholes'. *New York Post*. https://nypost.com/2014/07/14/is-google-glass-cool-or-just-plain-creepy/

Sedgwick, R., Epstein, S., Dutta, R., & Ougrin, D. (2019). Social media, internet use and suicide attempts in adolescents. *Current Opinion in Psychiatry 32*(6), 534-541. doi: 10.1097/YCO.0000000000000547

Starr, A., Fernandez, L.A., Amster, R., Wood, L. J., & Caro, M. J. (2008). The Impacts of State Surveillance on Political Assembly and Association: A Socio-Legal Analysis. *Qualitative Sociology, 31*, 251–270. https://link.springer.com/article/10.1007/s11133-008-9107-z

The White House (2022). The Impact Of Artificial Intelligence on The Future of Workforces in the European Union and the United States of America. https://www.whitehouse.gov/wp-content/uploads/2022/12/TTC-EC-CEA-AI-Report-12052022-1.pdf

Teo, A. R., & Gaw, A. C. (2010). Hikikomori, a Japanese culture-bound syndrome of social withdrawal?: A proposal for DSM-5. *The Journal of Nervous and Mental Disease*, *198*(6), 444–449. https://doi.org/10.1097/NMD.0b013e3181e086b1

Ullah, M. (2016). Impact of Drone Attacks Anxiety on Students at Secondary Level in North Waziristan Agency. *Public Policy and Administration Research*, *6*, 1-7. https://www.iiste.org/Journals/index.php/PPAR/article/view/28550

U.S. House of Representatives (March 2023). Hearing on the Weaponization of the Federal Government on the Twitter Files. Available at: https://judiciary.house.gov/committee-activity/hearings/hearing-weaponization-federal-government-twitter-files

Van Dijk, J. A. (2006). Digital divide research, achievements and shortcomings. *Poetics, 34*(4-5), 221-235.

Verma, P. (2022). Humans vs. Robots: The Battle Reaches a 'Turning Point'. The Washington Post. https://www.washingtonpost.com/technology/2022/12/10/warehouse-robots-amazon-sparrow

Wee, A. & Findlay, M. J. (September 14, 2020). A.I. and Data Use: Surveillance Technology and Community Disquiet in the Age of COVID-19. SMU Centre for A.I. & Data Governance Research Paper No. 2020/10. http://dx.doi.org/10.2139/ssrn.3715993

Wojcieszak, M. (2010). 'Don't talk to me': effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *New Media & Society*, *12*(4), 637-55. https://doi.org/10.1177/1461444809342775

Stanford
Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Science, Delusion, and Existential Risk

Andrew Nepomuceno [1*]

Citation: Nepomuceno, Andrew. Science, Delusion, and Existential Risk. *Proceedings of the Stanford Existential Risks Conference* 2023, 75-90. https://doi.org/10.25740/hc216sm8573

Academic Editor: Steve Luby, Trond Undheim, Dan Zimmer

Funding: N/A

Conflict of Interest Statement: N/A

Informed Consent Statement: N/A

Acknowledgments: Author thanks colleagues at Stanford Epidemiology and at Magic for insightful critiques.

Author Contributions: N/A

Abstract: What if delusion, fixed belief not amenable to change in light of conflicting evidence" is a foundation for multiple existential risks—including all termed "major concerns" in the call for papers? Is relative inattention to delusion a blind spot with which we limit effective framing and understanding of existential risks? Phenomena from emergent and resurgent infectious diseases to climate destabilization and mass extinction worsen when people hold delusional ideas about them. If delusion is an underpinning for multiple existential threats, it's simultaneously a critically important challenge and a unique opportunity. We can diminish existential risks we generate from delusion by shedding it to see ourselves and surroundings more accurately. With its mandates to conform ideas to evidence, identify pattern where it exists, eschew claims to it where it's lacking, and reject even the idea of immutable "truth," sciencing can be effective prevention and treatment for delusion, and powerful means to lessen multiple existential threats." Given this promise, promoting sciencing may warrant substantially greater attention from the community addressing existential risks.

Keywords: delusion, science, climate, infectious disease, existential risk

[1] Stanford Department of Epidemiology and Population Health, Alway Building 300 Pasteur Drive Stanford, California 94305 Mail code: 5405; andrewn1@stanford.edu.

* Correspondence: andrewn1@stanford.edu

1. Introduction

Throughout history, various cultures and societies have contemplated the possibility of catastrophic ends. The concept of apocalypse has been an element of religious traditions for thousands of years. In the past century humans have become aware of existential threats—*e.g.*, super-volcanoes (Costa *et al*, 2014), solar flares (Kaufmann *et al*., 2004), asteroid strikes (Shukolyukov and Lugmair, 1998)—of which we were previously ignorant. In addition, we've by our own hand intensified long-standing risks, such as climate disruption (United Nations) and plague (Fauci, 2005), and created new ones, such as artificial intelligence (Turchin & Denkenberger, 2020), engineered pathogens (Wickiser *et al*., 2020), and nuclear weapons. Today a growing number of us are asking: (1) How may we better assess the likelihood of global cataclysm? and (2) What can we do to prevent it, or if it's underway, stem its damage and recover from it (Bostrom, 2002)?
These questions lie within the domain of the science of human ecology, since both of them entail study of interaction between humans and the environment. Human ecology is an interdisciplinary field (Hens, 1998; Marten, 2001; Bates *et al*., 2022) comprising aspects of other scientific disciplines in which researchers have addressed x-risks (Koch *et al*., 2013; Grunwald *et al*., 2017; Lin, 2019; Pimentel *et al*., 2007). In this paper, I: (1) present a human ecological analysis; (2) use it to identify and describe an existential threat that is largely ignored, already exacting a toll, underpins a panoply of anthropogenic threats that comprise the lion's share of existential risk ($\sum$ [probability x adversity]), and compromises our ability to respond to other x-risks; and (3) suggest research to address this foundational threat.

2. Human Ecology Analysis

In its simplest form, a human ecology framework entails humans, environment, and interactions between these (Haeckel, 1866). In this framework, "environment" denotes everything with which humans interact, including for each of us: other people, other life, other elements of the natural world, and the artifact we've shaped from it; "interaction" means transfer of matterenergy (*e.g.*, sound, heat, light, oxygen, food, bodily wastes).

Human interaction with the environment is predicated upon information—genetic (Dawkins, 1982), epigenetic (Powledge, 2011), and experiential (Whiten *et al*., 1999)—embodied in our physical structure. While humankind has accumulated vast quantities of information extra-somatically (*e.g.*, books, digital media), individuals internalize different elements of this to varying degrees and use only that we've absorbed, however temporarily, to interact with the environment.

Concern about events that may result in human extinction or in prolonged and severe degradation of the quality of human existence springs from desires to sustain the human species and aspects of our current way of life deemed critical to our well-being. Doing these things entails maintaining a match between the information on which we act and the qualities of the environment. Species that have maintained this match by evolving our information through natural selection remain extant. Those that have failed to maintain the match, including at the time of this writing growing numbers lost each day (Ceballos *et al*., 2017), have become extinct.

Humans threaten our own match with the environment. As we confront the challenge to evolve our information in response to changes in our surrounds, we labor under two handicaps of our own making. First, being so many and so powerful, we ourselves are wreaking environmental changes—already rapid beyond precedent during our tenure and still accelerating—to which we're obligated to conform (Stokstad, 2019). Second, because a growing number of us are increasingly immersed in interaction with each other and human artifact, and separated from the rest of nature (Kesebir & Kesebir, 2017), we're insulated from, less able to discern, and consequently hampered in adapting to critically important environmental changes (*e.g.*, decreasing biodiversity, climate destabilization, ocean acidification) (Bekoff & Bexell, 2010).

Because human generations are relatively long, on the order of two decades, and because our capacity for precise, deliberate selection of genetic information is nascent and remains even in its currently limited, primitive form largely untested, evolution of genetic information is too slow to meet our adaptive challenge (WHO, 2021). We've persisted because we, like many other life forms, have evolved ability to augment genetic information with learned.

Culture, information we transmit behaviorally (Bonner, 1980), is uniquely important learning, especially in this era of near lightspeed global communication, massive extrasomatic information processing and storage, and pervasive networks connecting billions of us. Today unprecedentedly extensive culture can be more rapidly than ever before transmitted from any of a large and growing majority of individuals to nearly all of the human population.

Culture, however, is a two-edged sword, its adaptive elements potentially salvational (UNESCO, 2021), its maladaptive, potentially devastating (Southwell *et al.*, 2017). As we work to discern and promulgate the former and promote their widespread uptake, and to suppress the latter and encourage their abandonment, we've made our task more difficult by growing culture exponentially. Though once we may have adapted primarily by distinguishing accurate information from inaccurate, today we also need distinguish accurate information necessary to adaptation from a vast amount less critical. Furthermore, by granting broader access to channels of distribution and empowering individuals to pour more into them, we may have reduced the ratio of adaptive to maladaptive culture (Zollo, 2019).

Both edges of the cultural sword are sharper than ever before. Though advocates for free exchange of information celebrate movement towards anyone's being able to transmit anything to everyone, many agree that consequences of such movement are already problematic, as evidenced by onslaughts of disinformation and misinformation (Johnson *et al.*, 2020) and by actions based on these (*e.g.*, storming of the US Capitol Building, January 6, 2021).

When we rely more heavily upon cultural transmission of representations of experience, we lose feedback inherent to direct experience of phenomena. (*e.g.*, Hearing, "Fire burns!" is different from feeling flames.) In part as a result, we mistake partial knowing sufficient to describe selected aspects of self and world for operative, more complete knowing necessary to action. We recognize this by requiring direct experiential learning in everything from building trades apprenticeships to science class labs. Perils of shallow learning are evident (Pfeffer & Sutton, 2000) in countless

adverse events arising from failure to act on what we "know" (*e.g.*, drunk driving fatalities, loss of data without backup).

We're faced with necessity to rely on culture more heavily and to curate it more carefully. Only with culture can we adapt our information rapidly, broadly, and deeply enough to match environmental changes. Only by adept curation can we evolve culture to meet the challenge of adaptation. Such curation may entail reexamining foundational information. Among the most critical of this are ideas about what we want, how we can get it, and how we can accurately formulate these (Miles, 2015).

At the time of this writing in 2023, eight billion humans awaken each day, begin to ask more or less continuously, consciously and unconsciously, "What do I want?" and "How can I get it?" and act accordingly. With our growing numbers and waxing power, the world is increasingly a reflection of our actions and of ideas about what we want and how to get it in which we ground them. ("Climate change and human behaviour," 2022).

Ideas about what we want and how to get it—ideas about value or good—are future-oriented. They're based upon predictions that if we get what we want, we'll feel satisfaction, and that if we do what we think necessary, we'll succeed.

The polycrises (Homer-Dixon *et al.*, 2021) of our era, including many x-risks, are outgrowths of failure to predict accurately consequences of our behavior. In them we see reflected illusory knowledge of value. Developers of CRISPR, nuclear weapons, and artificial intelligence, think benefits of these technologies outweigh their drawbacks. How different will they prove from Thomas Midgley, Jr., inventor of stratospheric-ozone-depleting chlorofluorocarbons (Midgley, 1930), and human-intelligence-lowering tetraethylleaded gasoline (Boyd, 1953), or from the billions who've induced—until recently mostly unwittingly—climate catastrophe by burning fossil hydrocarbons?

We're overconfident in thinking we know what we want (Gilbert et al., 2009). Despite past and present—in many cases large and growing—admonitions in the form of unanticipated adverse impacts of our behavior, we persist in acting with confidence that we'll get what we want and want what we get. Though most of us aim to learn from errors as we become aware of them, few step back and ask a more fundamental question directed to the root of our overconfidence: "How do we know?" (Kuhn et al., 2000).

When we do this, we see that what we think we know about value is largely a result of chance—genes we inherited without choosing, and experience determined largely by circumstances of birth (Achen, 2002; Pew Research Center, 2020). While this information has sufficed for adaptation to this moment, its adequacy in a rapidly changing environment is far from assured and becomes less so with each acceleration of change.

While, as noted previously, we've limited ability to alter genetic information, we can become more aware of it, and we can accentuate or attenuate behavior informed by at least some of it (Henwood *et al.*, 2015). Differences between our reproductive behavior, which is deeply genetically informed, and that of the primate relatives with whom we share nearly all our genetic information (Kappeler, 2012) are compelling evidence.

Cultural information about value, though sometimes firmly ingrained—especially when we receive it early in life (*e.g.*, religion) and act repeatedly and publicly (*e.g.*, participating in group observances) to become identified with it—often can be more malleable than genes. We observe this in the rapidity and extensiveness of evolution of culturally-informed values (*e.g.*, music, dress, language, adoption of technology) of people around the world. To the extent that we rely upon culture to inform ideas about what we want and how to get it, we've opportunity to accelerate their evolution and with it, change to myriad behaviors we root in them.

Living things have been gaining capacity to learn for billions of years (Mitchell, *et al*., 2009). In recent centuries humans have made accelerating progress developing a set of best practices for learning. We term them "science" (Einstein, 1936; Walker, 1963). They include creative imagination, abiding skepticism, careful and extensive observation, sound reasoning, validating experimentation, and collegial collaboration with emphasis upon inviting others to confirm or contradict our findings (AAAS, 1990). Behaviors we denote by "science" together comprise the sole demonstrated means for predicting with success greater than we can achieve by chance.

Since successful adaptation requires that we interact with the environment to achieve specific outcomes, sciencing—with its unique capacity to improve prediction of consequences of our interactions—is key to human survival and thriving (Walker, 1963: 177-179). Only to the extent that we've predicted accurately are ideas about, and actions to realize the ends and means of our lives likely to result in satisfaction greater than we can achieve by random action (Schrom, 2004).

A kernel element of sciencing is admission to ignorance of any final "truth" and that knowledge is provisional (House, 1992; Griffiths *et al*., 1995; Zermelo, 2013; Understanding Science, 2022). We science, aware that just as a map is different from the territory we represent with it (Korzybski, 1958), ideas are different from what we represent with them. In the case of ideas about value—about what we want and how to get it—we science as we do with all other ideas, to conform them more nearly to what we represent with them and inform more accurate prediction. We science to be less wrong rather than to be right. By holding open to challenge whatever information we carry, we enable learning to change both it and how we act upon it, and increase our capacity to keep existential and other threats at bay.

With human ecology we discern our adaptive challenge to be maintaining a match between our information and the characteristics of our environment. We note that in the current rapidly changing environment we rely heavily upon culture to maintain this match. We identify science, behavior—including abjuration of claims to certainty—by which we improve prediction, as key to evolving adaptive culture, including that necessary to forestall existential threats yet to occur and to abate those already underway.

3. Delusion

Whenever we hold ideas beyond question, we inhibit sciencing. We've many terms for refusal to question behavior: dogmatism, motivated reasoning, delusion, ideological rigidity, close-mindedness, and more. One of these, delusion, stands apart as a diagnosis for mental illness.

Examining criteria for this diagnosis and considering how we might refine them to more accurately reflect incidence of the psychopathology we characterize with it, we gain insight to the pervasiveness of holding ideas beyond question, and to the role of this behavior in destruction by which we've already degraded human existence and by which, if we continue, we may hasten its end. We also shed light on inducements to delusion and on obstacles to recovery.

Mental health professionals have for centuries (de Sauvages, 1772) asserted that mental illnesses are diagnosable and treatable medical conditions that can be categorized on the basis of well-defined criteria. In 1952 they created the first *Diagnostic Statistical Manual* (American Psychiatric Association, 1952) to improve upon prior classifications by establishing clear, evidenced-based standards for identifying specific psychopathologies. Researchers and clinicians rely heavily upon these diagnostic criteria and continue to evolve them as they accumulate evidence for doing so.

In the most recent version of the *Diagnostic Statistical Manual (DSM-5)*, they write: "Delusions are fixed beliefs that are not amenable to change in light of conflicting evidence." *(American Psychiatric Association, 2013,* p.87*).* Other academic works (Garety, 1985; Jaspers, 1997) highlight the fixed nature of beliefs as a key element of delusion.

Though we might expect delusion to be frequently diagnosed, given how many people cling to fixed beliefs in the face of overwhelming contravening evidence, it's attributed to less than 0.2% of the population—fewer than one person in 500 *(American Psychiatric Association, 2013)*. To understand why, and to see more clearly the extraordinary threat we pose with delusion, we return to the *DSM-5*. There we discover a further diagnostic criterion: "The belief is not ordinarily accepted by other members of the person's culture or subculture (*i.e.*, it is not an article of religious faith)." (*American Psychiatric Association, 2013,* p.819*)*.

This is a remarkable departure from norms of modern medicine, analogous to declaring that a child with measles is cured if all the other children in the neighborhood contract the disease. With it, authors of *DSM-5* qualify popular opinion to trump evidence, and convert the criteria for diagnosing delusion from scientific to political.

Belief held beyond question despite evidence of its falsity is just as real when shared as when individual. An entire population or substantial fraction of one may meet this criterion, just as it may be diabetic, cancerous, or infected with plague. We've examples of whole villages in medieval Europe delusional with ergotism as a result of eating fungus-contaminated rye. In *Extraordinary Popular Delusions and the Madness of Crowds* (Mackay, 1841), Charles Mackay enumerates and analyzes several mass delusions. Examples from U.S. history include the Salem Witch Trials of the mid-17th century (Blumberg, 2007) and the Red Scare of the mid-20th (Murray, 1955). Today we've widespread delusion with respect to everything from vaccines, to climate, to corporate profitability (TEEB, 2010, Muller *et al*., 2011).

As noted earlier, when we embody ideas in brain structure, we're like any mapmaker or modeler, representing incompletely and imperfectly (Korzybski, 1958). If we imagine neurons and synapses embodying some idea to be complete and final truth, we're delusional regardless of how otherwise accurate those ideas may be. While we can make a case for concentrating attention on delusions that are "more wrong" based

upon perceived degree of inaccuracy, or "more dangerous" based upon perceived potential for harm, we risk missing these if we grant exemption from challenge.

From an ecological perspective, health is the match between an organism and its environment. A healthy organism sustains its functions by dint of that match. Disease and injury entail departures from that match, and compromise interactions necessary to full function. If we reduce the criteria for a psychopathology of delusion to "ideas held beyond question," we're consistent with human ecology because such ideas, reduce our capacity to adapt our information to match our environment.

Delusion is ubiquitous. Each of us exhibits it whenever we hold ideas beyond question. Only to the extent that we question can we science. To the degree that we inform our actions with delusion, we reduce our likelihood of achieving results we intend, and we set in motion unforeseen consequences, including many of today's least tractable problems (*e.g.*, x-risks). By depoliticizing delusion we may also destigmatize it. Rather than preserve it as a tool for legitimizing popular dogma and denigrating less common, we make it impetus for self-scrutiny and motivation to better sciencing by all. With this mindset, we might treat ourselves and others with more empathy when any of us appears stuck on one or another belief. As prominent literature on persuasion shows, non-judgmental attitudes are useful when reconsidering views or assisting others (Kalla & Broockman, 2019) in doing so.

Degrees of delusion from mild to severe and from narrow to broad are analogous to degrees of other mental and physical illnesses. An individual's skin lesion may be trivial or life-threatening. So may be a delusion. A person may harbor diverse infectious organisms or degenerative conditions. So may each of us hold a variety of delusions.

Just as we mount an immune response or alter behavior to fight and defeat infection and limit or reverse degeneration, so may we resist or abandon delusion. Ability to do these reflects both personal attributes and social context, and is inevitably constrained by the very nature of delusion, with which we disable the faculty necessary to shed it.

Humans have harbored delusions (*e.g.*, belief in supernatural phenomena) for many millennia, likely for as long as we've existed as a species. While costs (*e.g.*, torture and killing) of some past delusions (*e.g.*, belief that with these we served some supernatural entity) have been massive, we've evidence and reason to conclude that costs (*e.g.*, ocean acidification) of current delusions (*e.g.*, thinking that fossil fuel burning is beneficial) are much greater.

Eight billion people wielding technologies orders of magnitude more powerful than any prior to effect global change more rapid, more extensive, and more destructive than any previous have transformed delusion from narrow to existential threat. Where once some suffered hunger for unquestioned belief that one or another was able to summon rain that crops might flourish, today all plunge forward into cataclysm for unquestioned belief that we lack wherewithal to survive and thrive if we work substantially less to further narrow interests and substantially more to further common good.

The price we pay for failing to identify and shed delusion is immediate and future forfeiture of satisfaction and endurance of pain. In light of the current status and trends

of major ecosystem elements and processes (Meadows *et al*., 1972; Turner, 2008; Bardi, 2011; Herrington, 2021), delusion constitutes existential risk.

Often when we feel strongly about an idea, we're compromised in our ability to question it. Typically, we're even more hindered when we can gather around us others with similar deep feelings. Difficulty of questioning unrestricted gun ownership in much of Montana, authority in the military, or benefits of increasing GDP on Wall Street are illustrative.

The ideas in these examples, like many others deeply rooted and widely shared, are elements of worldview, an overarching framework of answers (Aerts *et al*., 2007) to fundamental questions about being and purpose whose basic components we typically acquire when too young to critically evaluate them. As we accumulate experiential information, we integrate these with, and layer it onto worldview, making its foundational elements more challenging to discern and to alter.

We confront the evolutionary task of simultaneously adapting to: (1) a social order in which shared ideas held beyond question are bedrock, and individual adherence or resistance to them may result in immediate rewards and punishments, and (2) a natural world of immutable law perfectly enforced which delivers many reinforcements and penalties collectively in the distant future.

In the early twenty-first century we've growing recognition that imperatives of human and natural law may be incompatible, even opposing. For example, we may lionize luxury automobiles as status symbols even as we acknowledge their existence to be an insult to the poor and both result and cause of environmental degradation by which we impoverish a common future.

In our discomfort we may retreat into fixed beliefs such as, "I can't make any difference," "I can't change," and "Everything will work out," that we defend, rather than hold open to challenge. Regardless of their merit, placing them beyond question is delusional, perhaps existentially so. Our ability to forestall x-risks may rest much more on how many of us think we can than upon any scheme hatched by a small number of us. Because accuracy of ideas about effectiveness of, and necessity for individual and collective action may depend to a large extent on how many share and act on them, testing impacts of delusions of self-negation and complacency is difficult.

In a society where individualism is more celebrated than individuality, both leadership in testing assumptions and rationalizations, as well as ability to rally in support of such leadership are unusual. By making them commonplace we may be able to shed delusion and enhance adaptivity.

The pandemic of delusion differs from many x-risks because it's underway—probability 100%—and has already yielded degradations of the human condition severe, extensive, and persistent on the scale of millions of years (Song & Dunhill, 2018)—the amount of time that has been necessary to recover biodiversity lost in prior mass extinctions comparable to the one now underway. It's unique among x-risks also in the extent that it amplifies the remainder that are anthropogenic, and adversely affects our ability to address others.

Many who contemplate action to counter x-risks concentrate on technology applied to the world beyond human consciousness (*e.g.*, monitor emergent pathogens; stockpile electrical transformers to replace those lost to space weather) including global interventions without precedent (*e.g.*, block sunlight with stratospheric aerosols; seed oceans with iron). If unquestioned, such heavy reliance on outward-facing, technological solutions generally, and in confronting existential risk specifically, is delusional.

With it we foreclose examination of limits to how well we can secure our future by emphasizing manipulation of the world without and giving little attention to the world within. Fixated on incorrigible belief, we convert the natural world around us to human biomass and artifact at ever-increasing rate, while making proportionately much smaller changes to hearts and minds. Conviction that the next round of technological innovation will secure satisfaction when all prior rounds have brought new threats may be a delusion with exceptionally clear connection to x-risk.

We've paved the path to, and sustained such delusion by tapping ever richer and more concentrated energy sources and using them to achieve ever more spectacular transformations of the rest of nature, thereby reinforcing delusions of limitless ability to substitute new for exhausted resource and to reshape nature, including our physical selves, to our liking. In the past century we've generated global impacts sufficiently detrimental for a substantial and growing fraction of humankind to question whether we can continue to rely for our well-being on technological innovation and escalating rearrangement of the biosphere to sustain a growing population (Feldstein, 2023). Thus may we be shedding delusion.

Yet even in 2023, exhortations abound to increase human population (Woo, 2021) and to grow GDP (Oulton, 2012) both locally and globally. Unquestioning beliefs in the goodness of one or both of these are among the most widespread, and given the abundant evidence of their adverse consequences, among the most pernicious of delusions. With them we increase environmental degradation and social conflict that are major factors in anthropogenic x-risks.

Importantly, both desire to reproduce to maximum potential (Bradshaw & McMahon, 2008) and closely related pursuit of dominance over others and the Earth are to a substantial degree genetically informed. Individuals of nearly all species studied generally behave to reproduce as many offspring as possible. Dominance hierarchies are evident in many social species (Sapolsky, 2005), including all primates. With motivated reasoning by which we justify pre-existing thoughts and feelings even as we pretend to be rationally interpreting evidence, we can camouflage delusion grounded in genes, making it more difficult to detect. Even when we're aware of it, such information can be difficult to suppress, despite great and mounting evidence for adaptive benefit of doing so.

As noted previously with respect to growth of population and GDP, despite obstacles and difficulties mentioned here and many more, people can and do surrender delusions and resist acquiring them. Pope Francis himself has encouraged such changes with his call to reject unquestioning condemnation of gay people (Rocca, 2023). For centuries Buddhists have practiced non-attachment, with which they imply openness to questioning. With the rise of sciencing, we've abandoned countless former delusions and with rejection of each made stronger commitment to persistent questioning. In

these and myriad other actions lies promise that we may be able to abate delusion sufficiently to secure a future worth inhabiting.

Human beings have gathered enormous momentum in ways of living through which we degrade the planet and impoverish the future. We've very effective interlocking and mutually reinforcing systems to protect power and privilege. Each of us confronts the infamous "prisoner's dilemma," (Poundstone, 1993) where individual action for common good may yield either: (1) its intended effect or (2) personal loss benefitting those who decline to cooperate. If shifting from collective defection to cooperation be adaptive, we'll discover this by sciencing.

Delusion is inimical to sciencing and poses an impediment to successful adaptation. Pandemic delusion in an ecosystem where human population is as large and commands as much power as we do today, qualifies in itself as an x-risk already unfolding. Sciencing can be means to reduce the x-risks of delusion and its offshoots.

4. Research

How might we learn to promote sciencing more effectively? Colleagues and I are evaluating three strategies: (1) recharacterize science to emphasize its universality and its uniqueness as a method for improving prediction; (2) cooperate with sub-groups in the general population who self-identify or are identified by others as non-scientists to strengthen their scientist identity; and (3) expand the domain of science to include questions of value.

Recharacterize science.
Albert Einstein wrote in the *Journal of the Franklin Institute*, "The whole of science is nothing more than a refinement of everyday thinking" (Einstein, 1936). In doing so, he invited all of us to join in sciencing more consciously, competently, and consistently. Colleagues and I are designing and testing interventions by which we recharacterize science to emphasize how people science continuously in everyday life to predict consequences of behavior and to choose actions we predict will yield desired results. By drawing attention to individuals' successes and failures and assisting people in becoming aware of ways of thinking by which we generate improved prediction, we inspire and inform better sciencing.

Strengthen scientist identity among "non-scientists."
Abating anthropogenic x-risks may require broad, creative participation in a scientific enterprise with widely shared leadership that taps the full potential of as many people as we can enlist. Characterization of science as how we improve prediction, sketched above, makes clear that the binary scientist/non-scientist (*Cambridge English Dictionary*) reflects in common language our collective failure to hinge such labels proficiency in sciencing, the essence of science, and our reliance instead upon loosely coupled surrogate measures (*e.g.*, formal education).

All of us science with varying degrees of competence in different aspects of our lives. Nobel prize-winners may be miserable spouses, and peasant women without formal schooling may maintain peace in entire villages. By emphasizing the obvious, that we're all scientists when we science, we invite each person to cultivate scientist identity. Encouragement—especially from people currently recognized as scientists—to embrace one's "inner scientist" may be means for those who currently identify as non-

scientists to overcome obstacles that we've emplaced and continue to strengthen by adherence to dichotomous characterization using non-essential criteria.

Expand the domain of science to include questions of value.
Ideas about what we consider good—what we want and how to get it, the ends and means of our lives—are necessarily future-oriented, since we can only fulfill want or act to fulfill it in the future, however proximate. They therefore rest on prediction. Those occasions when we get what we want and feel less satisfaction than we anticipated, or when we do what we think sufficient and fall short reflect failures of prediction. In them we see discrepancy between ideas about value and actual value. To enjoy more available satisfaction and avoid superfluous pain we conform idea to actuality. Science, by definition behavior by which we improve prediction, is how we do this.

Because ideas about value are a basis for much action, and the world is increasingly a reflection of our actions, sciencing to discern value more accurately and realize it more fully is means to increase the likelihood that with our behaviors we'll attenuate rather than exacerbate our problems, including x-risks.

Although professional scientists speak with one voice in affirming that we can use science to discover ever better means, they're equally unanimous in *denying* that we can use science to discern our ends (AAAS, 1990). In this we see reflected historical divisions of power in society rather than basis in fact. We've agreement that we can use science to ascertain the ends of every other life form, including our chimpanzee near relatives with whom we share 98+% of our DNA. Only because we cling to human exceptionalism do we to continue to pretend that we're barred from using it to illuminate our own ends.

Researching how to demonstrate benefits of valuescience and communicate them may be means to motivate large numbers of people to re-examine ideas about value, change them, and alter behavior accordingly. It's potentially a powerful lever to move humankind towards a more adaptive science-based culture better suited to allaying and avoiding x-risks.

5. Conclusion

In this paper I've taken a step back to view x-risks collectively with an eye to common elements that might be means to address them *en masse* and nearer their roots. As an initial step, I outlined a human ecology framework, by definition inclusive of other applicable science.

Using this framework, I identified cultural information as a lynchpin both for generating and for attenuating x-risks. I noted how humans more numerous and powerful than ever before rely heavily upon culture to inform action, and how the world, including numerous x-risks, increasingly reflects actions based upon misjudgment of outcomes. I then discussed how science, in which we commit to abiding skepticism, is means to improve prediction and thereby more effectively attenuate x-risks.

Turning attention to impediments to sciencing, I examined delusion, the essence of which is information held beyond question. I discussed the difficulty of effectively

addressing delusion in a context where its definition has been politicized to exclude much mass delusion, and recommended that we hew more closely to a scientific definition.

I showed how delusion is an unfolding phenomenon by which we've contributed to past and present degradation of the human condition, and by which we continue into collective impoverishment and towards extinction by amplifying and multiplying anthropogenic x-risks like climate degradation, nuclear war, and emergent infectious disease. I also observed that science can be both prophylaxis and treatment for delusion, despite our limited expertise in making it either.

Finally, I've outlined three proposals to test means to promote more conscious, consistent, competent sciencing. I look forward to growing partnership in learning how we can develop these and other ideas to secure a long and comfortable human future.

References

Achen, C. H. (2002). Parental socialization and rational party identification. *Political Behavior*, *24*, 151-170.

Aerts, D., Apostel, L., De Moor, B., Hellemans, S., Maex, E., Van Belle, H., & Van der Veken, J. (2007). World views: From fragmentation to integration.

American Association for the Advancement of Science. (1990). *Science for all Americans*.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Bahm, A. J. (1971). Science is not value-free. *Policy Sciences*, *2*(4), 391-396.

Bardi, U. (2011). The limits to growth revisited. Springer Science & Business Media.

Bates, D. G., Lozny, L. R., & Tucker, J. (2022). Editor's Note on the 50th Anniversary of Human Ecology: An Interdisciplinary Journal. *Human Ecology*, *50*(1), 1-9.

Bekoff, M., & Bexell, S. M. (2010). Ignoring nature: Why we do it, the dire consequences, and the need for a paradigm shift to save animals, habitats, and ourselves. *Human ecology review*, *17*(1), 70-74.

Blumberg, J. (2007). A brief history of the Salem witch trials. Smithsonian.

Bonner, J. T. (1980). The evolution of culture in animals (Vol. 2). Princeton University Press.

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

Boyd, T. (1953). Obituary. Thomas Midgley, Jr. *Journal of the American Chemical Society*, *75*(12), 2791-2795.

Bradshaw, C. J. A., & McMahon, C. R. (2008). Fecundity. In *Encyclopedia of ecology, five-volume set* (pp. 1535-1543). Elsevier Inc..

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, *352*(6282), 220-224.

Cambridge Dictionary. (n.d.). Non-scientist. In *dictionary.cambridge.org dictionary*. Retrieved April 9, 2023, from https://dictionary.cambridge.org/us/dictionary/english/non-scientist

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the national academy of sciences*, *114*(30), E6089-E6096.

Climate change and human behaviour. *Nat Hum Behav* **6**, 1441–1442 (2022). https://doi.org/10.1038/s41562-022-01490-9

Committee on Nomenclature and Statistics of the American Psychiatric Association. (1952). *Diagnostic and Statistical Manual*. American Psychiatric Association.

Costa, A., Smith, V. C., Macedonio, G., & Matthews, N. E. (2014). The magnitude and impact of the Youngest Toba Tuff super-eruption. *Frontiers in Earth Science*, *2*, 16.

de Sauvages, F. B. (1772). *Nosologie méthodique* (Vol. 5).

Dawkins, R. (1982). *The extended phenotype: The long reach of the gene*. Oxford University Press.

DeWitte, S. N. (2014). Mortality risk and survival in the aftermath of the medieval Black Death. *PloS one*, *9*(5), e96513.

Einstein, A. (1936). Physics and reality. *Journal of the Franklin Institute*, *221*(3), 349-382.

Fauci, A.S. (2005). "Emerging and reemerging infectious diseases: the perpetual challenge". Academic Medicine. 80 (12): 1079–85. doi:10.1097/00001888-200512000-00002. PMID 16306276. S2CID 17293745

Feldstein, S. (2023, May 4). Population Decline Will Change the World for the Better. Scientific American. https://www.scientificamerican.com/article/population-decline-will-change-the-world-for-the-better/

Funkhouser, D. (2015, July 25). *Battling 'the Largest Mass Poisoning in History'*. State of the Planet. Retrieved April 9, 2023, from https://news.climate.columbia.edu/2015/07/13/battling-the-largest-mass-poisoning-in-history/

Garety, P. (1985). Delusions: Problems in definition and measurement. *British Journal of Medical Psychology*, *58*(1), 25-34.

Gilbert, D. T., Killingsworth, M. A., Eyre, R. N., & Wilson, T. D. (2009). The surprising power of neighborly advice. *Science*, *323*(5921), 1617-1619.

Gopnik, A., Meltzoff, A. N., & Kuhl, P. K. (1999). *The scientist in the crib: Minds, brains, and how children learn*. William Morrow & Co.

Griffiths, Phillip A., et al. *On being a scientist: Responsible conduct in research*. Second, 1995.

Grunwald, S., Clingensmith, C. M., Gavilan, C. P., Mizuta, K., Wilcox, R. K. K., Pinheiro, É. F., ... & Ross, C. W. (2017). Integrating new perspectives to address global soil security: Ideas from integral ecology. *Global soil security*, 319-329.

Haeckel, E. (1866). Generelle Morphologie der Organismen [General morphology of organisms]. *Berlin: Reimer*.

Henwood, K. S., Chou, S., & Browne, K. D. (2015). A systematic review and meta-analysis on the effectiveness of CBT informed anger management. *Aggression and violent behavior*, *25*, 280-292.

Hens, L. (Ed.). (1998). Research in human ecology: An interdisciplinary overview.

Herrington, G. (2021). Update to limits to growth: Comparing the World3 model with empirical data. *Journal of Industrial Ecology*, *25*(3), 614-626.

Homer-Dixon, T., Renn, O., Rockstrom, J., Donges, J. F., & Janzwood, S. (2021). A call for an international research program on the risk of a global polycrisis. *Available at SSRN 4058592*.

House, E. R. (1992). Response to notes on pragmatism and scientific realism. *Educational Researcher*, *21*(6), 18-19.

Jaspers, K. (1997). *General psychopathology* (Vol. 2). JHU Press.

Johnson, C. R. (2021, June 18). *How the Black Death made life better*. Washington University in St. Louis Department of History. Retrieved April 9, 2023, from https://history.wustl.edu/news/how-black-death-made-life better

Johnson, N. F., Velásquez, N., Restrepo, N. J., Leahy, R., Gabriel, N., El Oud, S., ... & Lupu, Y. (2020). The online competition between pro-and anti-vaccination views. *Nature*, *582*(7811), 230-233.

Kalla, J., & Broockman, D. (2020). Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments. *American Political Science Review, 114*(2), 410-425. doi:10.1017/S0003055419000923

Kappeler, P. M. (2012). Male reproductive strategies: How do behavior, morphology and physiology act in concert to improve a male's reproductive success - and why is there so much variation among outcomes within and between species? *Nature Education Knowledge*, *3*(10), 82.

Kaufmann, P., Raulin, J. P., De Castro, C. G., Levato, H., Gary, D. E., Costa, J. E., ... & Correia, E. (2004). A new solar burst spectral component emitting only in the terahertz range. *The astrophysical journal*, *603*(2), L121.

Kesebir, S., & Kesebir, P. (2017). A growing disconnection from nature is evident in cultural products. *Perspectives on Psychological Science*, *12*(2), 258-269.

Koch, A., McBratney, A., Adams, M., Field, D., Hill, R., Crawford, J., ... & Zimmermann, M. (2013). Soil security: solving the global soil crisis. *Global Policy*, *4*(4), 434-441.

Korzybski, A. (1958). *Science and sanity: An introduction to non-Aristotelian systems and general semantics*. Institute of GS.

Krosnick, J. A. (2018). Questionnaire design. *The Palgrave handbook of survey research*, 439-455.

Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive development*, *15*(3), 309-328.

Lin, H. (2019). The existential threat from cyber-enabled information warfare. *Bulletin of the Atomic Scientists*, *75*(4), 187-196.

Mackay, Charles, 1814-1889. 1841. Extraordinary Popular Delusions and the Madness of Crowds.

Marten, G. G. (2001). *Human ecology: Basic concepts for sustainable development*. Earthscan.

Meadows, D. H., Meadows, D. L., Randers, J., & Behrens, W. W. (2018). The limits to growth. In *Green planet blues* (pp. 25-29). Routledge. Various versions published since 1972.

Midgley Jr, T., & Henne, A. L. (1930). Organic fluorides as refrigerants1. *Industrial & Engineering Chemistry*, *22*(5), 542-545.

Miles, A. (2015). The (re) genesis of values: Examining the importance of values for action. *American sociological review*, *80*(4), 680-704.

Mitchell, A., Romano, G., Groisman, B. *et al.* Adaptive prediction of environmental changes by microorganisms. *Nature* **460**, 220–224 (2009). https://doi.org/10.1038/nature08112

Muller, N. Z., Mendelsohn, R., & Nordhaus, W. (2011). Environmental Accounting for Pollution in the United States Economy. *The American Economic Review*, *101*(5), 1649–1675. http://www.jstor.org/stable/23045618

Murray, R. K. (1955). *Red Scare: A study in national hysteria, 1919-1920*. U of Minnesota Press.

Oulton, N. (2012). Hooray for GDP!

Pfeffer, J., & Sutton, R. I. (2000). *The knowing-doing gap: How smart companies turn knowledge into action*. Harvard business press.

Pimentel, D., Cooperstein, S., Randell, H., Filiberto, D., Sorrentino, S., Kaye, B., ... & Weinstein, C. (2007). Ecology of increasing diseases: population growth and environmental degradation. *Human Ecology*, *35*, 653-668.

Poundstone, W. (1993). *Prisoner's dilemma: John von Neumann, game theory, and the puzzle of the bomb*. Anchor.

Powledge, T. (2011). "Behavioral epigenetics: How nurture shapes nature". BioScience. 61 (8): 588–592. doi:10.1525/bio.2011.61.8.4

Rocca, F. X. (2023, February 5). *Pope Francis Reaffirms Calls for Acceptance of Gay People*. Retrieved April 9, 2023, from https://www.wsj.com/articles/pope-francis-reaffirms-calls-for-acceptance-of-gay-people-11675620797

Sapolsky, R. M. (2005). The influence of social hierarchy on primate health. *Science*, *308*(5722), 648-652.

Schrom, D. (2004). Can we use science to know our ends. *Bioscience*, *54*(4), 284-285.

Pew Research Center, Washington, D.C ."Shared beliefs between parents and teens."(September 10, 2020).

Shukolyukov, A., & Lugmair, G. W. (1998). Isotopic evidence for the Cretaceous-Tertiary impactor and its type. *Science*, *282*(5390), 927-930.

Simis, M. J., Madden, H., Cacciatore, M. A., & Yeo, S. K. (2016). The lure of rationality: Why does the deficit model persist in science communication?. *Public understanding of science*, *25*(4), 400-414.

Sinatra, G. M., & Hofer, B. K. (2021). *Science denial: Why it happens and what to do about it*. Oxford University Press.

Song, H., Wignall, P. B., & Dunhill, A. M. (2018). Decoupled taxonomic and ecological recoveries from the Permo-Triassic extinction. *Science advances*, *4*(10), eaat5091.

Southwell, B. G., Thorson, E. A., & Sheble, L. (2017). The persistence and peril of misinformation: Defining what truth means and deciphering how human brains verify information are some of the challenges to battling widespread falsehoods. American Scientist, 105(6), 372-375. https://doi.org/10.1511/2017.105.6.372

Stokstad, Erik (5 May 2019). "Landmark analysis documents the alarming global decline of nature". *Science*. AAAS.

Tamerius, K. (2018, December 24). Opinion | How to Have a Conversation With Your Angry Uncle Over the Holidays. The New York Times. https://www.nytimes.com/interactive/2018/11/18/opinion/thanksgiving-family-argue-chat-bot.html

TEEB (2010), The Economics of Ecosystems and Biodiversity Ecological and Economic Foundations. Edited by Pushpam Kumar. Earthscan: London and Washington.

Turchin, A., & Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. Ai & Society, 35(1), 147-163.

Turner, G. M. (2008). A comparison of The Limits to Growth with 30 years of reality. *Global environmental change*, *18*(3), 397-411.

Understanding Science. (2022, September 10). *Science aims to explain and understand - Understanding Science*. Understanding Science - How Science REALLY Works. . . https://undsci.berkeley.edu/understanding-science-101/what-is-science/science-aims-to-explain-and-understand/

UNESCO. (2021). Science for society. Geneva, Switzerland. https://en.unesco.org/themes/science-society

United Nations. (n.d.). Causes and Effects of Climate Change | United Nations. https://www.un.org/en/climatechange/science/causes-effects-climate-change

Walker, M. (1963). The Nature of Scientific Thought (pp. 177:179). Prentice Hall.

Whiten, A., Goodall, J., McGrew, W. C., Nishida, T., Reynolds, V., Sugiyama, Y., ... & Boesch, C. (1999). Cultures in chimpanzees. *Nature*, *399*(6737), 682-685.

Wickiser, J. K., O'Donovan, K., Washington, M., Hummel, S., & Burpo, F. J. (2020). Engineered pathogens and unnatural biological weapons: the future threat of synthetic biology. *CTC Sentinel*, *13*(8), 1.

Woo, R. (2021, May 11). China demographic crisis looms as population growth slips to slowest ever. Reuters. https://www.reuters.com/world/china/china-2020-census-shows-slowest-population-growth-since-1-child-policy-2021-05-11/

World Health Organization. (2021). Human genome editing: recommendations. Geneva, Switzerland. ISBN: 978-92-4-003038-1.

Young, G. L. (1974). Human ecology as an interdisciplinary concept: a critical inquiry. In *Advances in ecological research* (Vol. 8, pp. 1-105). Academic Press.

Zermelo, E. (2013). *Ernst Zermelo-Collected Works/Gesammelte Werke II: Volume II/Band II-Calculus of Variations, Applied Mathematics, and Physics/Variationsrechnung, Angewandte Mathematik und Physik* (Vol. 23). Springer Science & Business Media. pp. 263.

Zollo, F. (2019). Dealing with digital misinformation: a polarised context of narratives and tribes. *EFSA Journal*, *17*, e170720.

Stanford

Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

An Axiology of Aesthetics for Existential Risk

Ishan Raval [1]*

Citation: Raval, Ishan. An Axiology of Aesthetics for Existential Risk. *Proceedings of the Stanford Existential Risks Conference* 2023, 91-103. https://doi.org/10.25740/yr107mx0459

Academic Editor: Trond Undheim, Dan Zimmer

Funding: N/A

Conflict of Interest Statement: N/A

Informed Consent Statement: N/A

Acknowledgments: N/A

Author Contributions: N/A

Abstract: This paper examines foundational axiological questions in the context of existential risk. Axiology asks what is valuable, and what it is about what's valuable that makes it such. These matters get projected to pure, raw clarity in our present precarious state as a species and civilization: What is it for that we want existential risk to be minimized and the human enterprise to continue? So far, the preponderance of existential risk discourse has ranged from pleading agnosticism regarding axiological fundaments to avowing that the ultimate good resides in the experience of well-being for individuals, i.e., avowing utilitarianism, to argue for the worthiness of addressing existential risk. Such avowals can be questioned however, and as it happens, an alternative axiology emerges from the prior question pertinent to the context of existential risk. When we look into what good we want the world to continue to exist for, we find an aesthetic orientation to the way we value things, and from there, an objective and transcendent conception of the good. Such a conception, it is argued, would be better than that offered by utilitarianism for bringing forth passion toward and about the world with the urgency and intensity existential risks call for.

Keywords: existential risk, utilitarianism, aesthetics, value theory, the Good

[1] Independent researcher

* Correspondence: ishansraval@gmail.com

1. Extant Explicit X-Risk Axiology

The axiology guiding popular existential risk discourse has two main structuring elements: agnosticism and utilitarianism, the latter understood simply as normativity regarding the happiness/well-being of humans and/or individual sentient beings.

I'll take Nick Bostrom's seminal paper, 'Existential Risk Prevention as Global Priority', as case in point. Bostrom seeks to create a categorization of risks, and for this, needs a normative framework; it is on the basis of judging what is valuable that we can best classify risk. He chooses "a simplified version of one important class of normative theories" (Bostrom, 2013, p. 16), one specifically holding that "the lives of persons usually have some significant positive value and that this value is aggregative (in the sense that the value of two similar lives is twice that of one life)" (Bostrom, 2013, p. 16), with no relevant distinction made on their value based on whether they exist now or would exist in the future. In other words, he considers a broad utilitarianism to be the most agreeable way to think about existential risks.

Still, he acknowledges that we may not have figured out the best normative framework for guiding humanity. This agnosticism he avows in fact serves as an argument for addressing existential risk; as he explains it, there is "great option value" to ensuring that we continue to exist in a non-defunct collective state so that we can ultimately figure out what is truly of value and then exist accordingly.

Discussions about existential risk have continued within this matrix, especially in popular (and more societally influential) treatments, such as *The Precipice: Existential Risk and the Future of Humanity* by Toby Ord, and *What We Owe the Future* by William MacAskill.

I will take the latter to showcase this, as it is more overtly philosophical.[1] The first three words of the first chapter of the book, 'The Case for Longtermism', state the position that the whole book is based on: "Future people count" (MacAskill, 2022, Ch. 1) a position that is developed through the chapter. One section within it, 'The Value of the Future', gets at what about the future could be valuable. While not explicitly stated as such, it expresses utilitarianism, i.e., it places ultimate value in pleasant first-person experience: "We can build a world where everyone lives like the happiest people in the most well-off countries today, a world where no one lives in poverty, no one lacks adequate medical care, and, insofar as is possible, everyone is free to live as they want (MacAskill, 2022, Ch. 1)"; "If my best days can be hundreds of times better than my typically pleasant but humdrum life, then perhaps the best days of those in the future can be hundreds of times better again" (MacAskill, 2022, Ch. 1).

In the following chapter, 'You Can Shape the Course of History', MacAskill offers a framework for judging the value of some action that we can take today. The three variables that matter, he says, are the significance of the action, its persistence and its

[1] This book posits itself as being not about existential risk, but longtermism, "the view that positively influencing the longterm future is one of the key moral priorities of our time" (MacAskill, 2022, Appendix 2). MacAskill draws a distinction between this and the framing of existential risk because "much of [his] focus is on improving the values that guide the future," an idea that "doesn't fit neatly under the category of existential risk reduction" (MacAskill, 2022, Appendix 2). Still, it appears to me that the same big picture normative frameworks can and do apply regardless of whether the central concern is taken as the long-term future or existential risk.

contingency.[2] Significance is the variable that captures the normative dimension of a state of affairs, and in the example he gives, about the counterfactual in which glyptodonts (a kind of mammalian species that existed in South America) hadn't gone extinct about 12,000 years ago, he expands beyond a purely utilitarian framework: "[W]e would want to attend to all relevant aspects of glyptodonts' extinction: the intrinsic loss of a species on the planet, the loss to humans who could have used their shells or eaten their meat, and the impact on the ecosystems the glyptodonts inhabited" (MacAskill, 2022, Ch. 2). The first of the possibilities given there is of a non-utilitarian variety; however, it is put there in a way that still begs the question of what that is an intrinsic loss of, and thus, if it is a relevant loss, begs the question of a different axiology.

The second section of the book looks at changes to society's trajectory along these three variables. The chapter 'Moral Change' delves deeper into the topics of significance and contingency. It takes for granted that axiological significance is of an ethical sort (as opposed to aesthetic, or something else). The idea that individual happiness is significant (and contingent) is argued through a detailed exploration of (the abolition of) slavery. Now, it makes sense to think about slavery in terms of the experience of undergoing it. However, the rhetorical effect of this chapter, intended or not, is to suggest that all matters of value are like this, that what matters is being in a state of happiness or sadness. It thus implicitly favors and sets up a utilitarian account for the axiology we operate with for the long term, which is tempered in the book only by agnosticism, which is the backbone of the following chapter, 'Value Lock-In'. Here MacAskill appraises the risk of baking in a set of values for all of human society that can't be altered, but does so solely in terms of the implications such lock-in would have on first-person, individually reduced happiness, and argues that we should address existential risks so that we can then undertake a Long Reflection: "a stable state of the world in which we are safe from calamity and we can reflect on and debate the nature of the good life, working out what the most flourishing society would be" (MacAskill, 2022, Ch. 4).

It is in the final section, in the chapter titled 'Is It Good to Make Happy People?', that the book goes into maximal philosophical rigor and depth, providing an overview of the field of population ethics in the context of longtermism. Without appraising these ideas themselves, it can again be noted that they center people's well-being or happiness (terms he uses interchangeably) (MacAskill, 2022, Ch. 8), except that these people include those who could exist in the future, unlike with classical utilitarianism. It is by considering the positive first-person affect that could occur but which would be precluded by human extinction or civilizational collapse that MacAskill reaches one of his main conclusions, "that we should think of the nonexistence of future generations as a moral loss" (MacAskill, 2022, Ch. 2).

2. Conceptualizing Good Better

The one common element to dominant conceptions of value in existential risk and/or longtermist discourse is negative: There isn't an objective or "out there" value to reducing existential risk that exists independently or externally to people's experiences and is irreducible to them. Rather, we are exhorted to address existential risk because of all the possible positive experience there can be, or in order to get to a place where we can figure out what really is good and act from there.

[2] MacAskill uses the term "contingency" to denote the "noninevitability" of some change (MacAskill, 2022, Ch. 2). This is similar to the sense in which I deploy the concept of a contingency in this essay and other writings, but I take it in an entirely different direction that was entirely uninformed, when the ideas here came to me, about MacAskill's work.

This may be a sufficient motivator to act on existential risk with the urgency it requires. However, going about existence with the belief that good only resides in or about individual persons' positive experiences may instill a certain alienation toward the world. At this point, we may be as blind to it as the air we breathe. But at some level of our psyche, we may yet wonder: What matters when there's nothing good beyond or irrespective of some positive state of experience in and for individual sentient beings? A conception of the good that is greater than that could be at least valuable, if not necessary, for bringing forth the passion about the world and its transpersonal challenges in the urgency and intensity that are behooved. The same argument holds for the notion that we had better wait to get to a more secure position before devoting ourselves to getting to the bottom of core axiological truths. It is now that we may most need the impelling force of a good that is bigger than individual experience to conduct the action called upon us by this fleeting epoch.[3]

It may be useful here to unpack some concepts related to the good. The first is that of an intrinsic or ultimate good. An intrinsic good isn't valuable because it causes or constitutes another greater good, but as an end-in-itself. One may ask of something that is regarded positively, "What is that good for?" That may lead to some answer. That may again lead to the question, "But why is that good?" And so on, until there's something that is considered good for its own sake, not because it lends itself to some other good. That is an ultimate or intrinsic good. In the x-risk discourse covered earlier, the answer to the question of the intrinsic good is either deferred until the Long Reflection, or is implicitly or explicitly seen as being in or about happiness or well-being.

This leads to the second concept to unpack. Such utilitarian accounts of the good are subjective accounts. That is, they hold that it is in or about the *state* of happiness, rather than the *fact* of it, that the good ultimately resides. *A la* Nagel (1974), there is "something it is like" to be happy (or be in a state of happiness), and goodness is about that subjective experience. This can be contrasted with an objective conception of the good, in which it is not the first-person experience of something that makes something good, but the fact or phenomenon of that thing regarded from a third-person point-of-view. That which is being seen from the third-person point-of-view may involve subjective experience, but it is in and about the full phenomenon of its taking place, as can be seen from, say, the point-of-view of the universe, that it is good. That can include and involve the subjective aspect. Indeed, some accounts of the good have accounted for both aspects in an inextricable fashion, e.g., Aristotle's account of eudaimonia, often translated as "flourishing," in which this flourishing isn't valued just for what it's like to be in that state of flourishing, but because the fact of such a state reflects a life being lived rationally and virtuously, and one's function (*ergon*) being fulfilled, which, according to Aristotle, are positives from an objective perspective (Aristotle et al., 2009). Yet other accounts of the good posit it in a purely objective way; G.E. Moore, for example, described goodness as a simple, non-definable quality, existing in reality much as, say, the color yellow exists, which also can't really be defined in a meaningful sense but is recognized as its own thing when we see it (Moore, 1903, Ch. 1, § 10).

[3] This is aside from the point that the model of addressing existential risk first and then setting about a Long Reflection seems overly simplistic and unidimensional. We cannot ever act *without* some normative framework, whether it's one that posits an objective, transpersonal good or not. We have a choice about what axiology to proceed with today, and whether to do so implicitly or explicitly, but not about whether to proceed about things with an axiology. That there will be, no matter what, and whatever it is will prefigure the vectors followed by a Long Reflection.

While the above argument by Moore posits an objectivity to the good, his is not a *transcendent* good, which is the third and last term to unpack here. I am using the term transcendent in a way that I think there is a general intuition about: As that which is greater than or beyond ordinary experience, that which connotes an aspect or attribute that is of some Ultimate, True or Good character more than anything connected to our mundane being. Peace or justice may be appreciated objectively and considered good not just for the experiences such a condition brings about. However, that state is still to do with this plane of existence, and if one still asks of that state or of anything of this ordinary existence, "Well, what's so good about that?" it may seem like that question is still valid, lack as it does an innate connection to or constitution by that which is good as such in a pure, definitional way. That latter would be a transcendent Good, and about it, it would be meaningless to ask "Well, what's good about that?" as it is in its very essence the Ultimate Good.

Humans have generally claimed access to or knowledge of the transcendent in the domain of what is called in the Western imagination religion or spirituality. God, Brahman and the Dao, for example, have been considered three bearers, forms or substances of such Ultimate Value. Most people raised directly with or privy to some form of theism would recognize how there is an idea of something that is Good, very much with a big-G, to the idea of God; that Good is what is meant by a transcendental good. In the Western philosophical tradition, the most notable articulation of a transcendent good is given by Plato, for whom the highest among his Forms, all of which exist independent of individual experience (or lack thereof) of them, was the Form of the Good (Plato & Bloom, 1991). For contemporary times, John Schellenberg has offered an account of the transcendent meant as an alternative to theism, or belief in a personal God. This account, called ultimism, holds "that there is a reality ultimate in three ways: metaphysically, axiologically, and soteriologically" (Schellenberg, 2016, Sec. 2 ('Ultimism')), which can respectively be understood as that which is the most fundamentally real, that which is most valuable of its own right, and that which, if accessed, leads to the best possible outcome for that which accesses it. Such being I call "transcendent" in this reflection.

When I wrote above that "A conception of the good that is greater than [positive state of experience in and for individual sentient beings] could be at least valuable, if not necessary, for bringing forth the passion about the world and its transpersonal challenges in the urgency and intensity that are behooved, this is the kind of good that I meant: Irreducible to first-person experience or "what it is like" to be someone, i.e., objective, and standing beyond such phenomenal subjectivity in an Absolute way, i.e., transcendent. Human civilization faces a choice between existence and nonexistence in a way it never has. As the idea of God has served over the ages as an impelling force for monumental, even over-the-top, pursuits, we may today also need the impelling force of a truly greater Good, ideally in a conception that isn't articulated in a way that mandates a sensibility about the divine or the sacred as past conceptions have.

As it happens, one possibility for such a framework arises from a foundational question that the reality of existential risk raises forth, and it is to this that we will now turn.

3. Toward an Axiology of Aesthetics for Existential Risk

The question of the ultimate or intrinsic good can also be asked as follows: Why should the world exist? What is the positive thing or characteristic about the world or objective existence that ultimately makes the world something whose persistence we want? This is a framing that has a ring of pertinence in this era of intensifying and cascading existential

risk. Whilst prominent x-risk discourse has explicitly only avowed a utilitarian answer to this question (that is, when not pleading agnosticism), that isn't always how it approaches the issue. Tellingly, when MacAskill in *What We Owe the Future* and Ord in *The Precipice* aren't making a rigorous philosophical case (or arguing that we should wait a few centuries before doing more philosophy), but more rhetorically attempt to arouse in their readers a proper passion for the continuation of our future or sync with some deeply felt chord of truth within our psyches, their framings often take a spontaneously non-utilitarian form.

MacAskill (2022) for example, writes in the chapter 'Extinction':

> "Now and in all the coming centuries, we face threats that could kill us all. And if we mess this up, we mess it up forever. The universe's self-understanding might be permanently lost and, within just a few hundred million years more, the brief and slender flame of consciousness that flickered for a while would be extinguished forever. The universe might return to the state it occupied for much of its first thirteen billion years: cold, empty, dead" (Ch. 5).

Even more telling is that to commence this chapter, he quotes the opening of Arthur C. Clarke's book *Rendezvous with Rama*, which describes a catastrophe befalling humanity, and says about it, "Six hundred thousand people died, and the total damage was more than a trillion dollars. But the loss to art, to history, to science—to the whole human race, for the rest of time—was beyond all computation" (Clarke, 1973, as cited in MacAskill, 2022, Ch. 5).

Ord's project in *The Precipice* (2020), as noted earlier, isn't so much to make a philosophical case for saving the world. His book is a low-down on existential risks and how we can confront them. So, freed from the obligation to uphold utilitarianism in this book—although he is a professional philosopher whose scholarly work is within the utilitarian tradition (Ord, 2014, 2015)—Ord proceeds even more to rhetorically elicit his reader's passion for the world's continuation in a non-utilitarian vein. For example:

> "[B]ecause, in expectation, almost all of humanity's life lies in the future, almost everything of value lies in the future as well: almost all the flourishing; almost all the beauty; our greatest achievements; our most just societies; our most profound discoveries. We can continue our progress on prosperity, health, justice, freedom and moral thought. We can create a world of wellbeing and flourishing that challenges our capacity to imagine. And if we protect that world from catastrophe, it could last millions of centuries. This is our potential—what we could achieve if we pass the Precipice and continue striving for a better world" (Ord, 2020, Ch. 2).

Or:

> "[A] world without agony and injustice is just a lower bound on how good life could be. Neither the sciences nor the humanities have yet found any upper bound. We get some hint at what is possible during life's best moments: glimpses of raw joy, luminous beauty, soaring love. Moments when we are truly awake. These moments, however brief, point to possible heights of flourishing far beyond the status quo, and far beyond our current comprehension" (Ord, 2020, Ch. 1).

The latter is more overtly within a utilitarian framework, but it still raises the question, is the highest good being posited to be the subjective experience of the good life, or things that are possible during life's best moments, such as raw joy, luminous beauty and soaring love, regarded from a third-person perspective not for the state of experiencing them but

for the very fact of them? Does value reside in the experience of being truly awake, or that which we would be truly awake to?

That it might be the latter is suggested by statements such as: "We often treat the value of cultural traditions in this way. We see indigenous languages and ways of life under threat—perhaps to be lost forever to this world—and we are filled with a desire to preserve them, and protect them from future threats" (Ord, 2020, Ch. 2). And: "[I]f we can venture out and animate the countless worlds above with life and love and thought, then … we could bring our cosmos to its full scale; make it worthy of our awe. And since it appears to be only us who can bring the universe to such full scale, we may have an immense instrumental value, which would leave us at the center of this picture of the cosmos. In this way, our potential, and the potential in the sheer scale of our universe, are interwoven" (Ord, 2020, Ch. 8). We can thus see that when intuitively arguing the case that it would be preferable if the world goes on, even writers who otherwise espouse utilitarianism aren't necessarily writing to the effect of, "Imagine the awesome feelings we could have by experiencing the amazing things there could be," but rather often to the effect of "Imagine the amazing things there could be" (even if experiences are among those amazing things). In other words, they are appealing to a sensibility that values things not for what they are *subjectively* (or what it is like to experience them), but for what they are *objectively* (as they are regarded from a third-person point-of-view).

What such intuitive expressions also reveal, though, is that this sensibility is also one that values things not for some *ethical* value, but an *aesthetic* value. That which intuitively strikes us as good about the persistence of complex life or sentient civilization, about salubriously crossing the precipice of existential risk we stand at, or about the future that we could create, has some general positive quality that we may call "amazing," that may make us go "Wow!" or which may be best represented for different people by some other articulation, but which is essentially aesthetic. It is with an aesthetic disposition that we have in us evoked the sentiment that "*This* is why the world should exist!" or "*This* is why our enterprise should persist!"—which is, ultimately, the sentiment that Ord and MacAskill wish to bring forth in us through their writings.

What is the connection between the objective aspect to what strikes us as good and its aesthetic nature? We have but seen that there is in or about us a sensibility that intuitively, at least, sees that intrinsic or final value—that for which ultimately we want the world to go on—lies in things objectively. The mechanics of this are yet to be seen, i.e., why and how we value things in an aesthetic mode when we value them for their own sake as described. We will now turn to this.

4. Aesthetic Value and the Metaphysical Break

We are prone to having our aesthetic passions whetted not just while contemplating future possibilities, but also in the course of many experiences we may have in the present. These experiences aren't necessarily about art-objects, or things in a narrow conception of a domain of aesthetics (Dewey, 1934; Hepburn, 2004; Leddy, 2005; Saito, 2007). These may be experiences of the Earth—say, seeing the vista of Yosemite Valley or some other scenic wonder—or of the world of human creations—not just listening to a marvelous symphony but also appreciating a clean, simple, well-kept home. We may have some positive experience in these scenarios. But such experience is an epiphenomenon of finding these things good in some way. If we feel happy, it is because we find value in what we are experiencing, not in the happiness; indeed there wouldn't be happiness if that was all that we were to value.

Even our most cherished engagements with other humans are aesthetic matters: When we love someone, apart from our personal affect-ion and enjoyment of them, we tend be glad about their existence whether or not we are there to enjoy that existence; indeed, what separates love from a more casual liking may be that when we love someone, we regard them as a beautiful entity, worthy of an admiration or adoration that wills their existence regardless of whatever may happen to the one willing it, even if we are to die or not benefit from their being. When we genuinely, with our hearts, wish well for someone, we do so not because that is the morally correct thing to do, but because we find them a being of some quality that merits such wishes.

It is for such things, regarded objectively and aesthetically, that one has the sense of affirmation about the world that might be articulated in forms such as, "I want the world to exist so that there can be things like *this*" or "Yes, *this* is why there should be things." But what is it about the "*this*" that leads to such affirmation? What is the property of the objectivity we are seeing on account of which we value it?
One intrinsic fact about that "*this*" is that it is, by definition, the central variable in possible answers to questions such as "Why should there be things?", "What's the point to there being reality/anything?" or "Why should there be something rather than nothing?" Now, these questions, being questions regarding the point or purpose of some X, are necessarily asking about something that isn't X, as to ask "Why should there be X?" is to ask "For what that is not X should there be X?"

For the questions above, X is our scope of reality. So, whenever we have the sense that "*This* is why there should be things," whatever "*This*" specifically happens to be, it is something fundamentally unlike or standing apart from default or "status quo" reality. When we value something as an ultimate end, as that which calls for the existence of the world itself, it is something we take as a break, rupture, or discontinuity with reality. We see the natural course of reality as something that seemed to have nothing about it that necessitated this new entity coming about from it. Yet it has, and it is thus a *contingency*—something that need not have existed, yet does—and that we find amazing. What we experience phenomenologically as aesthetic quality is ontologically contingency.

Such a break with reality is inherently something—and the only kind of thing—that can make us think and feel that reality is justified or "worth it." After all, something is "worth it" or "has a point" when it is a necessary factor in bringing about something else that is fundamentally different from or in excess of being thus far. If that happens, i.e., if there comes to be some Y that need not have existed, thanks to (something about) the X which it emerges from, then X has had a point, its existence is in some way justified. This is the case for reality as such, and that is why we want there to be a world (or worlds) of this magical kind in which wonderful entities and connections thereof come to be.

We have thus seen how there is an objective valuation of some entities of reality—actual and possible—in an aesthetic mode; it is because we see contingency generated there. Seeing value in such entities of reality for their objective being, which we experience as valuing their aesthetic merit as an end-in-itself, leads to a pro-attitude toward the continued existence of the world. As such, the pro-attitude toward (things of) reality regarded aesthetically is thus the most basic psychological disposition with direct teleological linkage to the great task and mighty labor of applying ourselves to existential risk. If we want to maximize our chances that we don't all die or get rendered defunct far below our collective capacities, it's because we think that in, with or through us (or beings

like us), reality can bear great, ineffable beauty and other aesthetic marvels, which we value as an end-in-itself.

But while it is an aesthetic pro-attitude about things of the world that brings about a pro-attitude towards the world's existence, are these pro-attitudes reflective of a higher or absolute good about that which they are directed to? These amazing things in-themselves, rather than our experience of them, strike us as "good," at least in the sense of seeming like a "good" reason that the world should go on. But is this just something that's just "nice" or "cool," or is there something really Good, in an absolute sense, in a transcendent way, at play here?[4]

4. The Transcendent Good

Above we looked at one cognition that may underlie the experience of contingency, viz. "It's so great that the world exists (so that there can be such contingency)!", "*This* is why there should be things!" or something to that effect. That is a cognition that arises subsequent to this experience. There is a cognition, though, that constitutes such experience, even if it does not rise to its conscious fore. Namely, when we apprehend something of great aesthetic notability, somewhere and somehow we are cognizing: "What are the odds that something could be so X yet so Y?"

For example, on apprehending Yosemite Valley, one may be amazed due to some latent cognition going "What are the odds that something could be so compact yet so rich?", "What are the odds that something could be so casually positioned amidst nearby mundanity yet be so unlike it?" or some such probabilistically overwhelmed cognition. But whatever the latter term is for the given case of odds of contingency, it is something that has a positive essence. This can be observed from directly looking at the latter terms that comprise this cognition underlying aesthetic experience, and in looking directly at them, seeing their inadequacy: We may use terms such as "rich" or "grand" or "beautiful" or "amazing" to describe what we are experiencing for its aesthetic quality, but the question is always begged: Why does that perceived improbability strike us as being of a positive aesthetic nature? It does so because it is, ultimately and simply, good.

Indeed, no matter what traits may come across as more overt or apparent for any particular case of X and Y, all such instances are but variations on a single theme or truth of the aesthetic experience, captured at an essential level by this most fundamental cognition: "How can something be so *real* yet so *good*?"[5] Whether it's being awestruck by the vista of Yosemite Valley or filled with joy about someone we love, it is this we are

[4] Apart from this question, there may be two others that emerge for the reader: 1) Is the perception of contingency just that—an appearance that we find to be such—or is there actually, at a metaphysical level, contingency? 2) Why is the feeling of finding something to have aesthetic quality marked by pleasure or joy? The first is a truly tricky question. There is a way in which I think the actuality of contingency can be argued for. A large part of that argument involves an elaboration of the kind of subject that we can see that we are when we analyze the centrality of the aesthetic mode to our interest in existence; I haven't the scope to make this elaboration here, but in brief, as pertains to this question, if the subject has free will, even in a compatibilist mode, as I ascribe to, then there is actually contingency (of a kind compatible with and interwoven with necessity). The answer to the second question also requires, in part, the deferred elaboration about the kind of subject we can realize we are in this aesthetic context; however, a large part of the answer lies simply in the fact that there is, as is argued ahead, an encounter with the good qua good in the course of aesthetic experience.

[5] A conceptualization of the aesthetic that is parallel to this one may be the tension between the Apollonian and the Dionysian as described by Nietzsche (2008).

moved to aesthetic rapture in trying and failing to compute.[6] Whatever we are experiencing is real; it's right here in front of us so that much is self-evident. But how can something of this plane of plain reality be so, so... *good*?—that is the pure quality or principle we have come to sense and that we grasp at, regardless of the specific elements of realness with which the other aesthetic elements, seeming to mark a break with the real, seem so improbable.

While it would have still been meaningful to ask "Well, what's good about that?" of the prior account of contingency, with the sense of the amazing that we feel about that and the pro-attitude towards the world it leads to, it would make no sense to ask such a question here. It could only evoke a tautology: Good is good, and this is, indeed, good *qua* good that we have arrived at. It is not good because of some other property, and it is not good that is explainable or contained within the entities and experiences of phenomenal reality, though those are the only route through which we, being phenomenal beings, can access it. It is, in other words, the transcendent good. The good in-itself, as a quality or principle, is thus at the foundation of what we regard as the domain of the aesthetic. An axiology that offers a transcendent good is apparent to subjects such as us through aesthetics, though it grounds and substantiates aesthetics, and such an axiology we have now found.

5. Further Discussion

I will conclude by elucidating some possible connections between the ideas laid out here, and others that have been presented through civilizational history.[7] (This section may be of interest only to those of a scholarly bent, or those whose interest in this conception of the good depends on its substantiation through or linkage with others.)

Most obvious may be how this explanation of the good dovetails with perhaps the most famous account of a transcendent good: that of Plato. I could well explain the general positive-ness that came to be apparent at the foundation of the domain of the aesthetic as being the Platonic Form of the Good; after all, just as the Good comes to be an integral part of beauty and its experience here, Beauty and other Forms, for Plato, come from or are made of the Form of the Good (Plato & Bloom, 1991, Book VI).

But what of these other Forms that come from the Good? For Plato (1952), Beauty was the most directly accessible of the Forms, a position that also finds prominence in the value theory of a far more recent Platonist, Iris Murdoch. For Murdoch (1970), aesthetic experience gets us out of our ego-oriented existence. The kind of selfless, attentive perception that marks aesthetic experience has for her a moral value, as it takes us away

[6] The philosophically educated reader may not point out that the aesthetic quality I'm discussing is Kant's sublime: For Kant (2004), we have the aesthetic experience of the sublime when we can't wrap our minds around something. (The sublime he contrasts to beauty, with which the imagination and understanding can engage in a disinterested fashion to find order and harmony about it, experiencing which yields a pleasurable feeling.) I believe that a vaster range of aesthetic experiences are marked by the kind of failure to compute than which Kant ascribed to the sublime, such as a mountain on account of its size or a storm on account of its power; e.g., love for another person, as related above, also may involve a similar cognitive overload. For that reason, I prefer to talk about general aesthetic quality rather than the sublime, as I feel the latter has connotations of a narrower range of things than that which I'm describing, even if the mechanism underlying the broader range is very similar to what Kant described for the sublime.

[7] This will not include a discussion of the connection that I am personally most interested in, and which I thus can't help but mention. Namely, I believe that the conception of the good and the aesthetically oriented existence as given here may align with the Vedantic conception of the Ultimate (called Brahman in that tradition) as *Saccidānanda*, or made up of truth, consciousness and happiness. Explicating this possibility, however, would make for another essay altogether, and thus will have to be left aside for here.

from our self-centered headspace to things that are more real and valuable. While for Plato or Murdoch beauty isn't necessarily the only way to come to a broader valuation of the world apart from us—a valuation that wanting the world to not end is an exemplary specimen of—that is at least a choice way in which that happens.

I will remain agnostic for now on whether the aesthetic domain enjoys a special relationship with the Good. However, I find it arguable that it can only be in an aesthetic mode that we find something to be an ultimate or intrinsic end, i.e. something for the sake of which to want that the world goes on. Even if there is something Good about happiness, in a Platonic way whereby the Form of the latter is made up of the Form of the Good, the very question of what is the Good such that we want the world to exist (so that the former can manifest) involves looking at the contenders for the answer in an objective, third-person way. Having awareness of what is Good or has about it the Good is an activity that necessarily involves taking a third-person perspective towards things; even if an experience is in-itself Good, it has to be regarded objectively for its Good. And it may be that this objective standpoint of perception for or towards the Good *is* itself the aesthetic mode of orientation towards things. If Beauty enjoys an exclusive relationship with the Good, it may be only insofar as contemplating this possibility requires cognizing what the Good might be, which requires an objective, aesthetic standpoint towards reality.

So while for Murdoch the aesthetic is ethical in a sense that considers the Good synonymous with the ethical and sees that an aesthetic orientation is a way of accessing that, it could be as valid to think of the ethical as aesthetic, such that values that we consider ethical or moral are only Good in the form of an aesthetic consideration—we find peace, justice or kindness to be good because we find something beautiful about them. This is a view of ethics that has semblance with (at least a particular rendition of) Confucian philosophy. For Confucius (1938), a life well-lived—that is, one which actualizes traits such as *ren* (best translated as humaneness, or moral excellence) and *yi* (best translated as righteousness)—can only be coherently conceived of as such in the context of the social whole. This understanding not only gives Confucianism an objective edge—value resides at the level of the social plane, and thus cannot be experienced by an individual in the first-person—but also a distinct aesthetic edge, through the centrality of *li*, which could be translated as ritual propriety. The best existence is thus in any meaningful sense only collective, and manifests as a dance (Ihara, 2004)—it is performed in harmony with others, and has a quality to it that is aesthetic.

Of course, Confucius does not hold that aesthetic quality (or any single quality, really) has a special claim over a reified transcendent good. It's not that *ren* and *yi* should be cultivated, in this framework, *so that* there is *li*, with its aesthetic characteristics. I bring up the Confucian example as just a way of seeing that ethics too can be aesthetic, and thus accounting for ideals traditionally belonging to the domain of ethics, such as freedom, fairness or honesty, within the Good of this framework as well. With that, one has reason to retain their own certain ethical commitments while also coming to hold a self-conscious aesthetic orientation about what one ultimately wants.

After all, as we have seen here, it is such an objective, aesthetic orientation that fills us with a pro-attitude regarding the world's continued existence; the Good may not necessarily be exclusively or exhaustively aesthetic, but it is an aesthetic orientation through which we can contemplate, and thus be inspirited by, the Good. But advancing the strength and ubiquity of such a pro-attitude to the utmost may be critical to live up to the epochal challenges of existential risk, and as such, so might be fostering the aesthetic orientation that enables that attitude. To the extent that these ideas have validity, then, it is hoped that philosophical validity is of enough aesthetic appeal to play a nontrivial role

in fostering that orientation, for that, ultimately, is the singular objective in excavating this axiology.

6. Conclusion

We have assayed dominant existential risk discourse to discern how it holds that value (to addressing existential risk, and in general) resides ultimately in the states of well-being that could be experienced for much longer if the human enterprise doesn't meet a premature demise. There is a paucity to such an account of the good, it has been argued, and the position developed that in order to motivate the most vigorous action with respect to existential risk, we may be better off being able to believe in an ultimate good that is objective and transcendent. One such conception of the good is found by digging into the very question of why we want the world to keep existing. Having done so, we now have an idea not only about what may be at the bottom of a pro-attitude toward this existence—entities (and connections thereof) of aesthetic excellence, or in ontological form, contingency—but also how contingency is related to value *qua* value. A proposal has thus been made for a well-grounded axiology, which, it is hoped, can be harnessed to dispose us toward all the building and buoying that it will take to face up to the exigencies of existential risk.

References

Aristotle., Ross, W. D. 1., & Brown, L. (2009). *The Nicomachean Ethics*. Oxford University Press.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002

Clarke, A. C. (1973). *Rendezvous with Rama*. Gollancz.

Confucius (1938). *The Analects of Confucius* (A. Waley, Trans.). George Allen & Unwin.

Dewey, J. (1934). Art as Experience. Perigee Books.

Hepburn, R. (2004). Contemporary aesthetics and the neglect of natural beauty. In A. Carlson, & A. Berleant (Eds.), *The Aesthetics of Natural Environments* (pp. 44-62). Broadview Press.

Ihara, C. K., 2004, "Are Individual Rights Necessary? A Confucian Perspective," in K.-L. Shun & D. B. Wong (Eds.), *Confucian Ethics: A Comparative Study of Self, Autonomy, and Community* (pp. 11-30). Cambridge University Press.

Kant, I. (2004). *Critique of Practical Reason* (T. K. Abbott, Trans.). Dover Publications.

Leddy, T. (2005). The Nature of Everyday Aesthetics. In A. Light & J. M. Smith (Eds.), *The Aesthetics of Everyday Life* (pp. 3-22). Columbia University Press.

MacAskill, W. (2022). *What We Owe the Future: A Million-Year View*. OneWorld Publications.

Moore, G. E. (1903). *Principia Ethica*. Cambridge University Press.

Murdoch, I. (1970). The Sovereignty of the Good Over Other Concepts. In I. Murdoch, *The Sovereignty of the Good* (pp. 77-104). Routledge.

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914

Nietzsche, F. (2008). *The Birth of Tragedy*. Oxford University Press.

Ord, T. (2014). Global poverty and the demands of morality. In J. Perry (Ed.), *God, The Good, and Utilitarianism: Perspectives on Peter Singer*. Cambridge University Press.

Ord, T. (2015). A new counterexample to prioritarianism. *Utilitas*, *27*(3), 298–302. https://doi.org/10.1017/s0953820815000059

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.

Plato., & Bloom, A. (1991). *The Republic of Plato*. 2nd ed. Basic Books.

Plato. (1952). *Plato's Phaedrus*. Cambridge University Press.

Saito, Y. (2007). *Everyday Aesthetics*. Oxford University Press.

Schellenberg, J. (2016). God for All Time: From Theism to Ultimism. In A. Buckareff & Y. Nagasawa (Eds.), *Alternative Conceptions of God: Essays on the Metaphysics of the Divine*. Oxford University Press.

Section 2
Crises in the Earth System

Stanford
Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Navigating Cascading Planetary Boundaries: A Framework to Secure the Future

Tom Cernev [1]*

Citation: Cernev, Tom. Navigating cascading planetary boundaries: A framework to secure the future. *Proceedings of the Stanford Existential Risks Conference* 2023, 105-118. https://doi.org/10.25740/ww663gw1454

Academic Editor: Trond A. Undheim, Dan Zimmer

Funding: N/A.

Conflict of Interest Statement: N/A.

Informed Consent Statement: N/A

Acknowledgments: N/A.

Author Contributions: N/A.

Abstract: This paper examines the potential cascading effects into society of transgressed Planetary Boundaries. This paper provides a literature review of the Planetary Boundaries (climate change, novel entities, stratospheric ozone depletion, atmospheric aerosol loading, ocean acidification, biogeochemical flows, freshwater use, land-system change, and biosphere integrity), including an analysis of humanity's current position within each of the Earth systems. Using this, the potential for cascades across Planetary Boundaries, and from transgressed Planetary Boundaries into society is investigated. Extensive cascades exist, and agriculture and food security found to be at risk. It is investigated how the cascades through society could lead to increased Global Catastrophic Risk, and found that transgressed Planetary Boundaries have the potential to exacerbate known Global Catastrophic Risk (e.g. Weapons of Mass Destruction), and may lead to new and unforeseen Global Catastrophic Risks (e.g. as a result of the development of new climate change mitigating technology). From these findings, a framework is generated to identify and understand the cascades and associated Global Catastrophic Risks that may stem from crossed Planetary Boundaries, and also how to mitigate them. The framework emphasises the need for involvement from across industry, governments, and research institutions.

Keywords: planetary boundaries, cascading risk, sustainability, global catastrophic risk, climate change

[1] Research Affiliate, Centre for the Study of Existential Risk, University of Cambridge, United Kingdom; tcc38@cantab.ac.uk.

* Correspondence: tcc38@cantab.ac.uk

1. Introduction

Humanity is facing increasing environmental risks, and consequences, that stem not only from greenhouse gas emissions and climate change but from wider human impacts on Earth systems. Ultimately this is putting society at risk. Humanity's impact on Earth systems can already be observed, and there are far reaching consequences that could cascade across society. Human activity is further putting Earth systems, and thus wider society, at risk, with feedback loops that could potentially lead to tipping points, and high levels of uncertainty due to the complexity of both Earth system interactions and society making it difficult to quantitatively or qualitatively identify what the consequences may be.

A comprehensive representation of Earth systems is provided by the Planetary Boundaries (Rockström et al., 2009; Steffen et al., 2015). Consisting of nine Earth systems, the Planetary Boundaries provide a safe operating space for humanity with regards to the Earth systems involved (Steffen et al., 2015). Importantly, whilst it is acknowledged that the Earth systems that make up the Planetary Boundaries interact with one another, and thus do not exist in isolation, little is known about what interactions exist between the Planetary Boundaries (Steffen et al., 2015; Wang-Erlandsson, 2022), and subsequently what risk cascades could occur not only across the boundaries, but also into wider society, where cascading risk in this paper is considered from the point of view of understanding causal interdependencies and through the toppling dominos metaphor (Pescaroli & Alexander, 2018). This paper provides a background to the Planetary Boundaries, including humanity's current position in them, before assessing how the Planetary Boundaries interact with one another, and the possible risk cascades that may exist. Following this the potential for the Planetary Boundaries to impact wider society, and lead to Global Catastrophic Risk is examined by considering cascades out and into society. Using the findings from these investigations, a framework to help guide policy development and facilitate better understanding and mitigation of the cascading risks that may result from Planetary Boundaries that have been transgressed is generated, and possible next steps are considered.

2. Background

This section provides an overview of important background information, and existing information in the literature, for Planetary Boundaries, and Global Catastrophic Risk.

2.1 Planetary Boundaries

The nine Planetary Boundaries are based on critical Earth processes that regulate Earth system functioning, refer Figure 1, and aim to define a safe operating space for humanity, and provide a means for humanity to recognise that an Earth system boundary has been crossed and respond before any subsequent tipping point is reached (Steffen et al., 2015; Rockström et al., 2009; Wang-Erlandsson et al., 2022; Baum & Handoh, 2014). Whilst the Planetary Boundaries can be considered, and sometimes quantitated, on regional scales, the complexity and interconnectedness of modern day society a global approach is preferable despite the quantitative uncertainty when taking a global approach (Steffen et al., 2015). Notably, a Planetary Boundary is not tipping point, rather in each case the Planetary Boundary is placed upstream of any threshold, predominately to compensate for the high level of uncertainty in the Earth systems due to their complexity and underlying connectedness (Steffen et al., 2015), however, it is likely that interactions between Planetary Boundaries will reduce the boundary level (Lade et al., 2020). On the other side of the boundary is a zone of uncertainty (increasing risk), and beyond zone of uncertainty (high risk) (Stockholm Resilience Centre, 2021). Humanity's position across a

particular Planetary Boundary is determined by measurements against at least one control variable that is relevant to a particular Earth System (Steffen et al., 2015; Rockström et al., 2009).

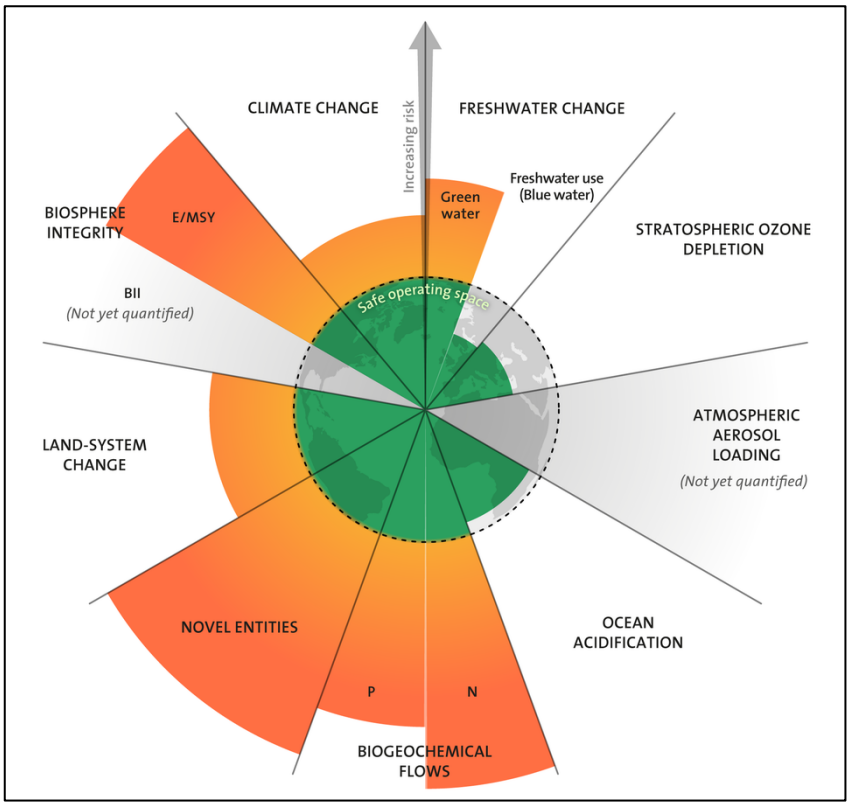


Figure 1: The nine Planetary Boundaries, and humanity's position within them (Azote for Stockholm Resilience Centre, based on analysis in Wang-Erlandsson et al 2022).

Whilst early assessments of the Planetary Boundaries had several boundaries crossed, or there existing large uncertainties as to humanity's position within them (Rockström et al., 2009; Steffen et al., 2015), recent assessments have illustrated that humanity has transgressed numerous Planetary Boundaries (Persson et al., 2022; Wang-Erlandsson et al., 2022), and currently resides beyond zone of uncertainty (high risk) for numerous Planetary Boundaries, refer Figure 1. In humanity's current position there is a high risk of non-linear, and irreversible change, and the potential crossing of tipping points. It has been suggested that transgressing a Planetary Boundary is high-risk when there are reinforcing feedback loops, however, there is a high-level of uncertainty in identifying and quantifying these (Beard et al., 2021; Kareiva & Carranza, 2018).

From Figure 1, whilst all Planetary Boundaries have at least one control variable associated to them, there still exists extensive uncertainty with regards to humanity's position within or beyond the boundaries. In two cases, there is not a quantification: that of the Atmospheric Aerosol Loading Planetary Boundary, and the Biodiversity Intactness Index control variable that contributes to the Biosphere Integrity Planetary Boundary. Furthermore, it is recognised that control variables for Planetary Boundaries will have to be further developed in order to better represent the Earth systems, including that of the Biogeochemical Flows Planetary Boundary. Updates to the Planetary Boundaries since their original publication (Rockström et al., 2009) have occurred, particularly for the Novel Entities, Land-system Change, and Freshwater Change Planetary Boundaries (Steffen et al., 2015; Stockholm Resilience Centre, 2020; Persson et al., 2022; Wang-Erlandsson et al., 2022).

2.2 Global Catastrophic Risk

Whilst definitions of what exactly constitutes a Global Catastrophic Risk differs (Baum & Handoh, 2014), generally Global Catastrophic Risks are risks with "the potential to inflict serious damage to human well-being on a global scale" (Bostrom & Cirkovic, 2008). Global Catastrophic Risks have the potential to result in significant loss of life, and in the event that one eventuates, will likely leave society in a state of reduced resilience, and hence increasingly vulnerable to future risks (Avin et al., 2018). Further to this. Existential Risk is "a risk that threatens the destruction of humanity's longterm potential" (Ord, 2020), whether this be through extinction or a permanent collapse of civilisation of which climate change is possibly one (Ord, 2020), such that it is not possible for humanity to recover (Bostrom & Cirkovic, 2008). Global Catastrophic Risks could result in fatalities up to approximately 10% of the global population, and the eventuation of a Global Catastrophic Risk could potentially start a chain reaction, and if a chain reaction does not start, then humanity would likely have lower levels of resilience and be more vulnerable to future risks (Turchin & Denkenberger, 2018). Such chain reactions could potentially result in or meet the definition of an Existential Risk.

Research to date has developed the means to identify and classify possible Global Catastrophic Risks, including through critical systems (Avin et al., 2018), and communication scales (Turchin & Denkenberger, 2018). Critical systems including sociotechnological, ecological, and biogeochemical have been identified that share similarities to the Planetary Boundaries (Avin et al., 2018). Whilst numerous Global Catastrophic Risks may be extracted through the definition of such critical systems, the following risks will be used for the purposes of this paper due to the inclusion of the Earth system related risks (Global Challenges Foundation, 2022), with the assumption that a severe Global Catastrophic Risk or chain reaction of Global Catastrophic Risks could result in Existential Risk or an Existential Risk event occurring.

3. Cascades across Planetary Boundaries

In this section the cascading effects from one Planetary Boundary to others as it is transgressed are considered and examined. The nine Planetary Boundaries do not exist in isolation, and there are extensive interconnections between them, whereby the transgression of one Planetary Boundary may contribute to the transgression of other Planetary Boundaries (Steffen et al., 2015; Wang-Erlandsson, 2022; Galaz et al., 2012; Lade et al., 2020). In order to develop effective policy, and fully understand the risks to society, the stabilizing and destabilizing feedbacks need to be identified (Steffen et al., 2015) as do any cascading effects. Concerningly, these feedbacks do not have to be linear, with it possible that the transgression of one Planetary Boundary may result in non-linear change for another Planetary Boundary (Rockström et al., 2009).

Furthermore, transgressing Planetary Boundaries significantly increases the risk of reaching thresholds or tipping points, should they exist for particular Earth systems (Rockström et al., 2009). In the case where a global threshold or tipping point does not exist for a particular Earth system, collapses in the system may still result in feedbacks that lead to a global threshold (Rockström et al., 2009), and the transgression of Planetary Boundaries at sub-global levels can have global effects (Steffen et al., 2015). Overall, it is likely that interactions between the Planetary Boundaries reduces the boundary levels, ultimately increasing the risk of transgressing Planetary Boundaries (Rockström et al., 2009).

From the nine Planetary Boundaries, two have been identified in the literature as core Planetary Boundaries, that of Climate Change, and Biosphere Integrity (Steffen et al., 2015; Wang-Erlandsson, 2022). Transgressing these Planetary Boundaries risks leading the

Earth systems towards an irreversible state (Wang-Erlandsson, 2022). Table 1 provides possible effects of the transgression of the boundary, or approach to the boundary of the Planetary Boundaries on each other. In Table 1, where a particular Planetary Boundary consists of more than one control variable, the higher risk control variable value has been used, i.e., for the case of Biosphere Integrity, which consists of E/MSY (transgressed) and BII (not yet quantified), for the purposes of Table 1 and the following analysis in this paper the overall boundary is considered to be transgressed.

Table 1: Cascades across Planetary Boundaries. As a Planetary Boundary in the left column has been transgressed (or is approaching such a state), possible effects are provided, that then have the potential to cascade onto humanity's position in the Planetary Boundary in the right column. Humanity's current position in Planetary Boundaries is illustrated by: red have been transgressed, green are below the boundary, and grey are yet to be quantified (Galaz et al., 2012; Steffen et al., 2015; Wang-Erlandsson, 2022; Rockström et al., 2009; Persson et al., 2022; Richards et al., 2021; Nash et al., 2017; Lade et al., 2020; Beard et al., 2021).

Planetary Boundary	Planetary Boundary Transgression Effects	Cascades onto Planetary Boundary:
Climate Change (core)	• Increases ocean acidification levels • Ecosystem change/destruction • Changing resource demands	⇒ Ocean Acidification ⇒ Land-system Change, Biosphere Integrity ⇒ Freshwater Change
Freshwater Change	• Ecosystem change/destruction • Reduced carbon sinks	⇒ Biosphere Integrity ⇒ Climate Change
Stratospheric Ozone Depletion	• Ecosystem change/destruction	⇒ Biosphere Integrity, Land-system change
Atmospheric Aerosol Loading	• Increased pollution	⇒ Biosphere Integrity, Climate Change
Ocean Acidification	• Reduced ocean CO_2 absorption capability • Ecosystem change/destruction	⇒ Climate Change ⇒ Biosphere Integrity
Biogeochemical Flows	• Ecosystem change/destruction	⇒ Biosphere Integrity
Novel Entities	• Increased pollution	⇒ Biosphere Integrity, Climate Change, Stratospheric Ozone Depletion, Atmospheric Aerosol Loading
Land-system Change	• Ecosystem change/destruction • New agricultural/industry activity	⇒ Biosphere Integrity ⇒ All Planetary Boundaries
Biosphere Integrity (core)	• Ecosystem change/destruction	⇒ Freshwater Change, Land-system change

Table 1 aims to identify the effects on a particular Earth system as its related Planetary Boundary is transgressed, and then how this subsequently may impact other Planetary Boundaries through interactions. These interactions, and the subsequent causal transgression of Planetary Boundaries due to an initial Planetary Boundary being transgressed can be considered amplifying effects, i.e. it is likely that the possible effects of one Planetary Boundary being transgressed would not lead to humanity moving into the safe zone below the boundary for another Planetary Boundary. Rather, transgressed Planetary Boundaries lead to more transgressed Planetary Boundaries (Lade et al., 2020).

From Table 1 some Planetary Boundaries have effects cascading onto them more than others. Importantly, from Table 1, the Biosphere Integrity Planetary Boundary is cascaded

onto by all other Planetary Boundaries, and the Planetary Boundaries for Climate Change, and Land-system Change have the most cascades onto other Planetary Boundaries. This presents the question of if the Land-system Change Planetary Boundary should be a core Planetary Boundary.

4. Cascades from transgressing Planetary Boundaries

4.1 Cascades into society

The previous section investigated the cascading effects that may occur across the Planetary Boundaries as they are transgressed. However, it is also important to consider how these may cascade into society. There has been increasing awareness that climate change may result in global catastrophe, and that understanding in these areas is underdeveloped (Richards et al., 2021; Kemp et al., 2022). The Planetary Boundaries include a wider perspective than solely climate change, and hence the understanding of how transgressed Planetary Boundaries may cascade into society is even less understood. Notably, the Planetary Boundaries provide a means to identify possible areas of Earth system risk, without providing extensive information on the impact to societies/humanity (Beard et al., 2021; Baum & Handoh, 2014). Whilst the societal effects of the transgression of Planetary Boundaries is likely to be similar to those of climate change, especially since the Climate Change Planetary Boundary is considered to be a *core* Planetary Boundary, and all Planetary Boundaries relate to Earth systems, there is a high level of uncertainty. Developing an understanding of what the future societal effects of Planetary Boundary transgression may be is important not only because it enables appropriate risk management and policy creation and implementation, but also because of the history of environmental change being a factor in societal collapses (Richards et al., 2021; Kemp et al., 2022). In Table 2, the third column, *cascades into society*, has been developed by asking the question: "From the transgression effects of a Planetary Boundary being crossed, what are the subsequent possible cascading effects into and across society?"

In developing Table 2 and understanding the possible cascading effects into society from transgressed Planetary Boundaries, a variety of sources were consulted that provide insight into the possible impact of environmental change on society (Global Challenges Foundation, 2022; Stockholm Resilience Centre, 2021; Richards et al., 2021; Kemp et al., 2022; United Nations Office for Disaster Risk Reduction, 2022; Beard et al., 2021). The sources provide differing levels of detail and focus, for example Richards et al. (2021) focuses on the causal interactions between climate change, food insecurity, and societal collapse, whereas Kemp et al. (2022) focus on climate endgame scenarios. From Table 1 cascades from transgressed Planetary Boundaries onto other Planetary Boundaries were found to generally be cascades onto the Climate Change Planetary Boundary, hence, many of the cascades into society from transgressed Planetary Boundaries will likely be those associated with climate change as explored by Richards et al. (2021), and Kemp et al. (2022).

From Table 2, the potential impact across society from the transgression of Planetary Boundaries is extensive, however, there exist commonalities. In particular, the cascading effects of the transgression of Planetary Boundaries through society are extensive for agricultural stability, and food security. This is highly relevant, as historical cases of societal collapse have been influenced by unsustainable agriculture, which subsequently effects population, and can cause migration and conflict (Richards et al., 2021). From Table 2, the Climate Change, Biosphere Integrity, and Land-system Change Planetary Boundaries are the most influential when considering cascades into society.

Table 2: Possible cascading effects into society from transgressed Planetary Boundaries, incorporating the results of Table 1.

Planetary Boundary	**Planetary Boundary Transgression Effects**	**Possible cascades into Society**
Climate Change (core)	• Increases ocean acidification levels • Ecosystem change/destruction • Changing resource demands	• Climate Change fueled events, famine, migration, conflict, agricultural disruption, food insecurity, political instability, human population health (Kemp et al., 2022; Richards et al., 2021; Beard et al., 2021; Global Challenges Foundation, 2022)
Freshwater Change	• Ecosystem change/destruction	• Famine, migration, conflict, agricultural disruption, food insecurity, political instability (UNESCO World Water Assessment Programme, 2023; UNESCO World Water Assessment Programme, 2021; Kemp et al., 2022; Richards et al., 2021)
Stratospheric Ozone Depletion	• Ecosystem change/destruction	• Human population health (Stockholm Resilience Centre, 2021)
Atmospheric Aerosol Loading	• Increased pollution	• Human population health, climate migration, agricultural disruption, food insecurity, political instability (Stockholm Resilience Centre, 2021; Steffen et al., 2015; Kemp et al., 2022; Richards et al., 2021)
Ocean Acidification	• Reduced ocean CO_2 absorption capability • Ecosystem change/destruction	• Agricultural disruption, food insecurity, political instability (Stockholm Resilience Centre, 2021; Pecl et al., 2017; Kemp et al., 2022; Richards et al., 2021)
Biogeochemical Flows	• Ecosystem change/destruction	• Agricultural disruption, food insecurity (Stockholm Resilience Centre, 2021)
Novel Entities	• Increased pollution	• Pollution leading to agricultural disruption, food insecurity, political instability (Persson et al., 2022)
Land-system Change	• Ecosystem change/destruction • New agricultural/industry activity	• Climate Change fueled events, famine, migration, conflict, agricultural disruption, food insecurity, political instability (United Nations, 2023; Kemp et al., 2022; Richards et al., 2021)
Biosphere Integrity (core)	• Ecosystem change/destruction	• Famine, migration, conflict, agricultural disruption, food insecurity, political instability, human population health (World Health Organization, 2023; Stockholm Resilience Centre, 2021; Pecl et al., 2017; Mace et al., 2014; Kemp et al., 2022; Richards et al., 2021)

The cascades into society presented in Table 2 have the potential to interact with one another to varying degrees. For example, ecosystem change and a reduction in biodiversity from transgressing the Biosphere Integrity boundary may result in agricultural disruption, which is further amplified by the transgression of the Freshwater Change Planetary Boundary. Furthermore, regional cascades into society, if there are

enough of high consequence could potentially cause other cascades to occur globally, i.e. conflict. In the event of sizeable impacts on society from the transgression of Planetary Boundaries, there will also likely be political instability (Beard et al., 2021). The cascades do not stop when a Planetary Boundary transgression effect first impacts society, there will likely be cascades throughout society and follow on effects. It is important to consider how these will impact the industrial and financial sectors, with the potential for supply chain issues, financial crises, and global/regional trade to be severely impacted. On the other hand, mitigating risks will require a concerted effort from all areas of society including the industrial, and financial sectors. Figure 2 summarizes the cascade process thus far.

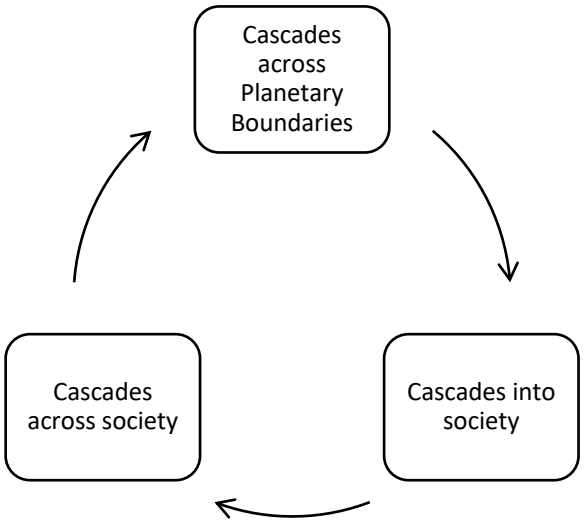


Figure 2: Cascades across Planetary Boundaries lead to cascades into society, which in turn lead to cascades across society. These then have the potential to create new cascades across the Planetary Boundaries.

4.2 Global Catastrophic Risk

Whilst humanity is resilient and adaptable, the complexity of modern society means that if several high impact events were to occur within a short time frame, then humanity could be subject to catastrophe (Beard et al., 2021). Section 4.1 highlighted the importance of several Planetary Boundaries in terms of what the cascading effects of transgressing their boundaries may be on society, in particular that of the Climate Change, Biosphere Integrity, and Land-system Change Planetary Boundaries, and agriculture and food security in society.

Research to date has included linking the Planetary Boundaries with Global Catastrophic Risk (Baum & Handoh, 2014), and other examinations of how the concept of Global Catastrophic Risk may apply to elements of the Planetary Boundaries, such as climate change and ecosystems (Beard et al., 2021; Cernev, 2022; Kareiva & Carranza, 2018). From the list of possible Global Catastrophic Risks provided in section 2.2 (Global Challenges Foundation, 2022), some are more relevant when considering the cascading effects through society of Planetary Boundaries that have been transgressed. Table 3 provides these Global Catastrophic Risks, and assesses whether their relevance. Ultimately, for cascading effects from the transgression of Planetary Boundaries to lead to civilizational collapse, there would have to be "a drastic reduction in human population, and the break-down of connections between surviving populations" (Turchin & Denkenberger, 2018). As identified earlier, it is assumed that a severe Global Catastrophic Risk or chain reaction

of Global Catastrophic Risk events could result in Existential Risk or an Existential Risk event occurring.

Given the large number of Global Catastrophic Risks that may eventuate from transgressed Planetary Boundaries and the cascading effects through society, the possibility of a chain reaction occurring is a concerning possibility. Such a chain reaction could potentially result in an Existential Risk event occurring, or reduce humanity's resilience significantly such that humanity is highly vulnerable to future risks.

Global Catastrophic Risk	**Relevant?**	**Comment**
Weapons of Mass Destruction	Highly Relevant	Earth system instability leading to political instability and conflict.
Pandemics	Highly Relevant	Migration of populations and emergence of new pathogens/viruses due to Earth system change. Erosion of wild spaces leading to the emergence of new pathogens/viruses.
Artificial Intelligence	Relevant	Accelerated developments in Artificial Intelligence in order to mitigate the worst societal impacts of Earth system change may have unforeseen consequences.
Near-Earth Asteroids	Relevant	Not relevant. However, it needs to be ensured that action is still taken to mitigate/minimise this risk in the event that preference is given to mitigating Earth system related Global Catastrophic Risks.
Climate Catastrophe	Highly Relevant	Climate Change is a core Planetary Boundary.
Supervolcanic Eruption	Relevant	Not relevant. However, it needs to be ensured that action is still taken to mitigate/minimise this risk in the event that preference is given to mitigating Earth system related Global Catastrophic Risks.
Ecological Collapse	Highly Relevant	Ecological Collapse is closely associated with Planetary Boundaries, including Biosphere Integrity.
Global Population Size	Relevant	The cascading effects of transgressing Planetary Boundaries could cause declines in global population in the event of conflict or major food insecurity and famine, this then increases the risk of societal collapse.
Climate Tipping Points	Highly Relevant	Planetary Boundaries are closely associated with tipping points, with increased risk of these being reached as Earth system boundaries are passed.

Table 3: Relevant Global Catastrophic Risks when considering the cascading effects of transgressed Planetary Boundaries through society, building on the information in Table 1, and Table 2 (Global Challenges Foundation, 2022; United Nations Office for Disaster Risk Reduction, 2022; Stockholm Resilience Centre, 2021; Richards et al., 2021; Kemp et al., 2022).

Even where it appears that a particular Global Catastrophic Risk may not be as relevant for the cascading effects through society of transgressed Planetary Boundaries, it should still be considered in order to ensure that resources are still provided to mitigate the risk, and not solely deployed for other risks. Furthermore, consideration should also be given to the emergence of possible future Global Catastrophic Risks. If the environmental situation continues to deteriorate as Planetary Boundaries are further transgressed, and societal impacts increase, then it is possible that new technologies, with or without Artificial Intelligence, will be developed to help combat environmental issues. These new technologies may be Earth or space based. The risks of such technologies are unknown, and they could potentially pose Global Catastrophic Risks or Existential Risks if either something were to go wrong, or they were to fail during operation, meaning that they no longer provide a solution to the problem they were developed to solve. Consideration should be given as to how an existing Global Catastrophic Risk (or a new risk) could capture the risks posed by such a new technology.

Global Catastrophic Risks stemming from the cascades of transgressed Planetary Boundaries can be considered in two groups: *primary* are those that stem directly as a result of transgressed boundaries, and *secondary*, which are those that arise due to cascades. *Primary* Global Catastrophic Risks include: Climate Catastrophe, Ecological Collapse, and Climate Tipping Points. *Secondary* include: Weapons of Mass Destruction, Pandemics, and Artificial Intelligence.

5. Framework generation

From the previous sections, there are extensive cascades between Planetary Boundaries, from Planetary Boundaries into society, and within society, which has the potential to increase the risk of different Global Catastrophic Risks, and potentially create some new ones. To aid in developing a better understanding of the cascading and reinforcing risks, effects, and feedback loops that may exist, and to help guide policy development and implementation, a framework is generated, refer Figure 3.

The framework presented in Figure 3 aims to provide a methodology to better understand 1) the possible interactions that exist between Planetary Boundaries, 2) the cascades that exist from the Planetary Boundaries into society, and 3) the Global Catastrophic Risks that may eventuate. Then from this understanding, to then leverage or identify critical points that exist in the system and develop and implement solutions accordingly. Since transgressed planetary boundaries are a global issue, the involvement of industry and governments to successfully implement solutions and ensure a safe future for humanity is essential, and the progress of industry and governments should be reviewed and held to account (i.e., through international organizations such as the World Trade Organization (WTO)).

The framework is aimed at an organization such as the United Nations (UN) and would be governed by an office such as the United Nations Environment Program (UNEP), the United Office for Disaster Risk Reduction (UNDRR), or potentially requires the creation of a new office to sit under the World Meteorological Organization (WMO), the purpose of which would be to monitor, mitigate, and respond to cascading environmental (Planetary Boundary) risks that have far reaching societal impacts, and Global Catastrophic Risk considerations. The UN office responsible for the overseeing and implementation of the framework, for building the industry partnerships, and for developing a means to hold industry and governments to account through partnerships with organisations such as the WTO or European Central Bank (ECB). The cost of such an endeavour largely depends on the method of implementation. Incorporation into an existing UN office is likely to have a smaller cost compared to that if a new office is set up.

Importantly, the framework prioritises involving industry and national governments, and developing a means to hold both industry and governments to account. Whilst the framework would be governed an overseen by an organization such as the UN, it involves industry and government such that the issue of when policy is generated by international organisations but not necessarily followed or adopted by industry and governments is addressed. The combination of involving industry and governments, and then having a means to hold them too account seeks to address this issue. Ultimately this is a preventative framework, in that it aims to minimise and mitigate possible risk rather than react to it should it eventuate. In this framework, industry is considered to consist of, but is not limited to, the manufacturing, resources, and financial sectors, and governments are considered to be national governments.

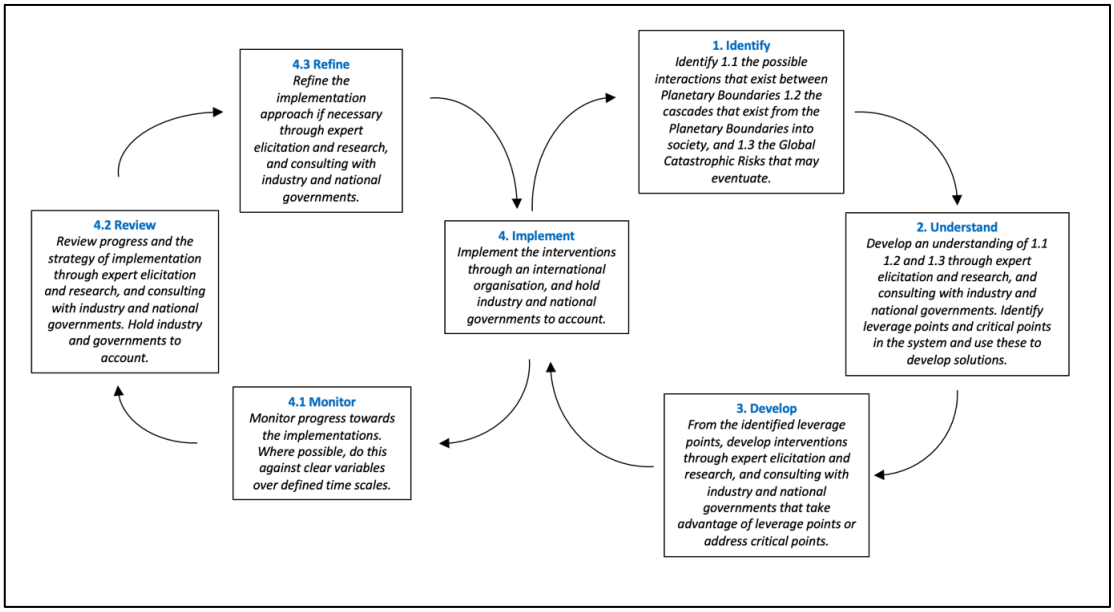


Figure 3: A framework to identify, categorise, develop, and implement solutions to the cascading effects of transgressed Planetary Boundaries, in particular the Global Catastrophic Risk consequences that may emerge.

As identified, the framework prioritises holding industry and national governments to account in the implementation phase. However, how to hold to hold these institutions to account is a difficult question, and whilst it would likely take form through an international organization like the United Nations or World Trade Organisation, it requires further investigation. Furthermore, this framework presents a proactive methodology to minimise and mitigate future risks rather than to necessarily respond to them. Consideration for the necessary reactionary solutions needs to be undertaken and solutions developed.

6. Conclusion

This paper has investigated the possible interactions between the Planetary Boundaries, the cascades that could flow from transgressed Planetary Boundaries into society, and the Global Catastrophic Risk that may exist as a result. It was found that transgressed Planetary Boundaries extensively cascade onto other Planetary Boundaries, in particular those of: Climate Change, Biosphere Integrity, and Land-system Change. Numerous cascades from transgressed Planetary Boundaries into and throughout society have been identified, with the most prominent being potential impacts to agriculture and food security.

These cascades into society have been found to present significant Global Catastrophic Risks, with potential for transgressed Planetary Boundaries to elevate existing risks including those associated with weapons of mass destruction, pandemics. To present a way forward, the developed framework provides a means for industry and national governments to work with the United Nations to identify, categorise, develop, and implement solutions to the cascading effects of transgressed Planetary Boundaries. The involvement of both industry and national governments is essential both due to the global nature of the Planetary Boundaries, and the far reaching cascades through society that have been identified in this paper.

Future work is required to further understand the cascades through society, and refining and development of the presented framework such that the cascading effects of transgressed Planetary Boundaries can be mitigated.

References

Avin, S., Wintle, B.C., Weitzdörfer, J., Ó hÉigeartaigh, S.S., Sutherland, W.J., & Rees, M.J. (2018) 'Classifying global catastrophic risks', *Futures*, *102*, 20-26. Available at: https://doi.org/10.1016/j.futures.2018.02.001 (Accessed: 10 February 2023).

Baum, S.D., & Handoh, I.C. (2014) 'Integrating the planetary boundaries and global catastrophic risk paradigms', *Ecological Economics*, *107*, 13-21. Available at: https://doi.org/10.1016/j.ecolecon.2014.07.024 (Accessed: 10 February 2023).

Beard, S. J., Holt, L., Tzachor, A., Kemp, L., Avin, S., Torres, P., & Belfield, H. (2021) 'Assessing climate change's contribution to global catastrophic risk', *Futures*, 127. Available at: https://doi.org/10.1016/j.futures.2020.102673 (Accessed: 12 February 2023).

Bostrom, N., & Cirkovic, M. (2008) *Global catastrophic risks*. Oxford: Oxford University Press.

Cernev, T. (2022) 'Global sustainability targets: Planetary boundary, global catastrophic risk, and disaster risk reduction considerations', *Progress in Disaster Science*, 16, 100264. Available at: https://doi.org/10.1016/j.pdisas.2022.100264 (Accessed: 15 February 2023).

Galaz, V., Biermann, F., Crona, B., Loorbach, D., Folke, C., Olsson, P., Nilsson, M., Allouche, J., Persson, Å., & Reischl, G., (2012) 'Planetary boundaries'—exploring the challenges for global environmental governance', *Current Opinion in Environmental Sustainability*, *4*(1), 80-87. Available at: https://doi.org/10.1016/j.cosust.2012.01.006 (Accessed: 11 February 2023).

Global Challenges Foundation. (2022) *Global Catastrophic Risks 2022: a year of colliding consequences*. Available at: https://globalchallenges.org/wp-content/uploads/2022/12/GCF_Annual_Report_2022.pdf (Accessed: 25 February 2023).

Kareiva, P., & Carranza, V. (2018) 'Existential risk due to ecosystem collapse: Nature strikes back', *Futures*, *102*, 39-50. Available at: https://doi.org/10.1016/j.futures.2018.01.001 (Accessed: 1 March 2023).

Kemp, L., Xu, C., Depledge, J., Ebi, K.L., Gibbins, G., Kohler, T.A., Rockström, J., Scheffer, M., Schellnhuber, H.J., Steffen, W., & Lenton, T.M. (2022) 'Climate Endgame: Exploring catastrophic climate change scenarios', *Proceedings of the National Academy of Sciences*, *119*(34), p.e2108146119. Available at: https://doi.org/10.1073/pnas.2108146119 (Accessed: 15 February 2023).

Lade, S.J., Steffen, W., De Vries, W., Carpenter, S.R., Donges, J.F., Gerten, D., Hoff, H., Newbold, T., Richardson, K., & Rockström, J. (2020) 'Human impacts on planetary boundaries amplified by Earth system interactions', *Nature sustainability*, *3*(2), 119-128. Available at: https://doi.org/10.1038/s41893-019-0454-4 (Accessed: 10 February 2023).

Mace, G. M., Reyers, B., Alkemade, R., Biggs, R., Stuarrt Chapin III, F., Cornell, S. E., Díaz, S., Jennings, S., Leadley, P., Mumby, P. J., Purvis, A., Scholes, R. J., Seddon, A. W. R., Solan, M., Steffen, W., & Woodward, G. (2014) 'Approaches to defining a planetary boundary for biodiversity', *Global Environmental Change*, 28, 289-297.

Available at: http://dx.doi.org/10.1016/j.gloenvcha.2014.07.009 (Accessed: 1 June 2023).

Nash, K.L., Cvitanovic, C., Fulton, E.A., Halpern, B.S., Milner-Gulland, E.J., Watson, R.A., & Blanchard, J.L. (2017) 'Planetary boundaries for a blue planet', *Nature ecology & evolution*, *1*(11), 1625-1634. Available at: https://doi.org/10.1038/s41559-017-0319-z (Accessed: 10 February 2023).

Ord, T. (2020) *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury.

Pecl, G. T., Araújo, M. B., Bell, J. D., Blanchard, J., Bonebrake, T. C., Chen, I., Clark, T. D., Colwell, R. K., Danielsen, F., Evengård, B., Falconi, L., Ferrier, S., Frusher, S., Garcia, R. A., Griffis, R. B., Hobday, A. J., Janion-Scheepers, C., Jarzyna, M. A., Jennings, S.,… Williams, S. E. (2017) 'Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being'*Science*, 355(6332). Available at: https://doi.org/10.1126/science.aai9214 (Accessed: 1 June 2023).

Persson, L., Carney Almroth, B.M., Collins, C.D., Cornell, S., de Wit, C.A., Diamond, M.L., Fantke, P., Hassellöv, M., MacLeod, M., Ryberg, M.W., & Søgaard Jørgensen, P. (2022) 'Outside the safe operating space of the planetary boundary for novel entities', *Environmental science & technology*, *56*(3), 1510-1521. Available at: https://doi.org/10.1021/acs.est.1c04158 (Accessed: 1 March 2023).

Pescaroli, G., & Alexander, D. (2018) 'Understanding Compound, Interconnected, Interacting, and Cascading Risks: A Holistic Framework', *Risk Analysis*, 38(11), 2245-2257. Available at: https://doi.org/10.1111/risa.13128 (Accessed: 2 June 2023).

Richards, C.E., Lupton, R.C., & Allwood, J.M. (2021) 'Re-framing the threat of global warming: an empirical causal loop diagram of climate change, food insecurity and societal collapse', *Climatic Change*, *164*(3-4), 49. Available at: https://doi.org/10.1007/s10584-021-02957-w (Accessed: 20 February 2023).

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S., Lambin, E., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C. A., Hughes, T., van der Leeuw, S., Roghe, H., Sörlin, S., Snyder, P. K., Costanza, R., Svedin, U.,… Foley, J. (2009) 'Planetary Boundaries: Exploring the Safe Operating Space for Humanity', Ecology and Society, 14(2), 302-333. Available at: https://www.jstor.org/stable/26268316 (Accessed: 10 February 2023).

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R., Carpenter, S. R., de Vries, W., de Wit, C., A., Folke, C., Gerten, D., Heinke, J., Mace, G. M., Persson, L. M., Ramanathan, V., Reyers, B., & Sörlin, S. (2015) 'Planetary boundaries: Guiding human development on a changing planet', *Science*, **347**(6223), 1-10. Available at: https://doi.org/10.1126/science.1259855 (Accessed: 10 February 2023).

Stockholm Resilience Centre. (2021) *The nine planetary boundaries*. Available at: https://www.stockholmresilience.org/research/planetary-boundaries/the-nine-planetary-boundaries.html (Accessed: 13 November 2021).

Turchin, A., & Denkenberger, D. (2018) 'Global catastrophic and existential risks communication scale', *Futures*, *102*, pp.27-38. Available at: https://doi.org/10.1016/j.futures.2018.01.003 (Accessed: 5 March 2023).

UNESCO World Water Assessment Programme. (2023) *The United Nations world water development report 2021: partnerships and cooperation for water*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000375724 (Accessed: 1 June 2023).

UNESCO World Water Assessment Programme. (2021) *The United Nations world water development report 2021: valuing water.* Available at: https://unesdoc.unesco.org/ark:/48223/pf0000375724 (Accessed: 1 June 2023).

United Nations. (2023) *15 Life on Land*. Available at: https://www.un.org/sustainabledevelopment/biodiversity/ (Accessed: 1 June 2023).

United Nations Office for Disaster Risk Reduction. (2022) *Global Assessment Report on Disaster Risk Reduction 2022: Our World at Risk: Transforming Governance for a Resilient Future*. Available at: https://www.undrr.org/media/79595 (Accessed: 10 March 2023).

Wang-Erlandsson, L., Tobian, A., van der Ent, R.J., Fetzer, I., te Wierik, S., Porkka, M., Staal, A., Jaramillo, F., Dahlmann, H., Singh, C., & Greve, P. (2022) 'A planetary boundary for green water', *Nature Reviews Earth & Environment*, *3*(6), 380-392. Available at: https://doi.org/10.1038/s43017-022-00287-8 (Accessed: 15 February 2023).

World Health Organization. (2023) *Biodiversity and Health*. Available at: https://www.who.int/news-room/fact-sheets/detail/biodiversity-andhealth#:~:text=Biodiversity%20loss%20can%20have%20significant, cause%20or%20exacerbate%20political%20conflict. (Accessed: 1 June 2023)

Stanford
Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Anthropocene Under Dark Skies: The Compounding Effects of Nuclear Winter and Overstepped Planetary Boundaries

Florian Ulrich Jehn [1, 2*]

Citation: Florian Ulrich, Jehn. Anthropocene Under Dark Skies: The Compounding Effects of Nuclear Winter and Overstepped Planetary Boundaries. *Proceedings of the Stanford Existential Risks Conference* 2023, 119-132. https://doi.org/10.25740/zb109mz2513

Academic Editor: Paul Edwards, Trond Undheim, Dan Zimmer

Funding: This work was funded by the Alliance to Feed the Earth in Disasters (ALLFED).

Conflict of Interest Statement: There are no conflicts of interest to disclose.

Informed Consent Statement: Given the nature of the project, informed consent was not required.

Acknowledgments: I would like to thank Ekaterina Ilin, Luke Kemp, Aron Mill, Paul Edwards, Juan García Martínez and David Denkenberger for valuable comments which improved this manuscript considerably.

Author Contributions: The author confirms sole authorship and contribution to all aspects of the work.

Abstract: The analysis of global catastrophic events often occurs in isolation, simplifying their study. In reality, risks cascade and interact. Therefore, it is essential to consider the interconnected nature of global risks. This investigation explores the interplay between nuclear winter and planetary boundaries. It may seem reasonable to assume that respecting planetary boundaries, which define a safe operating space for the planet, is preferable before a nuclear war. However, that does not always seem to be the case. For instance, increased nitrogen emissions today could serve as a nutrient buffer during nuclear winter. Contrastingly, mitigating climate change, means an even larger temperature drop in nuclear winter in comparison with pre-industrial times. This exploratory study also highlights planetary boundaries that could enhance human survival if we adhere to their limits, both presently and after a nuclear war. The best example being biosphere integrity, as conserving it has no direct downsides and would make the Earth system more resilient to resist the shock of a nuclear winter.

Keywords: Anthropocene, existential risk, food security, nuclear winter, planetary boundaries

1. Alliance to Feed the Earth in Disasters (ALLFED), USA
2. Justus-Liebig-University Gießen, Germany

* Correspondence: florian@allfed.info

1. Consequences of a nuclear war

Imagine a future after a full-scale nuclear war. An average person's life would dramatically change overnight. Many major cities could go up blazing in a firestorm, delivering large quantities of soot into the upper atmosphere (Coupe et al., 2019; Tarshish & Romps, 2022) and killing millions (Habbick, 1983). This would change the climate globally (Coupe et al., 2019). While there would be regions like Australia or New Zealand (Boyd & Wilson, 2022) which would still have bearable temperatures, other places like Eastern Europe or Canada would remain frozen for years (Coupe et al., 2019). Under these circumstances, billions of people might starve (Xia et al., 2022).

But it does not have to be this way. Nuclear winter would affect everyone, but the biggest impact would be felt in many of the world's richest countries. The United States and Central Europe would be devastated, both by the direct impact of the nuclear weapons and the indirect effects of the changing climate (Coupe et al., 2019). This gives a strong incentive for those nations to prepare and they have the resources to do so.

Now imagine a different future. A future where humanity is prepared for the worst case. While there are technical solutions, which allow us to scale up resilient food sources like single cell proteins from natural gas (García Martínez et al., 2022) or seaweed (Jehn et al., 2023), many of the problems we would have are linked to the way we are currently overusing the resources of our planet (Steffen et al., 2015). For instance, if we can avoid the overuse of fisheries through regulations now, humanity would be left with more food in a nuclear winter (Scherrer et al., 2020). If we limit our footprint on the planet now, we would have more resources to cope with catastrophes.

It is likely that fisheries are not the only part where being more modest in our resource use today, would allow us extra resources in worst case scenarios. Many of the Earth's systems are under considerable strain (Steffen et al., 2015). Relieving this strain would allow humanity more leeway during catastrophic events. This study explores the interactions between nuclear winter and planetary boundaries to identify which boundaries we should focus on from an existential risks perspective. Nuclear winter can be seen as standing in here for other abrupt sunlight reduction scenarios (ASRSs) such as impact winter or volcanic winter, which refer to sun blocking due to asteroid/comet (bolide) impacts or large volcanic eruptions respectively. While there are differences between those three events, they are likely similar enough to also have comparable interactions with planetary boundaries.

2. Connecting planetary boundaries and nuclear winter

Planetary boundaries are a framework to evaluate the carrying capacity of the Earth System (Rockström et al., 2009; Steffen et al., 2015). They highlight the parts of the earth system which ensure the habitability of Earth and how much strain they are under. This has shown that many important parts of the Earth System may be in a dangerous condition. Only three of the eight currently quantified planetary boundaries are in their safe operating space, which means in the state they had in the Holocene (last 12,0000 years). (Persson et al., 2022; Steffen et al., 2015). Especially, biodiversity and biogeochemical flows are beyond their safe limits (Steffen et al., 2015). This means that they are taxed beyond their capacity and will degrade over time. The more those planetary boundaries are overstepped, the more strain would be put on the Earth's systems that allow humanity to exist. Agriculture in particular would be significantly impacted due to its reliance on boundaries such as freshwater, climate, and phosphorus and nitrogen cycles.

Agriculture would also be massively impacted by nuclear winter (Xia et al., 2022) or other ASRSs. Those are caused by particles in the upper atmosphere blocking out sunlight. This can happen via bolide impact (Tabor et al., 2020), high-magnitude volcanic eruptions (Luterbacher & Pfister, 2015; Rougier et al., 2018), and nuclear war (Coupe et al., 2019; Turco et al., 1983). Given the lower rate of volcanic eruptions and bolide impact, nuclear war is the most likely candidate to lead to such a scenario. However, recent research also shows that volcanic eruptions might be more dangerous and likely than previously thought (Cassidy & Mani, 2022; Mani et al., 2021). Particles in the upper atmosphere would block incoming solar radiation, which would result in considerably lower temperatures and thus lower precipitation. This in turn would significantly decrease food production and make the current global system unviable. Recent research has highlighted that this could lead to global famine (Xia et al., 2022), though this could possibly be counteracted by implementation of resilient foods (Rivers et al., 2022) like sugar from fiber (Throup et al., 2022) single cell proteins from hydrogen (García Martínez et al., 2021), or leaf protein concentrate (Pearce et al., 2019). Still, it is very likely that a nuclear winter would bring a considerable strain on global food production.

Nuclear winter and planetary boundaries work on different time horizons. Overstepping planetary boundaries is a decadal-scale process that gets incrementally worse (Steffen et al., 2015). Nuclear winter on the other hand is sudden and devastating in comparison (Coupe et al., 2019). However, exploring their interaction is still valuable, as their difference in speed does not mean they cannot interact with each other. It merely means that every interaction identified, would get better or worse depending on how much humanity is able to stay clear of overstepping the planetary boundaries.

All this highlights that the main interaction of nuclear winter and planetary boundaries would most likely happen through agriculture. This fits into the classification of global catastrophic risks of Avin et al. (Avin et al., 2018), as this has also highlighted the food system as one of the elements of human society that is most at risk of global catastrophic events. Therefore, we need additional research that looks into possible problems in this area.

3. Other research looking into planetary boundaries and existential risks more broadly

I am not aware of any literature that is specifically looking into the interactions of nuclear winter and planetary boundaries. This is likely due to the fact that the existential risk studies field is relatively small, and has only really started in the last decade (Ord, 2020). Due to its novelty it also is somewhat separated from the traditional science around global problems, like planetary boundaries. In addition, planetary boundaries are still a relatively new concept as well (starting in 2009 (Rockström et al., 2009)). Nuclear winter has been known as a problem since the 1980s (Turco et al., 1983), but did not get much public attention between the end of the Cold War and the invasion of Ukraine. Still, there is some research that is already exploring ideas with a similar spin like this study here:

- Savitch et al. looked into how likely it is that exo-civilizations are creating their own version of an Anthropocene and use simple models to find interactions between civilizations and their planet. Those models might be adaptable to planetary boundaries (Savitch et al., 2021).
- Geoengineering and termination shock in nuclear winter, are hinted at in Tang and Kemp (Tang & Kemp, 2021).
- Kemp et al in their climate endgame paper briefly touch on interactions of climate change and nuclear war (Kemp et al., 2022).

- Thomas Cernev has done research on global catastrophic risk and planetary boundaries in general, but it is more abstract than the direct comparison made here (Cernev, 2022).
- Scherrer et al. have shown that if we make sure to not overuse natural resources (fisheries as the example in their study), the planet would have a bigger buffer to use up during a nuclear winter (Scherrer et al., 2020).
- Baum and Handoh established a framework (Baum & Handoh, 2014) that tried to combine global catastrophic risks and planetary boundaries, but it seems like this has not been built upon in recent years.

4. Interactions

4.1 Biosphere integrity

Biosphere integrity refers to the idea that changes in biodiversity both locally and globally can have significant impacts on the functioning of the Earth system (Steffen et al., 2015). These functions are important to humanity, as they offer ecosystem services like the cleaning of water or the pollination of plants. These services can only be maintained if enough of our environment can remain undisturbed (Mohamed et al., 2022). In the context of planetary boundaries this concept is split into functional and genetic diversity (Steffen et al., 2015). Functional diversity refers to the idea of how much the composition of the biosphere has changed since before the industrial revolution and genetic diversity to the totality of the genetic diversity between all species and individuals. It remains unclear how much biosphere integrity is already damaged by human influence. However, it seems likely that every reduction in functional and genetic diversity is likely to be detrimental to the ability of the biosphere to cope with nuclear winter, as increased biodiversity likely makes ecosystems more resilient to climate extremes (De Boeck et al., 2018). Nuclear winter would have an outsized impact on the global biosphere. The biosphere has survived a number of very large volcanic eruptions (e.g. the Toba eruption (Chesner et al., 1991)), which can also lead to a volcanic winter (Rampino, 2002). However, the mechanisms of volcanic winters and nuclear winters are different. Volcanic winters are mainly caused by sulfates (Luterbacher & Pfister, 2015), while nuclear winters are caused by soot (Coupe et al., 2019). This difference likely makes nuclear winters longer lasting (up to ten years) and therefore introduces a new challenge for the biosphere. The higher the biosphere's integrity, the greater its ability to recover following a long nuclear winter. Mitigating the impact of nuclear winter on humans by reducing starvation could spare some species that would otherwise be eaten by desperate humans or be unaffordable to save in zoos (Denkenberger & Pearce, 2015).

4.2 Climate Change

Climate change and nuclear winter can be seen as two sides of the same coin. Both are climatic changes driven by human actions, one making the planet too hot, the other making it too cold (Pittock, 1988). They are even simulated using the same models, like the Community Earth System Model (CESM) (Coupe et al., 2019; Kay et al., 2015). Current predictions estimate an average warming between 2.1 and 3.9°C by 2100 due to climate change (Liu & Raftery, 2021), while a nuclear winter caused by an all out nuclear war is estimated to cause a peak temperature drop of about 9 °C (Coupe et al., 2019). This means even a largely out of control climate change, would not be enough to counteract the whole cooling effect of a nuclear winter. Still, global warming could dampen some of the effects of a nuclear winter. However, the crops would likely be optimized (either through location or genetic control) to the warmer climate (Minoli et al., 2022), so a sudden temperature reduction would likely still be catastrophic. And this should not be seen as an argument that we should care less about climate change, as it might make us safer

against another catastrophic event. The climate system is immensely complex and has many complex feedback loops and tipping points (Armstrong McKay et al., 2022), and we have only limited research on higher temperatures (Jehn et al., 2021, 2022). Also, there simply is no research which looks at how exactly climate change and nuclear winter might interact. Still, we know that nuclear winter would likely influence large climatic patterns like El Niño (Coupe et al., 2021), whose fluctuations are already getting more intense and frequent due to climate change (Cai et al., 2021). Therefore, even though global warming might mitigate the cooling effect of nuclear winter, betting on climate change to solve nuclear winter would be a very risky proposal with unforeseeable consequences. In addition, restoration after a nuclear winter is likely harder if this has to happen in a world under pressure of strong global warming and a world ravaged by climate change has likely a higher probability of nuclear war to start with. Finally, there is a chance that a nuclear winter might push the Earth system in a new equilibria, lasting for hundreds of years. This has happened after large volcanic eruptions in the past (Newhall et al., 2018). As the effects of nuclear winter and volcanic winter are likely somewhat similar (Newhall et al., 2018; Özdoğan et al., 2013), this implies that a longer term shift might also be triggered by a nuclear winter and at least for the ocean system, there are modeling results that show that a longer term shift could happen after a nuclear war (Harrison et al., 2022).

4.3 Novel Entities

The term novel entities refers to the pollution of the environment with man made chemicals, which cause detrimental effects to humans and the environment (Steffen et al., 2015). A well-known example here is the usage of DDT in the 20th century, which almost led to the extinction of several species of birds of prey. As there is no background rate for such emissions, the planetary boundary for novel entities is defined as overstepped if globally more is produced than can be monitored, which is currently the case (Persson et al., 2022). The effects of most of the novel entities are chronic (Persson et al., 2022). This means that they would be detrimental to health during a nuclear winter as well, but not more so than they would have been otherwise. However, nuclear war itself would introduce additional novel entities into the environment, mainly in the form of fallout (Smith & Smith, 1981) and the toxic chemicals produced by fires (Alarie, 2002). In addition, toxic chemicals could be created and distributed through fires and explosions in industrial facilities. Therefore, this would push concentrations further outside of the safe operating space. Still, due to the different nature of emission before and during a nuclear war, it is unclear how much it would help in nuclear winter to stay below this boundary now. Novel entities could be seen as an additional stress factor, not a major disruption in and of itself.

4.4 Stratospheric ozone depletion

The ozone layer protects the Earth's surface from ultraviolet radiation. It was damaged by the release of ozone depleting substances (for example chlorofluorocarbons). After their ban by the Montreal Protocol the ozone layer started to regenerate and is now mostly intact again (Barnes et al., 2021; Rockström et al., 2009). This leaves ozone depletion as one of the few planetary boundaries which is currently in the safe operating space. However, this would change significantly after a nuclear war. Even the earliest nuclear winter research hypothesized that the ozone layer would be negatively impacted (Turco et al., 1983) and recent research has estimated that the ozone losses would be rapid and global average losses could be as high as 75 % (Bardeen et al., 2021). The same effect, but to a lesser extent, has also been found in simulations for smaller, regional nuclear wars (Mills et al., 2008). The main mechanism is reactions with nitrogen oxides, smoke and the general heating of the upper atmosphere (Bardeen et al., 2021). In the first few years the soot in the atmosphere would shield the surface from most of the incoming ultraviolet radiation.

However, at the same time as soot is cleared from the atmosphere, the ultraviolet radiation rises and could reach UV index values of 35-45 (Bardeen et al., 2021) (not going outside is recommended for UV index > 11). It is estimated to take 12-15 years to return to pre-war UV radiation levels (Bardeen et al., 2021). This means that it is important that we manage to keep the ozone layer intact, to not add to the potentially devastating effect of the nuclear war. However, the effect of nuclear war on the ozone layer could be in a different order of magnitude than problems with the ozone layer so far. This also shows that nuclear war would disrupt one of the few planetary boundaries we are currently managing to keep in safe operating space.

4.5 Atmospheric aerosol loading

This boundary is concerned with the totality of aerosols and their influence on human health and wellbeing. The aerosols also influence solar radiation by scattering it and hydrological cycles by altering cloud formation (Rockström et al., 2009). Both are important for nuclear winter. The main mechanism that could drive nuclear winter is the emission of soot by firestorms (Coupe et al., 2019). Those emissions would contribute significantly to the atmospheric aerosol loading. An all-out nuclear war may emit around 150 Tg of soot in a day to a week (Coupe et al., 2019), while the present-day global soot emissions per year are only around 4-22 Tg (Bond et al., 2004). It is not yet determined whether the planetary boundary for aerosol loading is overstepped now (Steffen et al., 2015). However, there is evidence that the scattering of incoming solar radiation cools the Earth today by a small amount (Bellouin et al., 2020). We also have further evidence for this cooling effect of atmospheric aerosol loading, as the decrease in sulfur content for ship fuel changed the forcing by ship emissions (Yuan et al., 2022). Therefore, removing aerosols now would result in an overall warmer planet, which in turn would not cool as much due to nuclear winter. This raises the same problems as the interaction between climate change and nuclear winter (section 4.2): Is it better to have a warmer planet now, to also have a warmer planet during nuclear winter?

4.6 Ocean acidification

Oceans absorb carbon dioxide as a part of the global carbon cycle. The level of carbon dioxide dissolved in the upper ocean is in equilibrium with the atmosphere and depends strongly on the temperature of the water. As the levels of carbon dioxide rise in the atmosphere, so does the amount of carbon dioxide in the oceans. This in turn decreases the pH in the water. The largest effect of this is the disruption of the life cycles of all organisms who build shells from calcium carbonate. In addition, there is evidence that ocean acidification influences the availability of carbon, nitrogen and phosphorus in the oceans, with unclear effects on the ecosystem (Doney et al., 2009). Since the beginning of the industrial revolution this has led to a drop of around 0.1 in the global average of ocean pH (Intergovernmental Panel on Climate Change, 2014). The most direct impact for humans would be the continuous decrease in the amount of catchable fish in the oceans, as the ecosystems get more and more out of balance and decline in productivity (Cooley & Doney, 2009).

Nuclear winter is predicted to increase the global ocean pH by about 0.05. The effect would mainly be driven by the decrease in sea surface temperature, which shifts the carbonate equilibrium in the water (Lovenduski et al., 2020). While this might seem like a positive effect, modeling results show that it would rather worsen the problem. Marine species would have to adapt to a sharp increase in pH that would only take around a year to shift. However, as the ocean heats up again, as the soot in the atmosphere clears, the pH drops to its previous level, or even lower due to the killed plant matter decomposing. Such a rapid change in ocean chemistry would put a considerable strain on marine

ecosystems. In addition, the cooling ocean during nuclear winter can dissolve more carbon dioxide, which in turn decreases the availability of carbonate even further (Lovenduski et al., 2020), which means that the increase in pH does not help shell building organisms.

Overall, the interactions between ocean acidification and nuclear winter would likely be negative. This implies that it is important to slow down ocean acidification now to leave ecosystems more room to adapt during a nuclear winter. This would also increase food availability today and after a nuclear war.

4.7 Biogeochemical flows

Biochemical flows mainly refer to the flows of nitrogen and phosphorus in the environment as two of the main nutrients for plants (Leinfelder et al., 2017). They are summarized under biogeochemical flows, as they are tightly connected. While both nitrogen and phosphorus are needed to sustain any ecosystem, they start to disrupt them as well once their levels change due to anthropogenic emissions (Rockström et al., 2009; Steffen et al., 2015). The main negative effects for both phosphorus and nitrogen are dead zones and shifts in species composition. Dead zones refer to parts of the ocean or other water bodies which have been depleted of oxygen, after eutrophication shifted their species composition and abundance (e.g. algae blooms) (Schindler & Vallentyne, 2008). The main emission pathway for both nutrients are fertilizers, which have been overapplied for decades, especially in major food production countries like Germany (Steffen et al., 2015).

There is no direct way that nuclear war would change biogeochemical flows. Still, there are possible interactions that have to be taken into account. Nuclear winter disrupts agriculture as it is practiced today by shifting climate zones globally and thus making agriculture very difficult if no adaptations are made (Xia et al., 2022). There are possibilities that allow us to still produce food, but those are under the assumption that enough nutrients remain available (Rivers et al., 2022). This leads to the counterintuitive conclusion that overstepping the biogeochemical boundary now, might make humanity more resilient to nuclear winter, as more nutrients are available without needing additional fertilizer, which are likely hard to come by after a nuclear war. Around half of currently used fertilizers are synthetic and any stress on energy and supply chains would be felt. This does not mean that the nutrients available in the environment would allow production levels of today, but they would add a buffer, which would give additional time to set up production and trade for fertilizer in a post nuclear war world. Greater fertilizer production now would also mean larger amounts in storage, which would be helpful in a catastrophe (Mörsdorf, 2021).

4.8 Freshwater use

This boundary is concerned with the influence of humans on the global water cycle. It is in the safe operating space when there is still enough water to sustain ecosystem services (Rockström et al., 2009). Currently this seems to be the case and the freshwater use planetary boundary is largely intact. However, future predicted water usage might bring it closer to its capacity (Rockström et al., 2009).

Nuclear winter generally leads to less evapotranspiration and thus less precipitation (Coupe et al., 2019). Therefore, the overall availability of water would decline, which means that full water storages now would give an additional buffer during nuclear winter. It is unclear how water usage would develop during nuclear winter. However, it might decline, as agriculture is one of the main water users and conventional agriculture

would not be possible anymore in many places (Xia et al., 2022). However, it could also be helpful for nuclear winter to have used more water now, as this implies a larger water infrastructure, which could be helpful to allow a better water distribution. Overall, freshwater use now has likely not a very large impact on nuclear winter either way, though both positive and negative impacts are possible.

4.9 Land-system change

Land system change is driven mainly by the expansion of agriculture and the conversion of forests and grasslands to agricultural land (Rockström et al., 2009). This threatens biodiversity and affects both the climate system in general and the hydrological cycle in particular. However, in relation to nuclear winter this boundary could be of lower importance. While deforestation leads to fewer biomass available in nuclear winter, the global amount of trees is so large that this likely remains not an issue (Denkenberger & Pearce, 2015). Also, there might be a positive effect of clearing more land now, which would be also available in nuclear winter. The other way around could be more important though. Nuclear winter would need a major shift in the way we produce food, which also includes relocating crops to warmer regions. In addition, the temperature drop in nuclear winter increases the area needed for crop production (Rivers et al., 2022). Also, nuclear war might cause large scale forest fires, which would at least temporarily change the land use of the affected areas. Therefore, land-system change would likely be accelerated in a nuclear winter. Large parts of currently unused land might need to be converted to agriculture, for example for greenhouses (Alvarado et al., 2020). While those changes may be reverted once the climate returns to normal after a nuclear winter, this would still be a significant change in those systems, because they would need a considerable amount of time to be able to return to their pre-war state.

5. Discussion and conclusion

Planetary boundaries are defined to highlight how we should treat the Earth to make it habitable for the long term. The included assumption here is that staying in the safe operating space is always better. This study was a first exploration of how this assumption holds true when the planetary boundaries interact with existential risks. The insights gained here show that this assumption is often true, but not always. Overstepping planetary boundaries can either increase or decrease nuclear winter survivability, depending on which boundary has been broken (Figure 1). In addition, all boundaries are interconnected, and fixing one boundary may have unintended consequences for others.

Overstepping the boundary on climate change results in an increase in temperature, which in itself has negative effects on the Earth system. However, this increase in temperature also means that during a nuclear winter, the planet would be cooled down from an elevated level, ultimately resulting in a lower peak cooling. This interaction might seem positive, but it remains unclear if it could lead to unforeseen consequences. Therefore, it is highly uncertain if this effect of climate change could be positive.

Overstepping the boundary on biogeochemical flows however might provide humanity with a nutrient buffer if overstepped, but it also has clear downsides today, like dead zones in the oceans. Therefore, it is essential to balance the present needs of human society with the long-term risks and benefits associated with overstepping planetary boundaries.

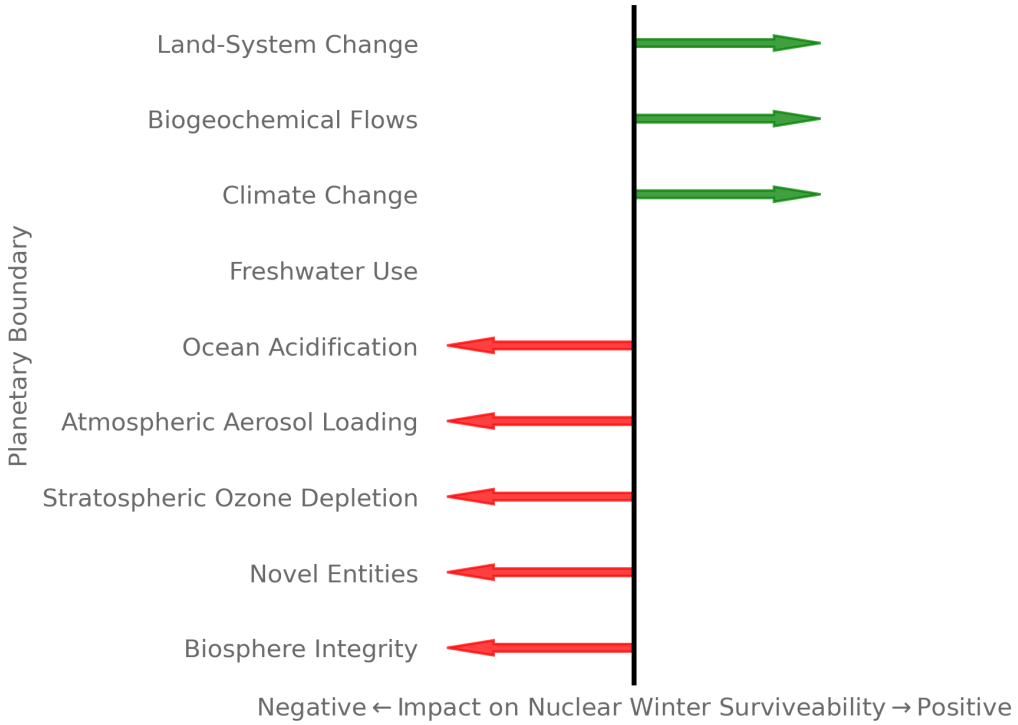


Figure 1: Visual summary and qualitative assessment of the impact of overstepping planetary boundaries on the chances of survival for humanity after a nuclear war.

On the other hand, certain planetary boundaries, if overstepped, likely have only a negative impact on nuclear winter survivability. Ocean acidification, for example, would be sensitive to the effects of a nuclear war and is already under stress, which diminishes global food production today and has been highlighted as a strong risk (Kareiva & Carranza, 2018). Therefore, stopping ocean acidification has clear upsides. However, it is also the case that planetary boundaries are interconnected, and ocean acidification is mainly caused by elevated carbon dioxide levels. Bringing those back to preindustrial levels would stop ocean acidification, but also remove the temperature buffer provided by climate change. All of those risks are connected and better results can be expected when their interactions and feedback loops are considered (Ward et al., 2022).

Changing the state of the earth relative to planetary boundaries would be an enormous undertaking. Therefore, directed existential risk reduction activities are likely more cost-effective. However, if mitigating global catastrophes could be used to nudge existing funding in this space towards work on planetary boundaries that would be most synergistic with global catastrophes, this may be promising.

These findings highlight the importance of identifying and staying withing boundaries that may provide upsides before and after a nuclear war. Stratospheric ozone depletion and biosphere integrity appear promising in this regard, as they could have a clear negative effect. But even here there are likely differences when it comes to costs and benefits. For example, the effect of nuclear winter on the ozone layer would be quite strong and likely dwarfs any reconstruction of the ozone layer now and biodiversity has more of a supporting role and its impact on human life are more indirect (Kareiva & Carranza, 2018). It is difficult to assess which planetary boundary should be given priority from a nuclear winter perspective. This problem gets even more difficult when we consider how boundaries might interact. For example, recent research has

highlighted that rising fertilizer prizes and thus lower fertilizer could increase the land area used for agriculture considerably (Alexander et al., 2023).

Given the tentative evidence presented here biosphere integrity could possibly be the planetary boundary with the highest net positive effect on nuclear winter survivability, albeit a diffuse one. Preserving biosphere integrity now is clearly positive, it does not have obvious, strong interactions with other boundaries and it would provide humanity with a more stable Earth system overall, both now and in the nuclear winter. Still, this paper here is just a first step in this direction and more research is needed, especially when it comes to interactions and feedback loops between the planetary boundaries themselves and nuclear winter.

References

Alarie, Y. (2002). Toxicity of Fire Smoke. *Critical Reviews in Toxicology*, 32(4), 259–289. https://doi.org/10.1080/20024091064246

Alexander, P., Arneth, A., Henry, R., Maire, J., Rabin, S., & Rounsevell, M. D. A. (2023). High energy and fertilizer prices are more damaging than food export curtailment from Ukraine and Russia for food prices, health and the environment. *Nature Food*, 4(1), Article 1. https://doi.org/10.1038/s43016-022-00659-9

Alvarado, K. A., Mill, A., Pearce, J. M., Vocaet, A., & Denkenberger, D. (2020). Scaling of greenhouse crop production in low sunlight scenarios. *Science of The Total Environment*, 707, 136012. https://doi.org/10.1016/j.scitotenv.2019.136012

Armstrong McKay, D. I., Staal, A., Abrams, J. F., Winkelmann, R., Sakschewski, B., Loriani, S., Fetzer, I., Cornell, S. E., Rockström, J., & Lenton, T. M. (2022). Exceeding 1.5°C global warming could trigger multiple climate tipping points. *Science*, 377(6611), eabn7950. https://doi.org/10.1126/science.abn7950

Avin, S., Wintle, B. C., Weitzdörfer, J., Ó hÉigeartaigh, S. S., Sutherland, W. J., & Rees, M. J. (2018). Classifying global catastrophic risks. *Futures*, 102, 20–26. https://doi.org/10.1016/j.futures.2018.02.001

Bardeen, C. G., Kinnison, D. E., Toon, O. B., Mills, M. J., Vitt, F., Xia, L., Jägermeyr, J., Lovenduski, N. S., Scherrer, K. J. N., Clyne, M., & Robock, A. (2021). Extreme Ozone Loss Following Nuclear War Results in Enhanced Surface Ultraviolet Radiation. *Journal of Geophysical Research: Atmospheres*, 126(18), e2021JD035079. https://doi.org/10.1029/2021JD035079

Barnes, P. W., Bornman, J. F., Pandey, K. K., Bernhard, G. H., Bais, A. F., Neale, R. E., Robson, T. M., Neale, P. J., Williamson, C. E., Zepp, R. G., Madronich, S., Wilson, S. R., Andrady, A. L., Heikkilä, A. M., & Robinson, S. A. (2021). The success of the Montreal Protocol in mitigating interactive effects of stratospheric ozone depletion and climate change on the environment. *Global Change Biology*, 27(22), 5681–5683. https://doi.org/10.1111/gcb.15841

Baum, S. 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

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., Boucher, O., Carslaw, K. S., Christensen, M., Daniau, A.-L., Dufresne, J.-L., Feingold, G., Fiedler, S., Forster, P., Gettelman, A., Haywood, J. M., Lohmann, U., Malavelle, F., Mauritsen, T., … Stevens, B. (2020). Bounding Global Aerosol Radiative Forcing of Climate Change. *Reviews of Geophysics*, 58(1), e2019RG000660. https://doi.org/10.1029/2019RG000660

Bond, T. C., Streets, D. G., Yarber, K. F., Nelson, S. M., Woo, J.-H., & Klimont, Z. (2004). A technology-based global inventory of black and organic carbon emissions from combustion. *Journal of Geophysical Research: Atmospheres*, 109(D14). https://doi.org/10.1029/2003JD003697

Boyd, M., & Wilson, N. (2022). Island refuges for surviving nuclear winter and other abrupt sunlight-reducing catastrophes. *Risk Analysis*, risa.14072. https://doi.org/10.1111/risa.14072

Cai, W., Santoso, A., Collins, M., Dewitte, B., Karamperidou, C., Kug, J.-S., Lengaigne, M., McPhaden, M. J., Stuecker, M. F., Taschetto, A. S., Timmermann, A., Wu, L., Yeh, S.-W., Wang, G., Ng, B., Jia, F., Yang, Y., Ying, J., Zheng, X.-T., … Zhong, W. (2021). Changing El Niño–Southern Oscillation in a warming climate. *Nature Reviews Earth & Environment*, 2(9), Article 9. https://doi.org/10.1038/s43017-021-00199-z

Cassidy, M., & Mani, L. (2022). Huge volcanic eruptions: Time to prepare. *Nature*, 608(7923), 469–471. https://doi.org/10.1038/d41586-022-02177-x

Cernev, T. (2022). Global catastrophic risk and planetary boundaries: The relationship to global targets and disaster risk reduction. https://www.undrr.org/publication/global-catastrophic-risk-and-planetary-boundaries-relationship-global-targets-and

Chesner, C. A., Rose, W. I., Deino, A., Drake, R., & Westgate, J. A. (1991). Eruptive history of Earth's largest Quaternary caldera (Toba, Indonesia) clarified. *Geology*, 19(3), 200–203. https://doi.org/10.1130/0091-7613(1991)019<0200:EHOESL>2.3.CO;2

Cooley, S. R., & Doney, S. C. (2009). Anticipating ocean acidification's economic consequences for commercial fisheries. *Environmental Research Letters*, 4(2), 024007. https://doi.org/10.1088/1748-9326/4/2/024007

Coupe, J., Bardeen, C. G., Robock, A., & Toon, O. B. (2019). Nuclear Winter Responses to Nuclear War Between the United States and Russia in the Whole Atmosphere Community Climate Model Version 4 and the Goddard Institute for Space Studies ModelE. *Journal of Geophysical Research: Atmospheres*, 124(15), 8522–8543. https://doi.org/10.1029/2019JD030509

Coupe, J., Stevenson, S., Lovenduski, N. S., Rohr, T., Harrison, C. S., Robock, A., Olivarez, H., Bardeen, C. G., & Toon, O. B. (2021). Nuclear Niño response observed in simulations of nuclear war scenarios. *Communications Earth & Environment*, 2(1), Article 1. https://doi.org/10.1038/s43247-020-00088-1

De Boeck, H. J., Bloor, J. M. G., Kreyling, J., Ransijn, J. C. G., Nijs, I., Jentsch, A., & Zeiter, M. (2018). Patterns and drivers of biodiversity–stability relationships under climate extremes. *Journal of Ecology*, 106(3), 890–902. https://doi.org/10.1111/1365-2745.12897

Denkenberger, D. C., & Pearce, J. M. (2015). Feeding everyone: Solving the food crisis in event of global catastrophes that kill crops or obscure the sun. *Futures*, 72, 57–68. https://doi.org/10.1016/j.futures.2014.11.008

Denkenberger, D., & Pearce, J. (2015). *Feeding everyone no matter what: Managing food security after global catastrophe.* Academic Press.

Doney, S. C., Fabry, V. J., Feely, R. A., & Kleypas, J. A. (2009). Ocean Acidification: The Other CO 2 Problem. *Annual Review of Marine Science*, 1(1), 169–192. https://doi.org/10.1146/annurev.marine.010908.163834

García Martínez, J. B., Egbejimba, J., Throup, J., Matassa, S., Pearce, J. M., & Denkenberger, D. C. (2021). Potential of microbial protein from hydrogen for preventing mass starvation in catastrophic scenarios. *Sustainable Production and Consumption*, 25, 234–247. https://doi.org/10.1016/j.spc.2020.08.011

García Martínez, J. B., Pearce, J. M., Throup, J., Cates, J., Lackner, M., & Denkenberger, D. C. (2022). Methane Single Cell Protein: Potential to Secure a Global Protein Supply Against Catastrophic Food Shocks. *Frontiers in Bioengineering and Biotechnology*, 10, 906704. https://doi.org/10.3389/fbioe.2022.906704

Habbick, B. (1983). Casualties in a nuclear war. *Canadian Journal of Public Health = Revue Canadienne de Sante Publique*, 74(1). https://pubmed.ncbi.nlm.nih.gov/6850478/

Harrison, C. S., Rohr, T., DuVivier, A., Maroon, E. A., Bachman, S., Bardeen, C. G., Coupe, J., Garza, V., Heneghan, R., Lovenduski, N. S., Neubauer, P., Rangel, V., Robock, A., Scherrer, K., Stevenson, S., & Toon, O. B. (2022). A New Ocean State After Nuclear War. *AGU Advances*, 3(4). https://doi.org/10.1029/2021AV000610

Intergovernmental Panel on Climate Change (Ed.). (2014). Carbon and Other Biogeochemical Cycles. *In Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 465–570). Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.015

Jehn, F. U., Dingal, F. J., Mill, A., Harrison, C. S., Ilin, E., Roleda, M. Y., James, S. C., & Denkenberger, D. C. (2023). Seaweed as a resilient food solution after a nuclear war. *Zenodo*. https://doi.org/10.5281/zenodo.7615254

Jehn, F. U., Kemp, L., Ilin, E., Funk, C., Wang, J. R., & Breuer, L. (2022). Focus of the IPCC Assessment Reports Has Shifted to Lower Temperatures. *Earth's Future*, 10(5), e2022EF002876. https://doi.org/10.1029/2022EF002876

Jehn, F. U., Schneider, M., Wang, J. R., Kemp, L., & Breuer, L. (2021). Betting on the best case: Higher end warming is underrepresented in research. *Environmental Research Letters*, 16(8), 084036. https://doi.org/10.1088/1748-9326/ac13ef

Kareiva, P., & Carranza, V. (2018). Existential risk due to ecosystem collapse: Nature strikes back. *Futures*, 102, 39–50. https://doi.org/10.1016/j.futures.2018.01.001

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., … Vertenstein, M. (2015). The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability. *Bulletin of the American Meteorological Society*, 96(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Kemp, L., Xu, C., Depledge, J., Ebi, K. L., Gibbins, G., Kohler, T. A., Rockström, J., Scheffer, M., Schellnhuber, H. J., Steffen, W., & Lenton, T. M. (2022). Climate Endgame: Exploring catastrophic climate change scenarios. *Proceedings of the National Academy of Sciences*, 119(34), e2108146119. https://doi.org/10.1073/pnas.2108146119

Leinfelder, R. R., Berndt, J., & Humboldt-Universität zu Berlin (Eds.). (2017). *Science meets comics: Proceedings of the Symposium on Communicating and Designing the Future of Food in the Antropocene*. Ch.A. Bachmann Verlag.

Liu, P. R., & Raftery, A. E. (2021). Country-based rate of emissions reductions should increase by 80% beyond nationally determined contributions to meet the 2 °C target. *Communications Earth & Environment*, 2(1), 1–10. https://doi.org/10.1038/s43247-021-00097-8

Lovenduski, N. S., Harrison, C. S., Olivarez, H., Bardeen, C. G., Toon, O. B., Coupe, J., Robock, A., Rohr, T., & Stevenson, S. (2020). The Potential Impact of Nuclear Conflict on Ocean Acidification. *Geophysical Research Letters*, 47(3), e2019GL086246. https://doi.org/10.1029/2019GL086246

Luterbacher, J., & Pfister, C. (2015). The year without a summer. *Nature Geoscience*, 8(4), Article 4. https://doi.org/10.1038/ngeo2404

Mani, L., Tzachor, A., & Cole, P. (2021). Global catastrophic risk from lower magnitude volcanic eruptions. *Nature Communications*, 12(1), Article 1. https://doi.org/10.1038/s41467-021-25021-8

Mills, M. J., Toon, O. B., Turco, R. P., Kinnison, D. E., & Garcia, R. R. (2008). Massive global ozone loss predicted following regional nuclear conflict. *Proceedings of the National Academy of Sciences*, 105(14), 5307–5312. https://doi.org/10.1073/pnas.0710058105

Minoli, S., Jägermeyr, J., Asseng, S., Urfels, A., & Müller, C. (2022). Global crop yields can be lifted by timely adaptation of growing periods to climate change. *Nature Communications*, 13. https://doi.org/10.1038/s41467-022-34411-5

Mohamed, A., DeClerck, F., Verburg, P. H., Obura, D., Abrams, J. F., Zafra-Calvo, N., Rocha, J., Estrada-Carmona, N., Fremier, A., Jones, S. K., Meier, I. C., & Stewart-Koster, B. (2022). Biosphere functional integrity for people and Planet (p. 2022.06.24.497294). bioRxiv. https://doi.org/10.1101/2022.06.24.497294

Mörsdorf, J. (2021). Simulating potential yield if industry is disabled: Applying a generalized linear modelling approach to major food crops [Master Thesis]. Justus-Liebig University.

Newhall, C., Self, S., & Robock, A. (2018). Anticipating future Volcanic Explosivity Index (VEI) 7 eruptions and their chilling impacts. *Geosphere*, 14(2), 572–603. https://doi.org/10.1130/GES01513.1

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books, 480 pp.

Özdoğan, M., Robock, A., & Kucharik, C. J. (2013). Impacts of a nuclear war in South Asia on soybean and maize production in the Midwest United States. *Climatic Change*, 116(2), 373–387. https://doi.org/10.1007/s10584-012-0518-1

Pearce, J. M., Khaksari, M., & Denkenberger, D. (2019). Preliminary Automated Determination of Edibility of Alternative Foods: Non-Targeted Screening for Toxins in Red Maple Leaf Concentrate. *Plants*, 8(5), 110. https://doi.org/10.3390/plants8050110

Persson, L., Carney Almroth, B. M., Collins, C. D., Cornell, S., de Wit, C. A., Diamond, M. L., Fantke, P., Hassellöv, M., MacLeod, M., Ryberg, M. W., Søgaard Jørgensen, P., Villarrubia-Gómez, P., Wang, Z., & Hauschild, M. Z. (2022). Outside the Safe Operating Space of the Planetary Boundary for Novel Entities. *Environmental Science & Technology*, 56(3), 1510–1521. https://doi.org/10.1021/acs.est.1c04158

Pittock, A. B. (1988). Climatic Catastrophes: The Local and Global Effects of Greenhouse Gases and Nuclear Winter. In M. I. El-Sabh & T. S. Murty (Eds.), *Natural and Man-Made Hazards* (pp. 621–633). Springer Netherlands. https://doi.org/10.1007/978-94-009-1433-9_42

Rampino, M. R. (2002). Supereruptions as a Threat to Civilizations on Earth-like Planets. *Icarus*, 156(2), 562–569. https://doi.org/10.1006/icar.2001.6808

Rivers, M., Hinge, M., García Martínez, J. B., Tieman, R., Jaeck, V., Butt, T., Jehn, F., Grillo, V., & Denkenberger, D. (2022). Food System Adaptation and Maintaining Trade Greatly Mitigate Global Famine in Abrupt Sunlight Reduction Scenarios. https://doi.org/10.21203/rs.3.rs-1446444/v1

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S. I., Lambin, E., Lenton, T., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P., Costanza, R., Svedin, U., … Foley, J. (2009). Planetary Boundaries: Exploring the Safe Operating Space for Humanity. *Ecology and Society*, 14(2). https://doi.org/10.5751/ES-03180-140232

Rougier, J., Sparks, R. S. J., Cashman, K. V., & Brown, S. K. (2018). The global magnitude–frequency relationship for large explosive volcanic eruptions. *Earth and Planetary Science Letters*, 482, 621–629. https://doi.org/10.1016/j.epsl.2017.11.015

Savitch, E., Frank, A., Carroll-Nellenback, J., Haqq-Misra, J., Kleidon, A., & Alberti, M. (2021). Triggering a Climate Change Dominated "Anthropocene": Is It Common among Exocivilizations? *The Astronomical Journal*, 162(5), 196. https://doi.org/10.3847/1538-3881/ac1a71

Scherrer, K. J. N., Harrison, C. S., Heneghan, R. F., Galbraith, E., Bardeen, C. G., Coupe, J., Jägermeyr, J., Lovenduski, N. S., Luna, A., Robock, A., Stevens, J., Stevenson, S., Toon, O. B., & Xia, L. (2020). Marine wild-capture fisheries after nuclear war. *Proceedings of the National Academy of Sciences*, 117(47), 29748–29758. https://doi.org/10.1073/pnas.2008256117

Schindler, D. W., & Vallentyne, J. R. (2008). The algal bowl: Overfertilization of the world's freshwaters and estuaries. *Earthscan*.

Smith, J., & Smith, T. (1981). Radiation injury and effects of early fallout. *Br Med J (Clin Res Ed)*, 283(6295), 844–846. https://doi.org/10.1136/bmj.283.6295.844

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R., Carpenter, S. R., de Vries, W., de Wit, C. A., Folke, C., Gerten, D., Heinke, J., Mace, G. M., Persson, L. M., Ramanathan, V., Reyers, B., & Sörlin, S. (2015). Planetary boundaries: Guiding human development on a changing planet. *Science*, 347(6223), 1259855. https://doi.org/10.1126/science.1259855

Tabor, C. R., Bardeen, C. G., Otto-Bliesner, B. L., Garcia, R. R., & Toon, O. B. (2020). Causes and Climatic Consequences of the Impact Winter at the Cretaceous-Paleogene Boundary. *Geophysical Research Letters*, 47(3), e60121. https://doi.org/10.1029/2019GL085572

Tang, A., & Kemp, L. (2021). A Fate Worse Than Warming? Stratospheric Aerosol Injection and Global Catastrophic Risk. *Frontiers in Climate*, 3. https://www.frontiersin.org/article/10.3389/fclim.2021.720312

Tarshish, N., & Romps, D. M. (2022). Latent Heating Is Required for Firestorm Plumes to Reach the Stratosphere. Journal of Geophysical Research: *Atmospheres*, 127(18), e2022JD036667. https://doi.org/10.1029/2022JD036667

Throup, J., García Martínez, J. B., Bals, B., Cates, J., Pearce, J. M., & Denkenberger, D. C. (2022). Rapid repurposing of pulp and paper mills, biorefineries, and breweries for lignocellulosic sugar production in global food catastrophes. *Food and Bioproducts Processing*, 131, 22–39. https://doi.org/10.1016/j.fbp.2021.10.012

Turco, R. P., Toon, O. B., Ackerman, T. P., Pollack, J. B., & Sagan, C. (1983). Nuclear Winter: Global Consequences of Multiple Nuclear Explosions. *Science*, 222(4630), 1283–1292. https://doi.org/10.1126/science.222.4630.1283

Ward, P. J., Daniell, J., Duncan, M., Dunne, A., Hananel, C., Hochrainer-Stigler, S., Tijssen, A., Torresan, S., Ciurean, R., Gill, J. C., Sillmann, J., Couasnon, A., Koks, E., Padrón-Fumero, N., Tatman, S., Tronstad Lund, M., Adesiyun, A., Aerts, J. C. J. H., Alabaster, A., … de Ruiter, M. C. (2022). Invited perspectives: A research agenda towards disaster risk management pathways in multi-(hazard-)risk assessment. *Natural Hazards and Earth System Sciences*, 22(4), 1487–1497. https://doi.org/10.5194/nhess-22-1487-2022

Xia, L., Robock, A., Scherrer, K., Harrison, C. S., Bodirsky, B. L., Weindl, I., Jägermeyr, J., Bardeen, C. G., Toon, O. B., & Heneghan, R. (2022). Global food insecurity and famine from reduced crop, marine fishery and livestock production due to climate disruption from nuclear war soot injection. *Nature Food*, 1–11. https://doi.org/10.1038/s43016-022-00573-0

Yuan, T., Song, H., Wood, R., Wang, C., Oreopoulos, L., Platnick, S. E., von Hippel, S., Meyer, K., Light, S., & Wilcox, E. (2022). Global reduction in ship-tracks from sulfur regulations for shipping fuel. *Science Advances*, 8(29), eabn7988. https://doi.org/10.1126/sciadv.abn7988

Stanford | Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Is Climate Change Ungovernable?

Paul N. Edwards[1]

Citation: Edwards, Paul N.. Is Climate Change Ungovernable? *Proceedings of the Stanford Existential Risks Conference* 2023, 133-146. https://doi.org/10.25740/yc096zw4572

Academic Editor: Dan Zimmer, Trond Undheim

Funding: This work was funded in part by Open Philanthropy.

Conflict of Interest Statement: I declare no financial or other conflicts of interest.

Informed Consent Statement: N/A

Acknowledgments: Stanford students Anuj Chetty, YuQing Jiang, Spencer Seay, and Keven Victoria provided invaluable research assistance for this article.

Abstract: This paper reviews the potential for catastrophic, civilization-threatening climate change within the next 2-3 centuries if climate sensitivity is on the high end of IPCC estimates and the thresholds of various tipping points are crossed. I argue that empirical evidence supports a substantial likelihood of future climate policy reversals by major emitters, resulting in continuing accumulation of greenhouse gases in the atmosphere. On high-end sensitivity estimates, Paris Agreement pledges to date are insufficient even if fully implemented. Policy is always reversible, and major reversals have already occurred. Climate denialism and misdirection can, and probably will, be amplified by artificial intelligence and social media. Finally, the focus on international governance mechanisms obscures the many levels of jurisdiction that must be engaged for strong climate policy to take effect. The paper concludes that while renewable energy progress presents a hopeful note, chances are high that current structures of climate governance will not succeed in preventing catastrophic levels of climate change.

Keywords: climate change; tipping points; climate policy; governance; artificial intelligence

[1] Co-Director, Stanford Existential Risks Initiative, Center for International Security and Cooperation, Stanford University; pedwards@stanford.edu

* Correspondence: pedwards@stanford.edu

1. Introduction: Climate Change Will Not Stop in 2100

A little-noticed, but deeply significant legacy of the post-1950 explosion of environmental knowledge is a *time horizon that ends in the year 2100*. In the IPCC Working Group I report from the Sixth Assessment cycle (hereafter AR6 WGI), the year 2100 is mentioned more than 1200 times. The year 2150 appears 45 times. 2300 appears 163 times, while 2200 and 2250 are barely mentioned. Most mentions of the year 2300 regard just two dimensions of climate change: sea level rise and ice sheet melting, both slow processes taking place on multi-decadal to centennial timescales. The scenarios used in most recent IPCC reports typically end in 2100, and the key phrase "long term" is deliberately restricted to the 2081-2100 timeframe.

In the mid-20th century, when modern climate science took shape, the year 2100 seemed very distant. Further, uncertainties inherent in projecting the evolution of human societies, technologies, geophysical systems, land and ocean ecologies, and the interactions among all of these seemed to render projections of the future beyond 2100 imprudent. As a result, the 2100 endpoint became an unacknowledged scientific convention that still dominates most projections of environmental change. Yet many children born today will be alive well into the 22nd century – and their own children will still be in the prime of life. For them, the "long term" will not end in 2100.

A similar story can be told about carbon dioxide doubling. When continuous measurements of CO_2 began in 1957, concentrations had risen by only about 35 ppm from their preindustrial (~1750 CE) level of around 275-280 ppm. Early analyses as far back as Arrhenius (1896) used doubling as a convenient benchmark for calculating warming potentials, and while larger, smaller, and fractional multipliers have often appeared in the literature, by the 1970s $2xCO_2$ had become a standard model experiment (Edwards, 2010). As a result, climate scientists often use "climate sensitivity" as shorthand for "equilibrium climate sensitivity (ECS) to CO_2 doubling over preindustrial concentrations," even though the formal definition of ECS references only a unit change in radiative forcing – not any specific time period, multiplier, or greenhouse gas.

The CO_2 doubling that seemed far away in the 1960s – when concentrations measured at Mauna Loa were rising at just 0.9ppm/yr – now appears much closer. In 2022, global CO_2 concentration reached 420ppm. If concentrations were to continue rising at the 2.4ppm/yr rate observed across the decade 2011-2020, CO_2 doubling (~550ppm) would occur around 2075. Further, concentrations of other greenhouse gases – especially methane, 80 times as potent as carbon dioxide on a 20-year basis , and nitrous oxide – are rising at record rates even as the rate of increase in carbon dioxide may be on the verge of diminishing (US National Oceanic and Atmospheric Administration, 2023).

Consider that the IPCC's 2021 "likely" (16-84% probability) range of equilibrium climate sensitivity (ECS) is 2.5-4°C, with a best estimate of 3°C. However, the high

end of the "very likely" (5-95% probability) ECS range is pegged at 5°C. Further, according to the report, "*a higher ECS than 5°C cannot be ruled out"* (Masson-Delmotte, V. et al., 2021, p. Chapter 7 p 111). In other words, there is about a 5 percent chance – 1 in 20 – that ECS actually *exceeds* 5°C.

Due to the slow response of oceans and ice sheets, climatic equilibrium would not be reached within the 21st century even if all anthropogenic emissions stopped tomorrow.[1] As a result, most scientific discussions of futures to 2100 disregard ECS in favor of other figures, also defined in terms of CO_2 doubling. For example, transient climate response (TCR) is modeled by adding greenhouse gases to preindustrial concentrations at 1% per year until they double, which takes 70 years (due to compounding). In AR6 WGI, this results in an assessed *likely* range of 1.4-2.2°C – narrower than AR5's 1.0-2.5, and lower at the high end (Masson-Delmotte, V. *et al.* 2021: Chapter 7). Both this quantity and other measures (such as transient climate response to emissions) may more accurately represent response of the climate system to emissions as they are added over time. However, ECS better reflects the longer-term response beyond 2100.

Estimates of anthropogenic CO_2 emissions seem to show them slowing, and they are widely anticipated to peak sometime this decade before beginning to decline as low-carbon energy sources take the relay. Yet post-AR6 evidence suggests that some emissions have been greatly underestimated and/or underreported (International Energy Agency, 2022; Mooney et al., 2022). And in fact, as already observed, atmospheric concentrations of CO_2, methane, nitrous oxide, and certain other greenhouse gases continue to rise (Ritchie et al., 2020), meaning that a substantial fraction of anthropogenic CO_2 emissions will remain in the atmosphere for centuries to millenia beyond (Archer & Brovkin, 2008; Knutti & Rogelj, 2015). In short, in 2023 there is still no *empirical* evidence that CO_2 doubling will not occur within this century. Should actual ECS prove to be 5°C or higher, doubled CO_2 (or higher) would lock in catastrophic levels of climate change in the centuries beyond 2100.

2. Climate Surprises, Underestimates, and Tipping Points

Climate modeling is based on physical theory and process understanding, but parameterizes many sub-grid-scale physical variables. Many model parameters are set "semi-empirically," i.e., derived from historical data about the real climate system. As a result, models in part reflect how those systems have behaved in the past. Knowledge of states outside historical experience comes primarily from studies of paleo time periods and of the climates of other planets, which are inherently less precise than instrumental records.

[1] In fact, climatic equilibrium is an abstract state that can be modeled, but is never attained in the real world due to natural fluctuations in radiative forcing factors – but for the purposes of my argument here, this is unimportant.

As Earth drifts further into a climatic state unprecedented in historical time, this empirical aspect of climate models may make their high-end results misleading. As Allen and Frame (2007) once put it,

> ... the properties of the climate system that we can observe now do not distinguish between a climate sensitivity… of 4°C and [one greater than] 6°C… Once the world has warmed by 4°C, conditions will be so different from anything we can observe today (and still more different from the last ice age) that it is inherently hard to say when the warming will stop.

So-called "emergent constraints" derived from observations helped to limit the ECS range in AR6 WGI, but these too are based on historical and paleo records (Williamson *et al.* 2021). If sensitivity is very high, climate surprises – unanticipated phenomena – become more likely. So do the chances of exceeding the thresholds of various "tipping points" at which a previously stable system abruptly reorganizes, becomes unstable, or "flips" from a negative to a positive feedback. While these possibilities are sketched in AR6 WG1, the report indicates that they are "not fully represented" (Masson-Delmotte, V. et al., 2021, p. Chapter 1, Box 11).

Even at today's roughly 1.2°C of warming, surprises have already emerged. Over the last 15-20 years, increased meandering or "waviness" of the jet stream – associated with Arctic warming and declining Arctic sea ice – has generated unprecedented weather in the northern hemisphere, such as long, extreme cold spells in North America and Europe related to the so-called "polar vortex" (Francis & Vavrus, 2015; Kim et al., 2014; Zhang et al., 2016), as well as recurring, concurrent heatwaves in specific regions of North America, Europe, and Asia (Bartusek & Kornhuber, 2022; Kornhuber et al., 2020; Mann et al., 2017; Rousi et al., 2022). These effects were not predicted by older climate models.

The rates of some known climatic changes have been underestimated as well. A post-AR6 study showed that Arctic warming is occurring at nearly 4 times the global average rate, rather than twice as fast, as previously believed (Rantanen et al., 2022). Recent estimates of the melt rates of glaciers around the world have risen considerably versus previous findings (Pörtner, 2022). Rapid melting can lead to catastrophic flooding. The horrific Pakistani floods of summer 2022 inundated one-third of the country and displaced some 32 million people (Mallapaty, 2022). They resulted mainly from extreme rainfall, but were amplified by melt-related outbreaks from multiple glacial lakes earlier in the summer (Nanditha et al., 2022). Melt rates of the Greenland ice sheet and the Thwaites "doomsday" glacier in East Antarctica have increased. Thwaites is set to unpin from the sea floor within this decade, allowing the glacier to slide more rapidly into the sea (Wild et al., 2022) (though it will be many decades before the entire glacier has melted into the ocean). Together, the Greenland ice sheet and Thwaites contain enough ice to raise global sea levels by 8 meters (18 feet). For context,

between 750 million and 1.1 billion people currently live on land that is less than 10 meters above sea level (MacManus et al., 2021).

Among many potential tipping points, rapid slowdown or collapse of the Atlantic Meridional Overturning Circulation (AMOC) would have major consequences for global climate. Strong evidence suggests that AMOC has weakened repeatedly in past climate regimes, including during the most recent deglaciation, possibly due to rapid inflows of freshwater from the melting Laurentide ice sheet. Meltwater from the Greenland ice sheet and other northern high-latitude glaciers could play a similar role within this century. Evidence assessed in both AR5 and AR6 suggests that AMOC has already weakened. AR6 WGI reported only "*medium confidence* that there will not be an abrupt collapse before 2100" (Masson-Delmotte, V. et al. 2021, TS-38). Looking further out, the IPCC report on oceans and cryosphere noted that "by 2300, an AMOC collapse is *about as likely as not* for a high emissions scenario" (Pörtner, 2022, SPM p 19).

Another potential tipping point is permafrost melting. As noted, Rantanen et al. (2022) found that since 1979, the Arctic has warmed almost 4 times faster than the globe as a whole, with some Arctic regions heating up to 7 times faster. Organic matter bound in Arctic permafrost is estimated to contain almost twice as much carbon as the entire global atmosphere (Pörtner et al. 2022, SPM p 6). As it thaws, increasing amounts of methane will be released – a process that is "irreversible at centennial [i.e. century] timescales (*high confidence*)" (Masson-Delmotte, V. *et al.* 2021: SPM, 28). The degree of warming conducive to thaw has already occurred in some permafrost regions more than 70 years prior to previous model projections (Mustonen, 2021). In a recent Earth system model simulation, a point of no return has *already* been reached in which "self-sustained thawing of the permafrost occurs for hundreds of years" *even without* further anthropogenic greenhouse gas emissions (Randers & Goluke, 2020) – a positive feedback that would contribute to ongoing global heating.

The litany of potential tipping points continues across numerous other climate-related phenomena, some with the potential to "flip" from negative to positive feedbacks. Beyond 3°C, forests may switch from carbon sinks to carbon sources (Scholze et al., 2006). Declining Arctic sea ice and faster-than-predicted spread of tree cover into the Arctic are rapidly changing the albedo of that region from a reflective state (negative feedback) to an absorptive one (positive feedback) (Berner & Goetz, 2022; Dial et al., 2022; Randers & Goluke, 2020). If these "flips" into positive feedback states occur in multiple major Earth systems, warming could continue in the coming centuries *even without* anthropogenic emissions.

3. Governing Climate Change

Global heating first emerged as a major political issue in the 1980s. During that decade, negotiation of the 1985 Vienna Convention for the Protection of the Ozone Layer was quickly followed by the 1987 Montreal Protocol on Substances that Deplete the Ozone Layer – among the most successful environmental treaties in

history. Signatories to these agreements agreed to meet regularly in Conferences of Parties (COPs), with the goal of a binding treaty regulating halocarbons and other substances responsible for anthropogenic ozone depletion in the stratosphere. 198 nations, including all UN member states, ratified the Vienna Convention. Numerous modifications adopted at COPs since 1987 have extended the Montreal Protocol's original regulations to more ozone-depleting chemicals – many of which are greenhouse gases up to 1000 times more potent than carbon dioxide.

The speed and effectiveness of this process led to optimism that a similar approach might work for climate change. Indeed, many scientists involved in the first IPCC report (1990) had previously helped establish both the reality of anthropogenic ozone depletion and the significance of its potential effects on human bodies, crops, and natural ecosystems. As a result, the 1992 United Nations Framework Convention on Climate Change (UNFCCC) closely paralleled the Vienna Convention. Conferences of Parties to the UNFCCC have been held each year since 1995.

3.1 From Kyoto to Paris: Top-Down vs. Bottom-Up Agreements

The expectation of a unified global agreement under the UNFCCC first bore fruit in the form of the Kyoto Protocol, "adopted" by its signatories in 1997. The complex Kyoto agreement set binding targets and timetables for some 41 OECD nations, plus "economies in transition" (Russia and its former Eastern European satellite states), to reduce their emissions of six major greenhouse gases to around 5 percent below than their 1990 levels by 2012. It also provided a "clean development mechanism" allowing these nations to receive credits toward their own emissions reductions by supporting projects that reduced emissions in less-developed countries instead.

Kyoto was originally intended more as a "proof of concept" (that a global agreement could work to mitigate climate change) than as an attempt to deeply cut global emissions. Successor agreements were widely expected to engage more stringent mitigation targets, and to be reached within a few years. Due to its complexity and to ongoing wrangling over implementation, Kyoto did not actually enter into force until 2005, some 8 years after its adoption. Although President Bill Clinton signed the protocol, consistent opposition in the US Senate – which requires a two-thirds majority to approve international treaties – made it impossible for the US to ratify the treaty. Even before Kyoto entered into force, the George W. Bush administration withdrew from the Protocol in 2002, citing harms to the US economy while also pointing to the protocol's failure to limit China's rapidly rising emissions. US withdrawal from Kyoto was widely regarded as a massive failure of leadership on the international stage. In 2012, when the long-awaited replacement agreement still had not materialized, the Kyoto Protocol was extended to a second 8-year commitment period with somewhat more ambitious emissions targets.

While it did succeed in setting the precedent of a global climate change agreement, post-hoc evaluations of its effectiveness suggest that Kyoto had little or no genuine effect on emissions of the nations it regulated. On top of that, *global* greenhouse gas emissions continued to increase throughout both Kyoto commitment periods.

The long-awaited successor agreement finally materialized 18 years after Kyoto. The Paris Agreement of 2015 succeeded in large part because it flipped the structure of national commitments. Where Kyoto was top-down and regulatory, with a complex structure requiring the agreement of all parties, the Paris Agreement was bottom-up and voluntary. At Paris, each UNFCCC signatory brought forward its own "Nationally Determined Contribution" (NDC), i.e., its plans to mitigate and/or adapt to climate change, or in the case of less wealthy nations, their ambitions and requests for international assistance to adapt and/or mitigate. The Paris mechanism thus relies essentially on peer pressure rather than law. Public commitments would make it difficult for nations to backtrack, and the Agreement includes a "ratcheting mechanism" of 5-yearly review with a "global stocktake" of progress toward the Paris goals and resubmission of NDCs, expected to be increasingly ambitious.

A key, and entirely intentional, motivation for the bottom-up structure was to eliminate any role for the Senate in the US commitment. Since each nation's NDC is independent, the Paris Agreement did not constitute an international treaty and therefore did not require a Senate vote to ratify. This structure allowed President Obama to sign the agreement as an executive order on behalf of the USA – but also permitted President Trump to withdraw from the agreement, which he did in June 2017. Because the Paris accord specified that nations could not withdraw until 3 years after it came into effect (in November 2016), and required a further 1-year waiting period after signaling an intention to leave the accord, the Trump withdrawal did not actually occur until November 2020, less than three months before his first term ended. As a result the withdrawal itself had little direct effect on US commitments (though the Trump administration did implement regressive environmental policies incompatible with the USA's NDC). President Biden then reinstated the US commitment to Paris in 2021 on his first day in office.

3.2 Paris Pledges Are Insufficient Even if Fully Implemented

Without a doubt, the Paris Agreement is a major milestone in the decades-long quest for global climate governance. Yet multiple evaluations of the NDCs submitted to date – even factoring in additional commitments added since 2015 – show them falling far short of meeting the 2°C Paris goal. The 2021 IPCC Working Group I report concluded that the NDCs presented through about 2019 were insufficient to meet the Paris goal of a 2°C maximum warming even if all pledges were fully implemented.

Most subsequent studies confirm this conclusion. Modeling of the many possible combinations of NDC pledges and IPCC socioeconomic scenarios produces a

wide range of results. Recent studies estimate at best about a 66 percent chance of limiting warming to between 2-3°C by 2100 if current policies are fully implemented – with a 20 percent chance of hitting 3-4°C or higher. Including the most recent pledges and stated national *long-term* goals (not usually counted in these analyses) seems to reduce the likelihood of going beyond 3°C (Ou et al., 2021), and one suggests that if all of these are counted, it may be possible to limit warming to "just below" 2°C (Meinshausen et al., 2022).

Yet these studies count on full implementation and "increased ambition" over time. They ask what will happen if nations meet not only their unconditional goals, but also their long-term (less firm and generally non-binding) goals as well as those conditioned on financial and/or technological assistance from the wealthy nations, most of which has not in fact materialized to date. By taking into account only current actually-implemented policies without regard to NDC goals for 2030 or longer-term targets, the major studies all place warming by 2100 at 1.7-3.9°C, with central estimates clustering around 2.5-3.5°C. None of them attempt to describe what will happen in centuries beyond 2100.

3.3 Policy Is Always Reversible, and Major Reversals Have Already Occurred

It is in the very nature of human governance that no policy is permanent. The United States withdrawal from the Paris Agreement, announced in 2017 and completed in 2020, is a glaring example of a policy reversal within the existing UNFCCC governance structure. Brazil under President Jair Bolsonaro is another; when running for office, he pledged to leave the Paris accord. Although he abandoned this pledge after taking office in 2019, Bolsonaro's policies dramatically increased deforestation in the Amazon and undermined other environmental protections – results entirely incompatible with Brazil's Paris pledges. Australia offers a third example, vacillating between stronger and weaker climate policies in a succession of governments over the last 20 years. At COP-26 in Glasgow (2021), it failed to submit any new pledges beyond its existing 26 percent emissions reduction by 2030. Yet in 2022, under a newly elected Labor government, the nation increased its pledged reduction to 43 percent by 2030. In 2022, during her brief tenure as UK Prime Minister, Liz Truss reversed numerous climate-related policies, such as a ban on fracking for natural gas, and introduced new, regressive ones, including a ban on the use of most UK farmland for solar energy projects.

No one who follows international news can be unaware that right-wing and far-right political parties have been gaining ground in most of the world's democracies and won power in Hungary, Poland, and most recently Italy (Buchholz, 2022). Like the leaders mentioned above, these parties tend in general to be hostile to strong climate policy. Their motives include combinations of denialism, conspiracy theories, right-libertarian ideologies, the powerful influence of fossil fuel lobbies, and simple resistance to major change in energy systems and consumerist lifestyles. Should they come (or return) to power in France, Germany, the United States, or other key nations, further policy reversals

can be expected. On the other hand, not all authoritarian governments advance climate-damaging policy: China under Xi continues to expand coal consumption (International Energy Agency, 2023), but has also far outspent every other world nation on renewable electric power and electric vehicles (BloombergNEF, 2023, p. 10).

3.4 Climate Denialism and Misdirection Will Be Amplified by Artificial Intelligence

Other political news forebodes further resistance to strong climate governance. The year 2022 saw the fossil fuel majors more than double their profits, which reached a record $219 billion (Bousso & Bousso, 2023). Under Elon Musk's ownership, Twitter reinstated numerous accounts previously banned for mis- and disinformation, including many that propagate misleading or outright false information about climate change, electric vehicles, and other targets (Calma, 2022; Fazackerley, 2023). Fossil fuel companies continue to finance denialist organizations and right-wing political candidates friendly to their goals, even as they spend minor amounts (relative to their profits) on new "climate tech" and promote dubious "solutions" such as direct air capture of CO_2 or carbon capture and storage (Winters, 2023).

At this writing, the remarkable power of large language models such as ChatGPT and GPT-4 engenders both utopian and dystopian speculations. Future AI systems show promise of original contributions to better climate technology and better understanding of potential pathways to net zero emissions. Yet at present, they also represent a major threat to social learning due to their ability to generate and promote mis- and disinformation. A recent study by the Center for Countering Digital Hate found that Google's AI system, called Bard, "was willing to generate text promoting a given [false] narrative" in 96 of 100 test cases. In 78 of those cases, "Bard did so without any additional context negating the false claims." Bard's answer to climate change: "Relax and enjoy the ride. There is nothing we can do to stop climate change, so there is no point in worrying about it" (Center for Countering Digital Hate, 2023). It seems inevitable that unscrupulous actors will exploit this possibility to generate confusion, distrust, and doubt.

3.5 The Many Levels of Jurisdiction Endanger Strong Climate Governance

Climate policymaking proceeds at nearly every level of government and governance. In addition to grand international agreements such as the Paris accord and its nationally determined commitments, many regional, state, county, and city governments create and coordinate climate policy. Ideally, each of these levels would interact synergistically with the others, but in very many cases delay, gridlock, and incoherence result instead.

One example is the three major and two minor regional, independently managed electric power grids in the United States. Making efficient use of intermittent renewable power sources requires the ability to transmit power to where it is

needed, even across large areas. Yet 2,020 gigawatts of renewable power — almost double the existing electric capacity — are currently unable to connect to these grids due to slow, contentious, bureaucratic permitting processes. Further, existing high-voltage transmission lines are far insufficient to carry rapidly growing loads, but construction of new lines also endures years-long approval processes and resistance by landowners,and local governments along the proposed rights-of-way (Clifford, 2023).

Building codes governing energy supply and consumption offer another example. Many jurisdictions now seek to promote electrification of home heating, cooling, and cooking by prohibiting the installation of new gas and oil appliances and new natural gas infrastructure. But because building codes are typically set at the state and city levels, even this modest move requires the independent approval of numerous jurisdictions – offering an equal number of leverage points for delay and political resistance (Johnson, 2023). These problems are compounded in cities such as San Francisco, where older housing stock typically lacks sufficient amperage for full electrification. Some building codes require more work (and expense) than is genuinely necessary to upgrade home electrical panels (Quackenbush, 2023).

4. Conclusion

It follows from these observations that no matter how dire much of it may appear, political analysis of climate change issues generally tends toward optimism. It depends on a set of assumptions whose validity is at least questionable. Much of it assumes the IPCC "likely" range of climate sensitivity (an ECS of 2.5-4°C), without regard for the high-end tail of the "very likely" range (ECS of 4-5°C or higher). It assumes that a net-zero emissions goal is widely shared. It assumes that the Paris NDCs represent achievable policy goals. It assumes that policy reversals by major emitters either will not occur, or will be short-lived if they do.

This article has asked the necessary question: what if some or all of these assumptions are wrong? If climate sensitivity is very high and CO_2 concentrations reach doubling or beyond, as now seems almost inevitable, the potential exists for 5° or more of warming over the next two or three centuries *even if emissions are eventually reduced to zero.* Xu and Ramanathan (2017), along with many others, characterize this level of warming as "beyond catastrophic." Large areas of the planet would become uninhabitable, ecosystems would be massively altered, and vast numbers of other species would go extinct (Ceballos et al., 2017; Wallace-Wells, 2019). It is already certain that baked-in sea level rise will gradually reduce available land area in some of the world's most populous nations, eventually submerging low-lying cities such as Shanghai and entire areas such as Bangladesh and southern Florida. Climate tipping point thresholds, if crossed, could cause a dramatic reorganization of the global climate system, with catastrophic consequences for human habitation and agriculture, including mass migration on an enormous scale as people flee uninhabitable areas. If tipping points such as permafrost melting or Arctic albedo change "flip" negative feedbacks into

positive feedback states, these might drive warming even higher. Could people adapt? Human ingenuity suggests that some, even many, might survive – but it is hard to imagine true human flourishing in such a damaged world.

The clear empirical evidence of policy reversals by major emitters suggests that further reversals are not just possible, but likely over time. The pressure of increasingly obvious and dangerous climate change effects *may* lead to popular resistance against such reversals. Yet the rise of right-wing parties and authoritarian governments in recent years suggests a plausible alternate course, in which such governments succeed in preventing any real contestation of their rule. Many state governments in the USA are currently engaged in exactly this project.

Is climate change ungovernable? Maybe. Clearly good governance is possible in principle, as demonstrated by the very successful Montreal Protocol and the Vienna Convention on ozone depletion. Indeed, the grim discussion here could be flipped on its head. A "silver buckshot" approach to climate change recognizes that many clumsy, imperfect, and partial climate policies can still add up to substantial action. Even with all their flaws and their snail's pace, the UN Framework Convention on Climate Change and the 2015 Paris Agreement represent highly visible, collective commitments to mitigating climate change – perhaps the best that is pragmatically achievable within the existing international system. The fact that many jurisdictions are capable of climate action means that while some may resist, others can lead and create inspiration for their willing counterparts elsewhere: the Global Covenant of Mayors for Climate & Energy, for example, currently represents over 12,750 cities comprising more than 1 billion citizens. Regional, multilateral, and bilateral agreements offer other possibilities, with European Union climate policies proving remarkably successful. During the Ukraine-Russia war that began in 2022, the EU not only accepted self-harming sanctions on Russian oil and gas, but used them to spur its existing shift to renewable energy sources. All of these are encouraging signs of resolve.

Yet science calls us to assess our future prospects based not only on hopes, promises, and expectations, but also on empirical trends. Expectations that emissions will reduce to net zero sometime this century depend upon optimistic views of governments' capacity not only to create and implement climate policy, but also to ratchet ambition upwards over time and avoid backsliding. Such views are not (yet) borne out by empirical trends. The greenhouse-gas emissions curve has flattened, but it is still on an upward trajectory; 2022 emissions were the highest ever (International Energy Agency, 2023). Plummeting prices for renewable energy offer a hopeful note, but they must contend with the gigantic installed base of fossil fuel infrastructures, along with that industry's enormous marketing and lobbying powers. Finally, the disquieting rise of nationalist authoritarian governments and far-right political parties around the world suggests the real possibility of future policy reversals. All of this evidence points to the likelihood of continuing resistance on the path to net-zero emissions, and

to the real possibility of catastrophic, civilization-threatening climate change within the next 2-3 centuries, if not before.

References

Allen, M. R., & Frame, D. J. (2007). Call Off the Quest. *Science*, *318*(5850), 582–583. https://doi.org/10.1126/science.1149988

Archer, D., & Brovkin, V. (2008). The millennial atmospheric lifetime of anthropogenic CO2. *Climatic Change*, *90*(3), 283–297. https://doi.org/10.1007/s10584-008-9413-1

Arrhenius, S. (1896). On the Influence of Carbonic Acid in the Air upon the Temperature of the Ground. *Philosophical Magazine and Journal of Science*, *41*, 237–276.

Bartusek, S., & Kornhuber, K. (2022). Analysing the 2021 North American heatwave to understand extraordinary heat events. *Nature Climate Change*, 1–2. https://doi.org/10.1038/s41558-022-01532-0

Berner, L. T., & Goetz, S. J. (2022). Satellite observations document trends consistent with a boreal forest biome shift. *Global Change Biology*, *28*(10), 3275–3292. https://doi.org/10.1111/gcb.16121

BloombergNEF. (2023). *Energy Transition Investment Trends 2023*. Bloomberg News.

Bousso, R., & Bousso, R. (2023, February 8). Big Oil doubles profits in blockbuster 2022. *Reuters*. https://www.reuters.com/business/energy/big-oil-doubles-profits-blockbuster-2022-2023-02-08/

Buchholz, K. (2022, April 22). *Where Right-Wing Populists Had the Most Success in Europe*. Statista Infographics. https://www-statista-com.stanford.idm.oclc.org/chart/20094/national-election-success-of-far-right-parties-europe

Calma, J. (2022, December 5). *Climate misinformation explodes on Twitter—The Verge*. https://www.theverge.com/2022/12/5/23494220/elon-musk-twitter-climate-misinformation-rise-analysis

Ceballos, G., Ehrlich, P., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proc Natl Acad Sci U S A*, *114*(30), E6089–E6096.

Center for Countering Digital Hate,. (2023, April 5). *Google's new Bard AI generate lies*. Center for Countering Digital Hate | CCDH. https://counterhate.com/research/misinformation-on-bard-google-ai-chat/

Clifford, C. (2023, February 17). *Why America's outdated energy grid is a climate problem*. CNBC. https://www.cnbc.com/2023/02/17/why-americas-outdated-energy-grid-is-a-climate-problem.html

Dial, R. J., Maher, C. T., Hewitt, R. E., & Sullivan, P. F. (2022). Sufficient conditions for rapid range expansion of a boreal conifer. *Nature*, 1–6. https://doi.org/10.1038/s41586-022-05093-2

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. MIT Press.

Fazackerley, A. (2023, May 14). Climate crisis deniers target scientists for vicious abuse on Musk's Twitter. *The Observer*. https://www.theguardian.com/environment/2023/may/14/climate-crisis-deniers-target-scientists-abuse-musk-twitter

Francis, J. A., & Vavrus, S. J. (2015). Evidence for a wavier jet stream in response to rapid Arctic warming. *Environmental Research Letters*, *10*(1), 014005. https://doi.org/10.1088/1748-9326/10/1/014005

International Energy Agency. (2023). *Coal 2022: Analysis and forecast to 2022*. IEA. https://www.iea.org/reports/coal-2022

International Energy Agency. (2022). *Methane emissions from the energy sector are 70% higher than official figures*. https://www.iea.org/reports/global-methane-tracker-2022

International Energy Agency. (2023). *CO2 Emissions in 2022*. IEA. https://www.iea.org/reports/co2-emissions-in-2022

Johnson, J. (2023, April 22). *California's push to ban natural gas hit a snag. Could it derail the entire effort?* San Francisco Chronicle. https://www.sfchronicle.com/climate/article/california-natural-gas-bans-17908841.php

Kim, B.-M., Son, S.-W., Min, S.-K., Jeong, J.-H., Kim, S.-J., Zhang, X., Shim, T., & Yoon, J.-H. (2014). Weakening of the stratospheric polar vortex by Arctic sea-ice loss. *Nature Communications*, *5*(1), Article 1. https://doi.org/10.1038/ncomms5646

Knutti, R., & Rogelj, J. (2015). The legacy of our CO2 emissions: A clash of scientific facts, politics and ethics. *Climatic Change*, *133*(3), 361–373. https://doi.org/10.1007/s10584-015-1340-3

Kornhuber, K., Coumou, D., Vogel, E., Lesk, C., Donges, J. F., Lehmann, J., & Horton, R. M. (2020). Amplified Rossby waves enhance risk of concurrent heatwaves in major breadbasket regions. *Nature Climate Change*, *10*(1), Article 1. https://doi.org/10.1038/s41558-019-0637-z

MacManus, K., Balk, D., Engin, H., McGranahan, G., & Inman, R. (2021). Estimating population and urban areas at risk of coastal hazards, 1990–2015: How data choices matter. *Earth System Science Data*, *13*(12), 5747–5801. https://doi.org/10.5194/essd-13-5747-2021

Mallapaty, S. (2022). Pakistan's floods have displaced 32 million people—Here's how researchers are helping. *Nature*, *609*(7928), 667–667. https://doi.org/10.1038/d41586-022-02879-2

Mann, M. E., Rahmstorf, S., Kornhuber, K., Steinman, B. A., Miller, S. K., & Coumou, D. (2017). Influence of Anthropogenic Climate Change on Planetary Wave Resonance and Extreme Weather Events. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/srep45242

Masson-Delmotte, V., P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (Eds.). (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Meinshausen, M., Lewis, J., McGlade, C., Gütschow, J., Nicholls, Z., Burdon, R., Cozzi, L., & Hackmann, B. (2022). Realization of Paris Agreement pledges may limit warming just below 2 °C. *Nature*, *604*(7905), 304–309. https://doi.org/10.1038/s41586-022-04553-z

Mooney, C., Eilperin, J., Butler, D., Muyskens, J., Narayanswamy, A., & Ahmed, N. (2022). *Countries' climate pledges built on flawed data, Post investigation finds*. Washington Post. https://www.washingtonpost.com/climate-environment/interactive/2021/greenhouse-gas-emissions-pledges-data/

Mustonen, T. (2021). *A Global State of Emergency: Extreme Weather Events in the Arctic and Beyond*. Climate Crisis Advisory Group. https://doi.org/10.13140/RG.2.2.15536.48644

Nanditha, J. S., Kushwaha, A. P., Singh, R., Malik, I., Solanki, H., Chupal, D. S., Dangar, S., Mahto, S. S., Mishra, V., & Vegad, U. (2022). The Pakistan flood of August 2022: Causes and implications. *Preprint*. https://doi.org/10.1002/essoar.10512560.1

Ou, Y., Iyer, G., Clarke, L., Edmonds, J., Fawcett, A. A., Hultman, N., McFarland, J. R., Binsted, M., Cui, R., Fyson, C., Geiges, A., Gonzales-Zuñiga, S., Gidden, M. J., Höhne, N., Jeffery, L., Kuramochi, T., Lewis, J., Meinshausen, M., Nicholls, Z., … McJeon, H. (2021). Can updated climate pledges limit warming well below 2°C? *Science*, *374*(6568), 693–695. https://doi.org/10.1126/science.abl8976

Pörtner, H.-O. (Ed.). (2022). *The Ocean and Cryosphere in a Changing Climate: Special Report of the Intergovernmental Panel on Climate Change*. https://doi.org/10.1017/9781009157964

Quackenbush, J. (2023, April 17). *Demand for heat pumps rises locally amid bans on gas appliances*. The North Bay Business Journal. https://www.northbaybusinessjournal.com/article/article/demand-for-heat-pumps-in-bay-area-rises-amid-bans-on-gas-appliances/

Randers, J., & Goluke, U. (2020). An earth system model shows self-sustained thawing of permafrost even if all man-made GHG emissions stop in 2020. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-75481-z

Rantanen, M., Karpechko, A. Y., Lipponen, A., Nordling, K., Hyvärinen, O., Ruosteenoja, K., Vihma, T., & Laaksonen, A. (2022). The Arctic has warmed nearly four times faster than the globe since 1979. *Communications Earth & Environment*, *3*(1), Article 1. https://doi.org/10.1038/s43247-022-00498-3

Ritchie, H., Roser, M., & Rosado, P. (2020). CO_2 and Greenhouse Gas Emissions. *Our World in Data*. https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

Rousi, E., Kornhuber, K., Beobide-Arsuaga, G., Luo, F., & Coumou, D. (2022). Accelerated western European heatwave trends linked to more-persistent double jets over Eurasia. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-31432-y

Scholze, M., Knorr, W., Arnell, N. W., & Prentice, I. C. (2006). A climate-change risk analysis for world ecosystems. *Proceedings of the National Academy of Sciences*, *103*(35), 13116–13120.

US National Oceanic and Atmospheric Administration. (2023, April 5). *Greenhouse gases continued to increase rapidly in 2022*. https://www.noaa.gov/news-release/greenhouse-gases-continued-to-increase-rapidly-in-2022

Wallace-Wells, D. (2019). *The Uninhabitable Earth: Life After Warming* (1st edition). Tim Duggan Books.

Wild, C. T., Alley, K. E., Muto, A., Truffer, M., Scambos, T. A., & Pettit, E. C. (2022). Weakening of the pinning point buttressing Thwaites Glacier, West Antarctica. *The Cryosphere*, *16*(2), 397–417. https://doi.org/10.5194/tc-16-397-2022

Winters, J. (2023, May 12). *Is carbon capture viable? In a new rule, the EPA is asking power plants to prove it*. Grist. https://grist.org/energy/is-carbon-capture-viable-in-a-new-rule-the-epa-is-asking-power-plants-to-prove-it/

Xu, Y., & Ramanathan, V. (2017). Well below 2 °C: Mitigation strategies for avoiding dangerous to catastrophic climate changes. *Proceedings of the National Academy of Sciences*, *114*(39), 10315–10323. https://doi.org/10.1073/pnas.1618481114

Zhang, J., Tian, W., Chipperfield, M. P., Xie, F., & Huang, J. (2016). Persistent shift of the Arctic polar vortex towards the Eurasian continent in recent decades. *Nature Climate Change*, *6*(12), Article 12. https://doi.org/10.1038/nclimate3136

Section III
Risk Intersections

Stanford | Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Investigating the Success Criteria for Dual-Use Biosecurity Education

Sofya Lebedeva [1*]

Citation: Lebedeva, Sofya. Investigating the Success Criteria for Dual-Use Biosecurity. *Proceedings of the Stanford Existential Risks Conference* 2023, 148-155. https://doi.org/10.25740/zk826vc3386

Academic Editor: Trond Undheim, Dan Zimmer

Funding: N/A

Conflict of Interest Statement: N/A

Informed Consent Statement: N/A

Acknowledgments: William Bracken Pearson Lund, Madison Lee McDougall, Clément Louis Messeri

Author Contributions: N/A

Abstract: The importance of preparedness for major biological events has been demonstrated substantially over the past few decades and was made strikingly necessary in recent years. Institutions need to fundamentally redesign their approach to biosecurity education to successfully mitigate biological threats and shortcomings. This issue is pressing due to its potential to mitigate or prevent the occurrence of events that cause the (near) extinction of humanity, thus drastically reducing the potential of future generations. Given the importance of this issue, an investigation of various approaches to biosecurity education was conducted. Following the investigation with web-based and database searches, the most relevant eight approaches emerged. They are as follows: UVienna Course, Local-People-Make-Sense Framework, EMR, NDMC Course, Train-The-Trainer, LNCV Course, QAU Course and the Karolinska Course. From this examination, seven key themes emerge as markers for the success of a course; length, context, personal connection and instructor excellency. From these takeaways, we can better understand the ways to promote responsible research practices while minimising the risks of catastrophic events resulting from the misuse of dual-use research of concern (DURC).

Keywords: biosecurity, education, responsible research practices, DURC

[1] Researcher, UC Berkeley, Weill Hall; sofya.lebedeva@berkeley.edu

[*] Correspondence: sofya.lebedeva@berkeley.edu

1. Introduction

Effective biosecurity education plays a crucial role in mitigating the risks associated with dual-use biosecurity research and promoting responsible practices. This article identifies key themes and examines the diverse approaches employed by academic institutions, shedding light on the evolution and effectiveness of biosecurity education.

Before discussing the need for better regulation of dual-use research, it is essential to establish the difference between biosafety and biosecurity and highlight recent funding initiatives aimed at strengthening both practices. Biosafety is defined as "the containment principles, technologies, and practices that are implemented to prevent unintentional exposure to pathogens and toxins, or their accidental release" (WHO, 2006). The WHO also defines Biosecurity as "the protection, control, and accountability for Valuable Biological Materials agents and toxins within laboratories, in order to prevent their loss, theft, misuse, diversion of, unauthorized access, or intentional unauthorized release" (WHO, 2006). For more extensive definitions the paper by Beeckman & Rüdelsheim provides an excellent summary of the various reports provided by other government agencies (Beeckman & Rüdelsheim, 2020).

Biosafety and biosecurity have recently received an influx of funding in the US. In a 2022 fact sheet from the White House, $1.8 billion was allotted to "enable the CDC and NIH to expand efforts to strengthen biosafety and biosecurity practices domestically and globally" (Fact Sheet FY 2023 President's Budget, 2022). This was backed by the Budget of the U. S. Government for the Fiscal Year 2023, which "provides $12.1 billion to NIH for research and development of vaccines, diagnostics, and therapeutics against high priority biological threats; biosafety and biosecurity" (Executive Office of the President, 2022, p. 67). This new influx of funding further illustrates a need for better education about biosecurity, as the budget makes no distinction between biosafety and biosecurity, and mostly stresses the need for the development of "safe and secure laboratory capacity and clinical trial infrastructure" without stressing the need to also educate on the importance of reducing or increasing the regulation of DURC research (Executive Office of the President, 2022, p. 67). It is necessary to increase the funding for DURC safety practices as it will be instrumental in preventing catastrophic pandemics which would affect billions of people from occurring. Investing in "professional biorisk management education, training, and accreditation" can help shift the scientists and the public's perception of this topic and can also "further increase the pool of experts available for biorisk consultations or review" (Greene et al., 2023). This will decrease the likelihood of an existential event due to biological misuse. This paper conducted a systematic literature review of approaches to biosecurity dual-use research education.

2. Methods

A web-based database search strategy was employed to investigate biosecurity education and identify academic interventions. Initially, Google Scholar was searched using key terms: "biosecurity education", "biosecurity education approaches" and "biosecurity education interventions" to retrieve relevant articles and research papers. Research Rabbit, a research aggregation platform, was utilized to expand the search to find adjacent papers. Keywords such as "dual-use research education" and "biosecurity training programs" were used to identify additional publications. An additional manual search was conducted by reviewing other works of the authors identified through the initial search process. This ensured the inclusion of relevant academic interventions that may have been missed initially. The search process was iterated until eight different interventions were identified, based on their relevance to biosecurity education, the student cohort size, and their potential to provide insights into different approaches and

themes. The identified interventions underwent careful review and analysis to extract key information, including intervention types, methodologies, target audiences, duration and any relevant contextual factors.

3. Results

Following the investigation of the various approaches to biosecurity education, the most relevant eight emerge. They are as follows: UVienna Course, Local People Framework, EMR, NDMC Course, Train-The-Trainer, LNCV Course, QAU Course and the Karolinska Course. From this examination, seven key themes emerge as markers for the success of a course; length, context, personal connection and instructor excellency. This section first provides a summary of the eight approaches and then proceeds to examine each of the interventions in more detail.

Figure 1: Overview of Biosecurity Implementation & Education Interventions

Number	Name	Location	Invention Type	Source
1	UVienna Course	University of Vienna, Austria	Educational Course	Revill et al., 2012
2	Local–People–Make Sense Framework	No Affiliation	Educational Framework	Dickmann et al., 2015
3	EMR	University of Bradford, UK	Open-Access Educational Resource	Minehata et al., 2013
4	NDMC Course	National Defense Medical College (NDMC), Japan	Educational Course	Minehata et al., 2013
5	Train-The-Trainer	University of Bradford, UK	Educational Course	Minehata et al., 2013
6	LNCV Course	Landau Network – Centro Volta (LNCV), Italy	Educational Course	Revill et al., 2012
7	QAU Course	Quaid-e-Azam University (QAU), Pakistan	Educational Course	Revill et al., 2012
8	Karolinska Course	Umea University, Karolinska Institutet & Uppsala University, Sweden	Educational Course & Biosecurity Network	Revill et al., 2012

3.1 UVienna Course

One of the earliest examples of biosecurity education is from Austria, where a laboratory safety course was introduced into the curriculum in 1995 and a complementary course on "biosafety, biosecurity and the safe handling of chemicals" was introduced in 2001 as an annual three European Credit Transfer and Accumulation System (ECTS) credit (Revill et al., 2012). This course covered relevant risk assessment and management methods, as well as presenting students with three different cases of misuse and requiring them to apply the methods learned in the previous parts of the course. The specific examples used were "The US Anthrax letters", "Iraq's biological weapons programmes" and "Aum Shinrikyo attack in Japan" (Revill et al., 2012). This approach advocates for integrating biosecurity issues into a more general course that addresses research safety as well as security issues to reach a wider audience.

3.2 Local–People–Make Sense Framework

Another approach to biosecurity education was tackled by the Local–People–Make Sense framework, which was developed by Dickmann et al. and aims to create a "sustainable, safer, and more secure laboratory capacity" through three principles of locally based decisions, the development of personal relationships and the improvement of laboratory safety procedures (Dickmann et al., 2015). This framework is designed to be used as a matrix to guide institutional decision-making. A limitation of this framework is the lack of specific guidelines, or as the authors describe it – "is not applying ticks to pre-set checklists", so while it is more nuanced and takes into account a greater number of stakeholders it is still too early in its development to be considered comprehensive (Dickmann et al., 2015). This framework also has the drawback of being standalone and not being supported by course material as in the example from the University of Vienna.

3.3 EMR

This is a more engaging online resource – an open-access Educational Module Resource (EMR) developed by the University of Bradford (Minehata et al., 2013). The website[1] aims to "assist university-level lecturers to incorporate material on biosecurity and dual-use issues into their life science courses" (Minehata et al., 2013). In comparison to the Local–People–Make Sense framework, this resource is significantly more detailed and is also available in three languages: English, Russian and Japanese (EMR, University of Bradford, 2023). The EMR provides 21 modules which focus on the Threat of Biological Warfare (BW), Biological Terrorism (BT), the International Prohibition Regime, The Dual-Use Dilemma and the Responsibilities of Life Scientists as well as the National Implementation of the Biological and Toxin Weapons Convention (BTWC). The open-access EMR is particularly useful as it can be "modified and tailored by users" to "fit the specific teaching modes and needs in various local educational contexts while also increasing the accessibility of the ideas of biosecurity by decreasing the need for academic institutions to reinvent the wheel when they want to design a biosecurity course (Minehata et al., 2013).

3.4 NDMC Course

The EMR has been implemented in the classroom at the National Defense Medical College (NDMC) in Japan at both the undergraduate and graduate levels (Minehata et al., 2013). The 21-module curriculum was presented over a 2-day course for the undergraduates and a 5-days course for the graduates. The significant difference between the two was that the undergraduate program focused more on the social aspects of biosecurity while the post-

[1] https://www.bradford.ac.uk/bioethics/educational-module-resources-emr/

graduate program presented specific cases of DURC research and asked students to consider practical preventative solutions "including the provision of national legislation" (Minehata et al., 2013). This case study shows the importance of tailoring the educational materials to the specific audience as "some highly advanced dual-use research examples were not successfully understood by earlier stage undergraduate students" due to the discrepancy in their scientific background and the complexity of the concepts.

3.5 Train-The-Trainer

Another example of a biosecurity educational resource is the Bradford online "Train-the-Trainer" programme (Minehata et al., 2013). This differs from the EMR, as the EMR is an educational resource and the "Train-the-Trainer" is an educational course. The EMR is a recourse for teaching biosecurity issues, while the "Train-the-Trainer" programme is a university-accredited educational course that aims to "create experts who can improve the utility of the existing EMR" (Minehata et al., 2013). The "Train-the-Trainer" aims to foster engagement with biosecurity rather than biosafety. This is important as biosecurity requires an understanding of the dual-use nature of research, while biosafety only asks the question if the researcher themselves are safe. The course also aims to encourage "participants to bring their ideas and experiences to the course" to teach them to contextualize DURC examples for their students (Minehata et al., 2013). Thus the educational resource and the educational course reinforce each other.

3.6 LNCV Course

Furthermore, the EMR has also been tested through its implementation by the Landau Network – Centro Volta (LNCV) in Italy. The module was tested as a series of seminars that were led by invited excerpts as well as in the form of a working group for students at both the undergraduate and graduate levels (Revill et al., 2012). The seminars "involved more than 100 students" with strong educator participation and a positive response on the feedback questionnaires (Revill et al., 2012). These courses also highlighted the importance of emphasising the "active role" that scientists have in creating a "culture of responsibility rather than risk" to make the discussions more constructive and to avoid "creating prejudices and generating a climate of suspicion towards science and research" (Revill et al., 2012).

3.7 QAU Course

An alternative to the EMR is currently being tested at Quaid-e-Azam University (QAU) in Pakistan. It is a course aimed at the graduate level titled "Bioethics, biosafety, biosecurity and dual-use education" (Revill et al., 2012). The course covers an understanding of ethical dilemmas and prepares students with the tools to assess and address the ethics in biotechnological research. It is also a core subject requirement for all the students pursuing the life sciences at QAU (Revill et al., 2012). It is important to note that this course takes into account the context of the students at also provides an "understanding of the ethical discourses that underpin Muslim and Western approaches to moral philosophies" (Revill et al., 2012).

3.8 Karolinska Course

The last example of a biosecurity education strategy is a series of seminars and the development of a national policy network in Sweden (Revill et al., 2012). The awareness-raising seminars were held in 2009 at Umea University, Karolinska Institutet and Uppsala University. They were organized by the Swedish Defence Research Agency (FOI) and "conducted by two invited experts, Professor Malcolm Dando … and Professor Brian

Rappert" (Revill et al., 2012). The paper provided little detail on the specifics of the content covered but stressed the importance of personal connections in fostering the informal network with the aim of "developing a national education approach" (Revill et al., 2012). The network has identified the Centre for Research Ethics and Bioethics (CRB) as a "unique resource for the network" which will be able to act as the centre point for "further development of topics in bioethics, dual-use and biosecurity education"(Revill et al., 2012). The CRB Website provides a few examples of courses which are currently being offered.[2] It is somewhat difficult to assess the effectiveness of this education strategy, but a the very least it has created an increase in awareness of the state of the biosecurity initiatives throughout Swedish institutions.

4. Discussion

Prior to proceeding with a discussion of the strengths and weaknesses of the specific interventions, it is important to note a weakness in the methodology of this research. The initial search on biosecurity education techniques had limited scope, focusing mainly on review papers. Further exploration of a wider range of educational approaches and their effectiveness is needed for a comprehensive understanding of this topic.

The overall success criteria which were identified in the eight interventions were: length, context, personal connection and instructor excellency. The UVienna course and the Local–People–Make Sense Framework had a broader reach but lacked collaborative projects and expert engagement that seemed to lead to success in the other interventions. The Karolinska Course also had a broader approach but was more successful in bringing in biosecurity experts and creating a biosecurity network, though its effectiveness is difficult to assess. The EMR and the Train-The-Trainer Program both had detailed content that was customizable to the particular audience at hand, though the latter had limited information on the effectiveness of its implementation. The effectiveness of the EMR was supported by its examples of implementation at the NDMC and LNCV, where it was tailored to audiences with varying depths of scientific knowledge. The LNCV course in particular, engaged students through seminars and took a more active approach, which was similar to the approach used by the QAU Course that integrated discussions on bioethics, biosafety, and biosecurity into the course in order to increase student engagement.

The implications of the findings shed light on the understanding of biosecurity education approaches, where the integration of biosecurity into existing courses can reach a wider audience. Whereas customizable resources, expert speakers and personal connections increase the engagement level of the students during the course.

5. Conclusion

Following the description of the eight approaches towards biosecurity dual-use research education, several key themes emerge; length, context, personal connection and instructor excellency. The length of the programs across the board varied from a few hours to a few weeks with the most important determining factor for this being context. The significance of context and tailoring the education approach to the target audience was present in the evaluation of every intervention. The context also affected the student's ability to uptake the material as well as to lead positive discussions. Personal connections between the students and the educators were crucial to both the set-up of the courses as well as their efficacy, naturally this ties in with the value of skilled instructors as both disseminators of knowledge and the regulators of cross-institutional networks such as the one in Sweden.

[2] https://www.crb.uu.se/education/

Further research on this topic could involve the testing of the frameworks such as the EMR and Train The Trainer at other institutions and improving their ability for customisation in length and content difficulty. A network of "highly skilled biosecurity educators" could also be established by continuing to foster the connections between institutions such as Bradford University, NDMC, LNCV, QAU and CRB. These key experiments will provide us with a better comprehension of how to encourage and scale responsible research practices to decrease the possibility of disastrous consequences arising from the exploitation of dual-use research.

References

Atlas, R. (2005). Biosecurity concerns: Changing the face of academic research. Chemical Health and Safety, 12, 15–23. https://doi.org/10.1016/j.chs.2004.10.004

Beeckman, D. S. A., & Rüdelsheim, P. (2020). Biosafety and Biosecurity in Containment: A Regulatory Overview. Frontiers in Bioengineering and Biotechnology, 8. https://www.frontiersin.org/articles/10.3389/fbioe.2020.00650

Dickmann, P., Sheeley, H., & Lightfoot, N. (2015). Biosafety and Biosecurity: A Relative Risk-Based Framework for Safer, More Secure, and Sustainable Laboratory Capacity Building. Frontiers in Public Health, 3. https://www.frontiersin.org/articles/10.3389/fpubh.2015.00241

Educational Module Resources (EMR)—Bioethics. (n.d.). University of Bradford. Retrieved April 16, 2023, from https://www.bradford.ac.uk/bioethics/educational-module-resources-emr/

Edwards, B., & Kelle, A. (2012). A life scientist, an engineer and a social scientist walk into a lab: Challenges of dual-use engagement and education in synthetic biology. Medicine, Conflict and Survival, 28(1), 5–18. https://doi.org/10.1080/13623699.2012.658659

Executive Office of the President. (2022). Budget of the U. S. Government, Fiscal Year 2023. Rowman & Littlefield Publishers, Incorporated. (n.d.).

Greene, D., Palmer, M. J., & Relman, D. A. (2023). Motivating Proactive Biorisk Management. Health Security, 21(1), 46–60. https://doi.org/10.1089/hs.2022.0101

House, T. W. (2022, March 28). FACT SHEET: The Biden Administration's Historic Investment in Pandemic Preparedness and Biodefense in the FY 2023 President's Budget. The White House. https://www.whitehouse.gov/briefing-room/statements-releases/2022/03/28/fact-sheet-the-biden-administrations-historic-investment-in-pandemic-preparedness-and-biodefense-in-the-fy-2023-presidents-budget/

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

Minehata, M., Sture, J., Shinomiya, N., & Whitby, S. (2013). Implementing Biosecurity Education: Approaches, Resources and Programmes. Science and Engineering Ethics, 19(4), 1473–1486. https://doi.org/10.1007/s11948-011-9321-z

Nixdorff, K. (2013a). Education for Life Scientists on the Dual-Use Implications of Their Research. Science and Engineering Ethics, 19(4), 1487–1490. https://doi.org/10.1007/s11948-013-9478-8

Nixdorff, K. (2013b). Education for life scientists on the dual-use implications of their research: Commentary on "implementing biosecurity education: approaches, resources and programmes." Science and Engineering Ethics, 19(4), 1487–1490. https://doi.org/10.1007/s11948-013-9478-8

Palmer, M. J., Fukuyama, F., & Relman, D. A. (2015). A more systematic approach to biological risk. Science, 350(6267), 1471–1473. https://doi.org/10.1126/science.aad8849

Research Rabbit. (n.d.). Retrieved April 15, 2023, from https://researchrabbitapp.com/home

Revill, J., Carnevali, M. D. C., Forsberg, Å., Holmström, A., Rath, J., Shinwari, Z. K., & Mancini, G. M. (2012a). Lessons learned from implementing education on dual-use in Austria, Italy, Pakistan and Sweden. Medicine, Conflict and Survival, 28(1), 31–44. https://doi.org/10.1080/13623699.2012.658624

Revill, J., Carnevali, M. D. C., Forsberg, A., Holmström, A., Rath, J., Shinwari, Z. K., & Mancini, G. M. (2012b). Lessons learned from implementing education on dual-use in Austria, Italy, Pakistan and Sweden. Medicine, Conflict, and Survival, 28(1), 31–44. https://doi.org/10.1080/13623699.2012.658624

University, © Stanford, Stanford, & California 94305. (n.d.). The Biorisk Management Casebook: Insights into contemporary practices. Retrieved June 21, 2023, from https://purl.stanford.edu/hj505vf5601

World Health Organization. (2006). Biorisk management: Laboratory biosecurity guidance (WHO/CDS/EPR/2006.6). World Health Organization. https://apps.who.int/iris/handle/10665/69390

Stanford

Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Existential Risks Associated with Dual-Use Technologies

Ashok Vaseashta [1*]

Citation: Vaseashta, Ashok. Existential Risks Associated with Dual-Use Technologies. *Proceedings of the Stanford Existential Risks Conference* 2023, 156-170. https://doi.org/10.25740/zy474yf0050

Academic Editor: Trond Undheim, Dan Zimmer

Funding: N/A

Conflict of Interest Statement: N/A

Informed Consent Statement: N/A

Acknowledgments: N/A

Author Contributions: N/A

Abstract: Dual-Use Research of Concern articulates research and innovation that is intended to provide useful knowledge, information, or products; however, it could also be misused to present existential risks to society, with broad potential consequences. Thus far, it has been considered in the context of nuclear and biological research, however, the principles can be applied to other fields. With the emergence of new and exponential technologies, there is a significant potential to enrich our lives, however, the risks associated with their use for nefarious purposes remain inextricable from their benefits. A comprehensive overview of existential risks and threats posed by dual-use technologies adaptation is outlined in this article. A special emphasis is placed on information technology and artificial intelligence which has transformative implications and has produced a wide range of capabilities to stay ahead of threats. Unfortunately, in an open society, cybercriminals take advantage of the same technology for malicious purposes. Hence, in the new and complex security environment, these new technologies must be examined carefully, especially regarding their potential dual-use nature. Furthermore, extracting threats from the benefits of exponential technologies is a complex undertaking, and is addressed in this report to lay the groundwork with the intent to curb dual-use technology proliferation.

Keywords: dual-use technologies, catastrophic risk, existential risk, threat intelligence, nano-biotechnology

[1] Researcher/Professor, Office of Applied Research, International Clean Water Institute, 9108 Church Street, P. O. Box 258, Manassas. VA 20110, USA; prof.vaseashta@ieee.org

* Correspondence: prof.vaseashta@ieee.org

1. Introduction: Framing the Challenge

In the last few decades, the war and battlefield theatre dynamics and their modus operandi have changed significantly [Vaseashta et al. 2014]. The current battlefield is far more complex and may not be, even, a physical and geographical space. The adversaries are either inviable, unknown, or perhaps, even "among us", and may even be betting on the "red team". The overall conflict construct has gradually morphed from conventional, to unconventional and now to a hybrid modality, which has many definitions but mainly refers to unrestricted warfare, i.e., using all available means to accomplish political and military objectives. There are no rules: no chivalry, no ethics, no Geneva Convention (United Nations, 1949), hence everything goes, a.k.a. the "jungle rules". This includes cyberspace, chemical-biological weapons, improvised explosive devices (IEDs)/incendiary devices, dirty bombs containing traces of nuclear materials, electromagnetic pulses (EMP), and even artificial intelligence (AI) based psyops – psychological operations. The addition of nanobiotechnology to produce synthetic biological weapons with highly transmissive and contagious capabilities, render such threats more potent, especially while deployed in conjunction with neuropharmacology [Qazi et al. 2019] and chemogenomics [Yang et al., 2021]. While the former impacts how nanosized drugs affect cellular-level functions in the nervous system, the latter may be used to deconstruct behavior, as a neuroweapon. As an example, a Receptor Activated Solely by a Synthetic Ligand (RASSL), or Designer Receptor Exclusively Activated by Designer Drugs (DREADD) can potentially be used for G-protein-coupled receptors (GPCRs), built specifically to allow for precise spatiotemporal control of GPCR signaling, which will allow for the new forms of controlling engineered tissues *in-vivo* [Iverson et al., 2013] and potentially *in-vitro*.

From the realms of informational technology, capabilities such as networked computers, the internet, supercomputers, and Internet-of-everything (IoT) provide exploits and vulnerabilities across the political, military, economic, social, informational, and infrastructure spectrum regarding cyber-security threats. The spectrum of threat vectors has evolved further into distributed denial of service (DoS) attacks, intrusion of the electronic virus (viral signature) into the networks to produce pre-calculated viral effects on the networks and associated local physical hardware, exploitation through social media for transmission of misinformation (such as propagation of fake news for political gains and dominance, deep fake videos for misinformation/disinformation dissemination, resistance movements), and even for human exploitation. As a historical example, Maskirovka (Russian: маскировка, lit. 'disguise') – is a Russian military deception that represents a military doctrine developed from the start of the 20th century. The doctrine covers a broad range of measures for military deception, from camouflage to denial and trickery that include concealment, imitation with decoys and dummies, and maneuvers intended to deceive, deny, and propagate disinformation. The 1944 Soviet Military version outlines such doctrine akin to "*means of securing combat operations and the daily activities of forces; a complexity of measures, directed to mislead the enemy regarding the presence and disposition of forces...*". Later versions of the doctrine included strategic, political, and diplomatic means including manipulation of "the facts", situations, and perceptions to affect the media and opinion around the world, so as to achieve or facilitate tactical, strategic, national, and international goals.

The doctrine has also been deployed into practice in peacetime, with denial and deception operations in such events as the Cuban Missile Crisis, the Prague Spring (Frommer, 2022), and the annexation of eastern regions of Ukraine. Recent advances in computational

sciences and algorithm-based data compression, and pattern recognition have led to the use of AI, machine learning (ML), and deep learning. These technologies can swiftly analyze millions of data sets to track any embedded malicious codes and cyber threats. AI and ML allow the identification and tracking of cyber-criminals and automated threat detection more effectively than conventional software-driven or manual techniques. AI-based cybersecurity systems provide the latest knowledge of global as well as industry-specific threats to formulate vital decisions-based prioritization, not merely based on what could be used to target systems but also, based on what is the most likely target. Unfortunately, cybercriminals also have access to the same platforms, making it a constant challenge to thwart adversarial cyber-attacks and counter threats. Other threat vectors include the electromagnetic spectrum in the form of electromagnetic pulse (EMP), which can render systems ineffective if used towards the electrical grid, aviation, or transportation systems. Hence, the real-world threats landscape is far more complex now than ever before. Even with the prevalence of exponential use of technologies to improve quality of life, global catastrophic and existential risks to human survival from emerging, hybrid, and dual-use threats have gained importance to the point of potential civilizational collapse or even human extinction. Major concerns are the use of AI, nuclear materials, natural and synthetic chemical and biological materials, big data, and data analytics, cyber-physical systems, advances in pharmaceuticals, exponential use of plastic, and ever-shifting weather patterns, a.k.a. climate change. One thing that is common among many of these cascading and existential risks (X-risks) is that with the beneficial use of every technology, there is a finite probability of its adverse impact in one form or the other.

Before delving further into this topic, it is instructive to understand the scope and general definition of dual-use technologies (Pandya & Cognitive World, 2019). In politics and diplomacy, dual-use is a technology that can be used for both peaceful and military ambitions. In general, dual-use goods are products and technologies normally used for civilian purposes but may have military applications. Additionally, research that, based on current understanding, can be reasonably anticipated to provide knowledge, products, or technologies that could be directly misapplied by others to pose a threat to public health, agriculture, plants, animals, the environment, or material, according to National Science Advisory Board for Biosecurity, is also classified as dual-use research. Modern concerns about dual-use technologies emerged with fears about the proliferation of nuclear weapons-related technologies in the early days of the Cold War. More recently, a group of stakeholders has initiated efforts to prevent the biological sciences from being used to develop weapons whose destructive effects against humans, animals, and plants could, under certain circumstances, rival those of nuclear weapons. The focus has now shifted to nonnuclear technological contexts—including, information technology—in which ongoing research and development has dramatically advanced human social and economic well-being, though at the cost of generating potential to be harnessed for nefarious purposes. In addition, a few other technologies, such as unmanned aerial vehicles (UAVs) and additive manufacturing, have tremendous advantages for the common good, but also have been used for unintended uses, and are discussed below.

2. Spectrum of Dual-Use technologies

In a conventional war, the battle is conducted with conventional military weapons and battlefield tactics between two or more states in an open confrontation. The weapons (or means) primarily deployed target the opposing army and the war is fought using conventional weapons, not including chemical, biological, and nuclear weaponry,

improvised explosive devices (IEDs), and/or the cyber domain. The main objective is to weaken or destroy the opponent's military capabilities. Standard rules of engagement (ROE) apply, and conventions are followed during and after the war, if officially concluded. Unconventional warfare (UW), on the other hand, is an attempt to achieve military victory through acquiescence, capitulation, or clandestine support for one side of an existing conflict. On the surface, UW contrasts with conventional warfare in that forces or objectives are covert or not well-defined, tactics and weapons intensify subversion or intimidation, and the general or long-term goals are coercive or subversive to a political body. Unconventional warfare, by nature, is an asymmetric operation, which is without any ROE, covert, poses technology surprise, inexpensively allows multiple simultaneous and coordinated attacks, and more importantly, the adversary is not necessarily opting for a successful outcome. From a tactical standpoint, such scenarios do not involve any design/development of tabletop exercises (TTX) activity, best practices playbook on capacity building, or an all-out preparedness for such "unknown" events, and hence they are more difficult to predict, track, and eradicate.

A low-intensity conflict (LIC), on the other hand, is defined by the U.S. Army, as a political-military confrontation between contending states or groups. Generally, LIC is below conventional war and above routine, and it involves protracted struggles of competing principles and ideologies. It is waged by a combination of means, employing political, economic, informational, and military instruments, and are often localized, generally in the Third World, but contain regional and global security implications. Unconventional warfare using the cyber domain, is far below LIC and yet needs an appropriate classification, such that international laws - such as those applicable to Cyber Warfare as described in the Tallinn Manual (Schmitt, 2017), may be applied. Furthermore, the countering of UW requires a set of tools, such as predictive intelligence (Thorrison et al. 2020) and decision support tools [Vaseashta et al. 2014, 2016, 2018] to align a level of countermeasures consistent with the nature of the conflict. It is far too often the case where there is a mismatch between the offensive and counter-offensive. Decision support and practice intelligence tools provide a balance between action and reaction, providing necessary balance. In addition, AI is deployed in behavioral patterns to learn how outside entities work within secure environments and also for the "observe, orient, decide and act" (OODA) loop (Pearson, 2022), as a decision-making process to enhance security posture. Due to dependence on intention, dual-use technologies may be regarded as unconventional and border LIC, and may also be compared with hybrid threats.

2.1 Nuclear

During the early days of the cold war, the concern about dual-use technologies emerged due to fear of the proliferation of nuclear weapons-related technologies. The major threat was due to stolen fissile materials – such as Cobalt (^{68}Co) for dirty bomb applications. There were several known cases of nuclear material poisoning, viz. Polonium (^{210}PO), allegedly against Mr. Alexander Litvinenko (Cotton, 2016), Mr. Yasser Arafat (Lallanilla, 2013), and Mr. Lal Bahadur Shastri (Vereshchagin, 2013). In terms of dual use, separated plutonium and highly enriched uranium (HEU) are the two nuclear materials of greatest proliferation significance. Both have a few nonmilitary applications, most notably in reactor fuel, but can also be used as fuel for a nuclear weapon. Uranium "enriched" to >20% of the isotope ^{235}U is capable of sustaining the uncontrolled chain reaction used to generate energy in a nuclear weapon, while practical warhead-making considerations dictate the use of HEU with an enrichment level of ~>80%. The International Atomic Energy Agency (IAEA) treats almost all plutonium as a weapon capable. Enrichment and

reprocessing are the most sensitive nuclear technologies and require tremendous capital, hence are less prone to proliferation but still remain dual-use because they can be involved in the production of weapons-grade materials. Some materials and technologies that do not involve uranium or plutonium have proliferation concerns as well. Those that also have nonnuclear applications are often described as "dual use." This usage of the term is subtle but confusingly different from describing plutonium as dual use because it has both civilian and military applications. Hence the primary concern remains the use of low-grade material for "dirty bombs" from facilities with low-level security, such as medical facilities and universities.

2.2 Chem.-Bio Threat Vectors

There is much emphasis on dual-use technologies at present. Historically, however, the "duality of science" (more specifically the "duality of Chemistry") existed for a long time. During WW1 (1914-18) – the idea of producing chemicals for explosives was for civil purposes only. However, the use of Ammonia was exploited for nitro-based explosives. Similarly, the production of acetone by fermentation of sugar – later became ABE (Acetone, Butanol, Ethanol) fertilizers for producing higher concentrations of Sulphuric acid. Some other chemicals were produced to make tear gas, viz. chlorine, phosgene, and diphosgene. Phosgene is a highly poisonous gas and was produced by the production of polycarbonate as a starting monomer for polyurethanes. Similarly, Hyprite – a.k.a. Mustard gas, although first produced in 1820, was studied as an anticancer agent. Lewisite produced by acetylene and arsenic trichloride – was extremely toxic but later was synthesized as neoprene–synthetic rubber. Hydrogen cyanide (HCN) was used in grenades but is currently used as a pesticide and insecticide. With the technological advances in nanomaterials (Vaseashta, 2012), there are many instances of its dual use. The use of nanomaterials for biological applications is described below, however, the use of nanomaterials in Nanoenergetics (combination of Fe and Al oxides) posed severe threats and was suspected of enhancing explosive reactions in Word Trade Center (Teschler, 2012). Certain organometallics, Tetramethyltin (TMT) as an example, have great optical characteristics for window glasses but present severe neurological consequences (Hamilton et al, 1988).

The next largest dual-use research concern (DURC) arises due to synthetic biology. With the introduction of DIYBIO toolkits, non-profit foundations such as iGEM (International Genetically Engineered Machine) (iGEM, 2023), BioBrick (Lego-based Synthetic biology toolkit to code RNA or Protein using elementary DNA sequence by abstraction and modularization) (Biobrick Foundation, 2023), and SynBERC (Synthetic Biology Research Center) EBRC (Engineering Biology Research Consortium) (SYNBERC, 2023) consortium, since there has been an enormous interest to build "desired" biological components and assembling them into integrated systems to accomplish specific tasks. In the broadest sense, synthetic biology is the deliberate modification of cells, organisms, populations, or major sub-systems thereof. The platforms mentioned above, such as iGEM, BioBrick, and SynBERC provide academic forums to create biological systems that operate in living cells. Despite their potential, the major drawback of these, otherwise academic social interest, groups is the engagement of students/learners/special interest groups with a possibility of "not so social" intentions. The categories and characteristics of biological agents are provided in Table 1.

As much as the emphasis has been on projects that have developed the foundational understanding and technologies needed to build large numbers of useful biological

systems from standard interchangeable parts, however, the technological innovations are not limited to beneficial uses ONLY! Undoubtedly, the potential of synthetic biology ranges from large-scale drug synthesis, biofuel production, environmentally friendly "green" chemical manufacturing, and non-allergenic, drought-resistant genetically modified crops (Seleiman et. al., 2021), advanced sensor systems (Vaseashta, 2012), "attack micro-organisms" targeted applications, gene-editing to manipulate specific functions, and stem cell development, however, synthesis capabilities far exceed design capabilities. Certain experimental projects such as minimum genome (Mizoguchi et al. 2007) and neuropharmacology (Garner, 2021) have heightened security concerns due to the ease of creating such organisms with a minimum set of genes for metabolism and replication and for remote control of neurons. Hence, with the enormous potential offered by synthetic biology, the potential of risk is equally (in fact more) real. Hence the concern for dual-use research presented by synthetic biology is enormous, especially in conjunction with nanotechnology due to unknown or unclear *in-vivo* genotoxicity and carcinogenicity.

Table 1: Categories and characteristics of biological agents

Cat.	Priority Rating	Characteristics	Biological Agents
A	Agents that pose a threat to national security bacause they -	▪ Can be easily disseminated or transmitted person-to-person ▪ Cause high mortality, with potential for major public health impact ▪ Might cause panic and social disruption ▪ Require special public health	Variola major (smallpox), *Bacillus anthracis* (anthrax), *Yersinia pestis* (plague), *Clostridium botulinum toxin* (botulism), *Francisella tularensis* (tularemia), *Filoviruses*, Ebola hemorrhagic fever, Marburg hemorrhagic fever, Arenaviruses, Lassa (Lassa fever), Junin (Argentine hemorrhagic fever), ...
B	Second highest priority agents that -	▪ are moderately easy to disseminate ▪ cause moderate morbidity and low mortality ▪ require specific enhancements of CDC's diagnostic capacity and enhanced disease surveillance	*Coxiella burnetti* (Q fever), Brucella species (brucellosis), *Burkholderia mallei* (glanders), Alphaviruses, Venezuelan equine encephalomyelitis eastern / western equine encephalomyelitis, Ricin toxin from *Ricinus communis* (castor bean), Epsilon toxin of Clostridium perfringens, *Staphylococcus enterotoxin B*. Food and water borne - Salmonella species, Shigella dysenteriae, Escherichia coli O157:H7, Vibrio cholerae Cryptosporidium parvum
C	Third highest priority agents include emerging pathogens that could be engineered for mass dissemination in the future because of -	▪ availability ▪ ease of production and dissemination ▪ potential for high morbidity and mortality and major health impact	Nipah virus, Hantaviruses, Tickborne hemorrhagic fever viruses, Tickborne encephalitis viruses, Yellow fever, Multidrug-resistant tuberculosis, and others not listed here.

2.3 Information Technology as Dual-Use Research Concern

Recent advances in computational sciences and algorithm-based data compression, and pattern recognition have led to the use of AI, ML, and deep learning. These technologies can swiftly analyze millions of data sets to track any embedded malicious codes and cyber threats. AI and ML allow the identification of cyber-criminals and automated threat detection more effectively than conventional software-driven or manual techniques. AI-based cybersecurity systems provide the latest knowledge of global as well as industry-specific threats to formulate vital decisions-based prioritization, not merely based on what could be used to target systems but also, based on what is the most likely target. Unfortunately, cybercriminals also have access to the same platforms, making it a constant challenge to thwart adversarial cyber-attacks and counter hybrid threats. Today, questions are being raised about how to manage the potential threat posed by information technology (IT), as the growth and spread of IT, some believe may position cyber weapons, parallel to nuclear and biological weapons capable of unleashing massive harm. Hence, the scope of dual-use technology has now expanded to information and computational technologies.

2.4 Artificial Intelligence: Pros, Cons, and Dual-Use

The rapid progress and developments in AI are prompting grave speculation about its dual-use applications and cascading security risks, despite the fact that AI offers numerous advantages and benefits across multiple domains. Notable advantages of AI include automation thus reducing human effort and increasing efficiency; handling complex calculations, data analysis, and decision-making processes at a much faster pace and with accuracy leading to more strategic and creative activities; processing vast amounts of data to provide valuable insights to aid in decision-making by analyzing patterns, trends, correlations in data; and in rendering data-driven decisions. AI systems are especially beneficial in sectors that require constant monitoring or rapid response times, such as cybersecurity or customer support. AI algorithms can analyze user preferences, behavior, and historical data to provide personalized experiences and recommendations. This is particularly helpful in areas such as fraud detection, predictive maintenance, and market analysis. AI-powered chatbots and virtual assistants can provide instant and personalized customer support, improving response times and overall customer satisfaction. AI can also be used to develop intuitive user interfaces, making products and services more user-friendly. AI techniques, such as machine learning and deep learning, enable the exploration and analysis of complex datasets that are difficult for humans to process manually, which opens new possibilities for scientific research, healthcare diagnostics, and uncovering hidden insights.

While AI brings numerous advantages, there are also several disadvantages and challenges associated with its implementation. One of the primary concerns is its potential to automate jobs, leading to job displacement. AI algorithms operate based on predefined rules and patterns, lacking human reasoning and intuition. This can limit their ability to handle complex or unexpected situations that require contextual understanding, empathy, or ethical decision-making. AI algorithms are trained on large datasets, and if those datasets contain biased or discriminatory information, the AI systems may learn and perpetuate those biases leading to unfair outcomes or discriminatory practices. With the increased integration of AI into various aspects of our lives, there is an increased risk of security breaches and privacy violations. AI systems may be vulnerable to attacks, and the collection and analysis of vast amounts of personal data raise concerns about data protection and misuse. Some AI algorithms, such as deep learning neural networks, can be complex and difficult to interpret. This lack of transparency raises concerns about accountability and the ability to understand and explain the decisions made by AI systems. As our reliance on AI systems grows, there is a risk of becoming overly dependent, and should these systems malfunction, encounter unexpected situations, or are inaccessible, it can disrupt operations and negatively impact productivity, particularly if backup systems or human alternatives are not readily available. AI raises a range of ethical dilemmas, including issues around privacy, transparency, bias, and potential misuse for malicious purposes. Determining the ethical guidelines and regulations for AI development and deployment remains a significant challenge. While AI excels at tasks that involve data analysis and pattern recognition, it often lacks creativity, intuition, and the ability to think abstractly. As AI systems become more autonomous, questions arise regarding the ethical and legal responsibility for their actions. Determining liability and accountability in cases of AI-related accidents, errors, or harm can be complex and may require the establishment of new legal frameworks.

The empirical risks associated with AI include adversarial attacks involving deliberately manipulating or fooling AI systems by introducing subtle changes to input data. Attackers

can exploit vulnerabilities in AI algorithms, such as image recognition or natural language processing, to deceive the system and generate incorrect results. This poses greater risks in critical domains like autonomous vehicles, facial recognition systems, and cybersecurity. Yet another security risk is data poisoning, i.e., intentionally tampering with or manipulating training data to inject biased or malicious information to manipulate the learning process and compromise the integrity and reliability of AI models. Furthermore, AI models and algorithms are valuable assets, and their theft can have severe consequences, such as attackers' ability to reverse engineer, gain unauthorized access, or perpetrate insider threats. Any unauthorized access to AI algorithms can be exploited or used to gain a competitive advantage or launch malicious attacks, and lead to privacy concerns as sensitive or personal information can be collected, analyzed, and potentially misused. The cascading risks include a lack of explainability since it becomes intriguing to understand why an AI system made a particular decision. Other risks include backdoor attacks where malicious actors manipulate the training process to introduce hidden vulnerabilities triggered by specific inputs or conditions, data inference by subtly modifying the training data or introducing strategically crafted inputs to influence the system's behavior, Distributed Denial of Service (DDoS) where malicious actors overload the system with excessive requests, causing it to become unavailable or slow down, social engineering to manipulate the system to extract sensitive information or deceive users, and creating a sense of undermining trust and confidence in the technology.

Organizations must ensure the reliability, integrity, and security of AI systems through robust authentication mechanisms, secure development practices, and regular security audits. Addressing these security challenges requires a comprehensive approach, including secure AI model development, data protection measures, robust authentication, and access controls, regular security assessments, and ongoing research to mitigate emerging threats. The challenges and complexities of evolving AI threats and security have crossed the barriers of space, ideology, and politics, necessitating a constructive collaborative effort of all stakeholders across nations. With the exponential advancements in information science and technology and associated subjects such as AI, ML, DL, DA, and big data, unfortunately, misuse of IT is now also classified as dual-use, since it is evident by exploits of social media by terrorists for recruitment and manipulation of the weak and vulnerable for promoting terrorist ideology. Other risk areas include (potentially AI-generated) misinformation and disinformation – feeding the rise of fascism, racism, and authoritarianism while undermining reality-based governance. Information science and technology play a major role in cyber defense, since adversaries always look for vulnerabilities in security posture, while cyber-defenders seek the same vulnerabilities from the viewpoint of either "plugging" them or finding a patch to deter penetration. Due to the vast spectrum of threat vectors arising from cyber, it is critical to use advanced technologies, computational tools, foresight, and decision support tools for cyber defense to ensure security and resiliency.

With the growing use of AI, there is even a threat to democracy by platforms such as ChatGPT (Chat Generative Pretrained Transformer) – an AI-powered Chatbot built on top of OpenAI's GPT-3.5 and GPT-4 families of large language models and has been fine-tuned using both supervised and reinforcement learning techniques, there are already discussions concerning bias in programming (ChatGPT, 2023; Buchholz, 2023; Sheth, 2023; Dastin; 2023). The potential implications are election interference, indoctrination of the younger generation, the spread of misinformation, and propaganda arms for nefarious groups. AI techniques have already generated recipes for thousands of toxins similar to

chemical warfare agents; guerilla armed forces, terrorist groups, or even individuals deploy these tools to construct deadly pathogens by modifying genomes.

2.5 Virtual and Augmented Reality

AI and virtual reality (VR) are two distinct technologies, but they can complement each other and enhance the overall user experience when combined. AI contributes to VR by enhancing natural language interaction, enabling computer vision capabilities, automating content generation, personalizing experiences, creating intelligent virtual agents, providing real-time analytics, and enabling realistic training and simulation scenarios. These synergies help create more immersive, interactive, and personalized VR experiences. VR and augmented reality (AR) are related technologies but have distinct differences. However, VR can play a role in the development and evolution of AR. While VR and AR have distinct purposes and use cases, the advancements made in VR technology, user experience design, content creation, and interaction paradigms can serve as a foundation for the development and evolution of AR experiences, enabling more immersive, interactive, and contextually rich applications. However, the integration of VR and AR presents several security concerns that need to be addressed to ensure user safety and protect against potential risks. To address these security concerns, VR and AR developers and platform providers should prioritize robust security measures, conduct regular security audits, and follow industry best practices. Collaboration with cybersecurity experts, adherence to privacy regulations, and transparent communication with users about data collection and usage can build trust and ensure safe integration.

2.6 Other Dual-Use Technologies

2.6.1: Unmanned Aerial Vehicles

More recent concerns about dual-use technologies consist of UAVs, as these autonomous and unmanned aerial vehicles are a great invention. The author has used drone-based platforms for emergency rescue missions, remote sensing, sample collection from otherwise inaccessible locations, agriculture, and aerial surveillance, however, there are many instances where these unmanned aircraft have been used with many negative consequences and thus it was necessary to use restrictions for its deployment. There have been reports for aerial chemical dispersions, unauthorized surveillance, inappropriate collection of data, and delivery of IEDs, drugs, etc. The use of drones at the Southern border of the United States by cartels is a common feature where the gangsters are trying to evade law enforcement to smuggle illegal migrants into the US. They also use these drones to map paths for human trafficking and drug delivery. War in Ukraine is a classic example where suicide drones have wreaked havoc in certain oblasts.

2.6.2: Electronic Warfare (EW)

Department of Energy has focused on research, preparedness, response, and recovery activities related to potential threats to the nation's critical energy infrastructure from electromagnetic pulses (EMPs) for a long time. EMPs can be caused by a high-altitude detonation of a nuclear device. This type of High impact, low-frequency (HILF) event can destabilize the nation's power grid and damage equipment such as large power transformers critical to the nation's power grid. While the likelihood of such an event is low, it is critical to understand the risks associated with such pulses.

A nuclear EMP is more energetic and has a shorter burst. A solar flare EMP may also be referred to as a Coronal Mass Ejection (CME) or a geomagnetic storm. Solar flares vary widely in intensity from simply causing bright "northern lights" to potentially destroying some or all of the power grid. Smaller EMPs may cause power grid blackouts. The non-lethal nature of electromagnetic weapons finds its use far less politically damaging than that of conventional munitions and therefore broadens the range of military options available. For weapons purposes, EMP-producing sources other than nuclear detonations, have been successfully developed. Several nations have developed non-nuclear bombs capable of generating EMPs. Electromagnetic bombs (E-bombs) are specialized, non-nuclear tools designed to destroy information systems. These devices are primarily intended for battlefield application, and their effects would be restricted to a relatively small area. An EMP shock wave can be produced by a device small enough to fit in a briefcase. High Power EMP generation techniques and high-power microwave technology have matured to a point where practical E-bombs are becoming technically feasible, with new applications in both strategic and tactical warfare. Although much of this work is classified, it's believed that current efforts are based on using high-temperature superconductors to create intense magnetic fields. The development of conventional E-bomb devices allows their use in non-nuclear confrontations to defeat an enemy without causing loss of life. Regardless of the method of delivery, experts agree that EMPs can be powerful enough to cripple electronic wiring and circuitry over a geographic area as large as several square miles, posing a real threat to the nation's critical infrastructure. In addition, the reliance on satellites and commercial computer equipment to conduct real-time command and control functions around the world also puts military operations at risk. Depending on the power of the explosion or solar flare, an EMP could disable, damage, or destroy TVs, radios, and other broadcast equipment; power grid transformers and substations; telephones and smartphones; vehicle and aircraft control systems; computers and all internet-connected devices; refrigerators; generators; and satellites within the range of the EMP.

3. Discerning Inextricable Existential Risks and Security Concerns

Due to our reliance on technology, our own strengths and capabilities can become weaknesses and blind spots, and potential pathways for the future of existential risks. Discerning risks from opportunities depends on intentions – which are neither quantifiable nor measurable. There are endless academic discussions on measuring intentions, however, in the security arena any incorrect assessment could result in loss of life and hence countering such threats by technological means is the best option. However, academic research which started with the Technology Acceptance Model (TAM) (Davis, 1989), is a great piece of work from the late '80s that enabled the measurement of 'behavioral intention', which is (understandably) a person's intention to use a certain technology (Venkatesh and Davis, 2000). In the original and subsequent uses of the TAM, it was identified that "Perceived Usefulness" and "Perceived Ease of Use" were the main influences on a user's behavioral intention. These can be translated as Perceived Usefulness (PU) – i.e. what does the technology do, and does it help me? Perceived Ease of Use (PEoU) – refers to how effortless is the beneficial outcome to achieve with the technology? Several thousands of studies have since found that if PU and PEoU are high, users have a higher behavioral intention – which generally translates into the actual use of the technology.

3.1 Countering Such Threats

Countering threats, especially when it is based on unknown intentions, is a tremendous undertaking. In addition to technological platforms, it is critical to enhance system resilience to mitigate threats. Being resilient is important because no matter how well a system is engineered, Murphy's law always prevails in any of its functions, such as availability, capacity, interoperability, performance, reliability, robustness, safety, security, and usability. Due to inevitable disruptions, availability and reliability become insufficient; hence, a system must be resilient. Furthermore, the system must recover rapidly to maximum capacity from any harm that those disruptions might cause. The pathways for resilience against hybrid threats should be augmented through several interdependent pillars: viz. multidisciplinary technology, policy implementation, the whole of the government approach, and educational platforms. It is beyond the scope of this publication to outline details of all available means to counter dual-use threats, however, it is intuitive to mention a few platforms to mitigate dual-use threats and increase systemwide resilience.

3.1.1: Use of Technological Platforms

To enhance resilience, the use of technological platforms is critically important to security and resilience. Although such tools and protocols enhance resilience, they themselves are prone to attacks against algorithms and protocols. The platform requires multiple inputs using sensors as IoTs (Internet of things) devices, computational tools, and intelligence information using an ecosystem of innovations. Invariably, concurrent to such innovation pathways are factors such as risk assessment and management, and consequence management, i.e. mitigation and environmental management

3.1.2.: Convergence of Technologies

System resilience is a complex topic that spans multiple disciplines. Information systems and cyber-physical systems have abstractions that span multiple domains. Albeit significant progress has been made in our understanding of these subjects to form an interplay of complexity science, stochastic processes, and with a certain level of uncertainty (especially in dealing with events that are uncertain, irregular, and random), the interplay of convergence technologies produce enhanced situational awareness.

3.1.3: Bacterial and Viral Forensics for Genomics and Bioinformatics

To mitigate the threat of synthetic biology, complex and wide-spectrum bioinformatics [Baldwin et al, 2019] tools are required for genome sequencing and proteomics to eradicate threats at the point of origin, such as decision support for predictive intelligence, and identification of leading indicators that describe enemies' course of action (COA). One of the tools that were experimented with was termed "TRUST" – Tools for Recognizing Unconscious Signals of Trustworthiness and was used for neural activation for recognition of behavioral patterns, physical signals, and neurophysiological signals

3.1.4: Information Technology

For IT/cyber, system resilience should include a combination of modalities. The current focus is on software development, database security, Blockchain technology, and machine learning to better detect, respond, recover, and adapt after the attack. A set of approaches

are commonly employed at the local, state, and even federal level which includes persistent network traffic monitoring/management, technology and support to track, surveil and de-mask the violators, use of AI in data mining, de-centralized social networks for monitoring, HUMINT (Human Intelligence) and OSINT (Open source intelligence) to identify the sources, academic-private partnerships, technologically advanced infrastructure, developing resiliency by design (RbD), enhancing information literacy, responsible IoT manufacturers to actively search for vulnerabilities, enhanced cooperation, and having a chief information security officer (CISO) for every organization.

3.1.5: Social-cultural-behavioral Modeling

The persistent growth of social media has necessitated observation of their activities. Social media provide great resources for personal growth and social connectivity, while these platforms are also used for nefarious activities for terrorism and human trafficking activities. Hence observing their dynamics in terms of decision-making using complexity sciences, worldwide monitoring of malfeasance, and developing strategic and tactical level global view is critical to monitor their dual-use. As stated earlier that a detailed discussion is not within the scope of this brief publication, however, it is intuitive to list a spectrum of platforms to mitigate dual-use threats and increase systemwide resilience.

4. Ethical Concerns

Regulation of Dual-use research occurs at the expense of the complete autonomy of the individual scientists, institutional control, slow progress, independent authority, and Governmental compliance. Further, in the case of synthetic biology, the threat vector comes at the expense of encouraging discoveries, viz. keeping them from proliferation and dissemination. Unfortunately, pervasive and latent cyber threats continue to challenge our communication, information systems, infrastructure, democracy, and society as a whole. The sophistication of threats requires modern security solutions, viz. the use of advanced computing technologies to keep up with malicious actors. Unfortunately, due to a free and open society, cybercriminals take advantage of the same technology-based systems for malicious purposes that we use for daily activities, such as commerce, education, entertainment, communication, and democracy. While AI has the potential to identify and stop cyberattacks, cybercriminals also use the same AI platforms to conduct complex attacks. This means that AI can be used to create more complex, adaptable, and malicious software. These concerns limit academic freedom and come at the cost of adversaries having access to the same literature. Furthermore, AI systems achieve near-human performance, yet lack human ethics and values. In dual-use, intention plays a major role and is an unquantifiable measure. This makes decision-making subjective, which itself poses an ethical dilemma.

5. Conclusion and Future Recommendations

It is demonstrated that with the use of new, innovative, and exponential technologies, there is always a potential for its use for unintended purposes. Some notable examples include the use of nuclear materials, chem.-bio agents, synthetic biology, IT, cyber, and computational algorithms that may be used for nefarious activities by adversaries, equipment and agents that could be misappropriated and misused for weapons development and production; and even the generation and dissemination of scientific knowledge that could, otherwise, be re-purposed. In fact, the concept and approach to security have generally evolved around the use of force and the territorial integrity of

geospace. In stark contrast, innovations in IT, AI, and cyberspace have redefined security and have fundamentally altered traditional battlefields in geospace from within or across its geographical boundaries, and are now outdated and need to be revised. The emergence of AI has triggered fear, uncertainty, competition, and an arms race that is leading us toward a new battlefield that has no boundaries or borders, which may or may not involve humans and will be impossible to understand and perhaps control.

The challenges and complexities of evolving AI threats and security have crossed the barriers of space, ideology, and politics, demanding a constructive collaborative effort from all stakeholders across nations. Furthermore, imposing restrictions poses ethical conundrums of restricting the spread of knowledge, which contradicts the basic tenet of academic institutions with academic freedom, and free speech, while most agree that enacting some guardrails is absolutely essential. While the debate on the structure, role, and dual use of AI will continue in the coming years, any attempt to redefine AI security needs to begin with identifying, understanding, incorporating, and broadening the definition and nature of AI security threats. Some of the suggested recommendations include that the AI developers should evaluate a model's potential to cause "extreme" risks at the developmental stages, before starting any implementation. Also, developers should let external auditors and researchers assess the AI model's risks before and after it's deployed. As generative AI adoption grows at record-setting speeds and computing demands increase, hybrid processing is recommended with a mix of cloud and edge devices, to scale and reach its full potential. Hybrid AI will allow generative AI developers and providers to offer the benefits of performance, personalization, privacy, and security at a global scale. Lastly, incorporating traceability is a step similar to the one proposed by the author for biological weapons. The potential of synthetic biology in terms of gene editing, and stem cell research translates to operational readiness, human performance optimization for overtraining and muscle endurance, biomarkers response, etc., however, dual-use concerns tend to limit the endless potential. Similarly, other technologies such as 3D printing, UAVs, and information technology find similar limitations.

The dual-use nature of these technologies highlights the need for responsible and ethical development, regulatory frameworks, and international cooperation to ensure their positive contributions while mitigating potential risks and misuse. The proliferation risks of dual-use research range from intentional, inadvertent, and accidental to nefarious acts since the intention is not quantifiable. Is the solution to such challenges "superior technology" and threat intelligence? Rooted in data, threat intelligence provides attack surface situational awareness, their motivation, and capabilities, and what might be the indicators of compromise (IOCs) to be aware of allowing informed decisions. Another solution is "better resiliency" and not accepting such threats as the "new normal". Since most malign activities operate in the "gray zone" and are constantly increasing, deterrence may be the wrong analytical lens through which to examine such threats, hence resilience may be a better paradigm.

References

Baldwin, J., Noorali, S., Vaseashta, A., (2109). Wide Spectrum Bio-threats Identification and Classification. Conference: *Chemical, Biological, Radiological and Nuclear Defense - Modernizing the Future Fight: Accelerate & Adaptation*, DE, USA.

Biobricks Foundation, (2023). Building with Biology to Benefit All People and the Planet https://biobricks.org/ accessed June 2023.

Buchholz, K. (Jan. 24, 2023). ChatGPT sprints to one million users. Statista. https://www.statista.com/chart/29174/time-to-one-million-users/ Accessed June, 2023.

ChatGPT.org, (2023). What is ChatGPT? https://chat-gpt.org/ accessed June 2023.

Cotton, S., (2016). Litvinenko poisoning: polonium explained, The Conversation, published online January 21, 2016 https://theconversation.com/litvinenko-poisoning-polonium-explained-53514

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 319-340. https://globalassistant.info/technology-acceptance-model-davis-1989-pdf-download-link-free/ accessed June 2023.

Dastin, J. et al. (Feb. 22, 2023). For tech giants, AI like Bing and Bard poses billion-dollar search problem. Reuters. https://www.reuters.com/technology/tech-giants-ai-like-bing-bard-poses-billion-dollar-search-problem-2023-02-22/ Accessed June 2023.

Frommer, F., (2022). When Soviet-Led Forces Crushed the 1968 'Prague Spring' https://www.history.com/news/prague-spring-czechoslovakia-soviet-union , Accessed June 2023.

iGEM (2023). *What is iGEM?* https://igem.org/ accessed June 2023.

Garner K.L. (2021). Principles of synthetic biology. *Essays Biochem*.; 65(5):791-811. https://www.doi.org/10.1042/EBC20200059 . PMID: 34693448; PMCID: PMC8578974.

Iverson, N., Barone, P., Shandell, M. et al., (2013). In vivo biosensing via tissue-localizable near-infrared fluorescent single-walled carbon nanotubes. Nature Nanotech 8, pp: 873–880. https://doi.org/10.1038/nnano.2013.222

Hamilton, L.D., Medeiros, W.H., Moskowitz, P.D., & Rybicka; K., (1988). Toxicology of tetramethyl tin and other organometals used in photovoltaic cell manufacture. *AIP Conference Proceedings*; 166 (1): 54–66. https://doi.org/10.1063/1.37131

Lallanilla, M., (2013). Yasser Arafat's Death Linked to Radioactive Polonium, Live Sciences. https://www.livescience.com/41008-yasser-arafat-assassination-polonium.html (published online Nov. 06, 2013). Accessed June 2023.

Mizoguchi H, Mori H, Fujio T. (2007). Escherichia coli minimum genome factory. *Biotechnol Appl Biochem*; 46(Pt 3):157-67. http://www.doi.org/10.1042/BA20060107 . PMID: 17300222.

Pandya, J., & Cognitive World, (2019). Excerpts from an interview: The Dual-Use Dilemma Of Artificial Intelligence, *Forbes*, https://www.forbes.com/sites/cognitiveworld/2019/01/07/the-dual-use-dilemma-of-artificial-intelligence/?sh=488c297e6cf0. Narrative: Dual Use Technology Dilemma and National Security" with Prof. Dr. Ashok Vaseashta. Accessed on June 2023.

Pearson, T., (2022). *The Ultimate Guide to the OODA Loop*, https://taylorpearson.me/ooda-loop/ , accessed: June 2023.

Qazi, R., Gomez, A.M., Castro, D.C., et al. (2019). Wireless optofluidic brain probes for chronic neuropharmacology and photo-stimulation. *Nat Biomed Eng*. 3, pp: 655–669. https://doi.org/10.1038/s41551- 019-0432-1

Saalbach KP. Gain-of-function research. *Adv Appl Microbiol*. 2022; 120:79-111. http://doi.org/10.1016/bs.aambs.2022.06.002 Epub 2022 Jul 13. PMID: 36243453.

Schmitt, M. (2017). *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (2nd ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316822524

Seleiman M.F., Al-Suhaibani N., Ali N., et al., (2021). Drought Stress Impacts on Plants and Different Approaches to Alleviate Its Adverse Effects. *Plants*;10(2):259. doi: http://doi.org/10.3390/plants10020259 PMID: 33525688

Sheth, S. (Feb. 25, 2023). Generative AI drives an explosion in compute: The looming need for sustainable AI. Silicon Angle. https://siliconangle.com/2023/02/05/generative-ai-drives-explosion-compute-looming-need-sustainable-ai/ Accessed June 2023.

SYNBERC, (2023). Synthetic Biology Research Center, https://ebrc.org/synberc/ accessed June 2023.

Teschler, T, (2012). Another blow for WTC conspiracy theorists, *Machine Design*, Published online, July 16, 2012. https://www.machinedesign.com/home/article/21830429/another-blow-for-wtc-conspiracy-theorists, Accessed June 2023.

Thorrison, H., Baiardi, F., Angeler, F., Teveter, K., Vaseashta, A., Rowe, P., Piotrowicz, W., Polmateer, T., Lambert, J., (2020). Applying Resilience to Hybrid Threats: Integrating Infrastructural, Digital, and Social Systems. In, *Resilience and Hybrid Threats - Security and Integrity for the Digital World*, Vol. 55, pp.1-12 NATO Science for Peace and Security Series - D: Information and Communication Security, http://doi.org/10.3233/NICSP190017

United Nations (1949). Geneva Convention relative to the Treatment of Prisoners of War. https://www.ohchr.org/en/instruments-mechanisms/instruments/geneva-convention-relative-treatment-prisoners-war , accessed: June 2023.

Vaseashta A., (2012). Technological Innovations to Counter CBRNE Threat Vectors and Ecotage. In: Vaseashta A., Braman E., Susmann P. (eds) *Technological Innovations in Sensing and Detection of Chemical, Biological, Radiological, Nuclear Threats and Ecological Terrorism*. Pp: 3-23. NATO Science for Peace and Security Series A: Chemistry and Biology. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-2488-4_1

Vaseashta, A., Susmann, P., Braman, E., (2014). *Cyber Security and Resiliency Policy Framework*. 38 of NATO Science for Peace and Security Series - D: Information and Communication Security, IOS Press, The Netherlands. ISBN: 978-1-61499-445-9 (print) | 978-1-61499-446-6 (online).

Vaseashta, A., (2014). Advanced Sciences Convergence based methods for surveillance of emerging trends in science, technology, and intelligence. 2014, *Foresight*, 16(1), pp: 17- 36, https://doi.org/10.1108/FS-10-2012-0074

Vaseashta, A., Braman, E., and Vaseashta, S.B., Mitigating Unconventional Cyber-Warfare: Scenario of Cyber 9/11, (2016). Handbook of Research on Civil Society and National Security in the Era of Cyber Warfare. pp: 238-260. In *Handbook of Research on Civil Society and National Security in the Era of Cyber Warfare*. Eds. Metodi Hadji-Janev, Mitko Bogdanoski. http://DOI.org/10.4018/978-1-4666-8793-6.ch012

Vaseashta A., (2018). Roadmapping the Future in Defense and Security: Innovations in Technology Using Multidisciplinary Convergence. pg. 3-14. In: Petkov, P., Tsiulyanu, D., Popov, C., Kulisch, W. (eds) *Advanced Nanotechnologies for Detection and Defence against CBRN Agents*. NATO Science for Peace and Security Springer, Dordrecht. https://doi.org/10.1007/978-94-024-1298-7_1

Vereshchagin, A. (2013). Lal Bahadur Shastri's death in Tashkent still raises questions. *Russia Beyond*, https://www.rbth.com/arts/2013/10/02/lal_bahadur_shastris_death_in_tashkent_still_raises_questions_29837, accessed June 2023.

Viswanath Venkatesh, Fred D. Davis, (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science* 46(2):186-204.

Yang, X., McGlynn, E., Das, R., Paşca, S. P., Cui, B., Heidari, H., (2021). Nanotechnology Enables Novel Modalities for Neuromodulation. *Adv. Mater*. 33, 2103208. https://doi.org.10.1002/adma.202103208

Stanford | Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Fairness in AI and Its Long-Term Implications on Society

Ondrej Bohdal, [1]* Timothy Hospedales, [2] Philip H.S. Torr, [3] Fazl Barez [4]

Citation: Ondrej Bohdal, Timothy Hospedales, Philip Torr, Fazl Barez. Fairness in AI and Its Long-Term Implications on Society. *Proceedings of the Stanford Existential Risks Conference* 2023, 171-186. https://doi.org/10.25740/pj287ht2654

Academic Editor: Trond Undheim, Dan Zimmer

Funding: Ondrej Bohdal and Fazl Barez were supported by the EPSRC Centre for Doctoral Training in Data Science (grant EP/L016427/1) and EPSRC Centre for Doctoral Training in Robotics and Autonomous Systems (EP/L016834/1) respectively, funded by the UK Engineering and Physical Sciences Research Council and the University of Edinburgh. Philip Torr was supported by Turing AI Fellowship EP/W002981/1.

Conflict of Interest Statement: N/A

Informed Consent Statement: N/A

Acknowledgments: We are grateful to Trond Undheim for providing highly useful and detailed suggestions for improving our paper as well as to Charlotte Siegmann and Shahar Avin for their feedback on our earlier draft. Their insights and comments have greatly improved the quality of this work.

Author Contributions: Ondrej Bohdal and Fazl Barez are the primary contributors. Timothy Hospedales and Philip Torr helped improve the ideas presented in the paper.

Abstract: Successful deployment of artificial intelligence (AI) in various settings has led to numerous positive outcomes for individuals and society. However, AI systems have also been shown to harm parts of the population due to biased predictions. AI fairness focuses on mitigating such biases to ensure AI decision making is not discriminatory towards certain groups. We take a closer look at AI fairness and analyze how lack of AI fairness can lead to deepening of biases over time and act as a social stressor. More specifically, we discuss how biased models can lead to more negative real-world outcomes for certain groups, which may then become more prevalent by deploying new AI models trained on increasingly biased data, resulting in a feedback loop. If the issues persist, they could be reinforced by interactions with other risks and have severe implications on society in the form of social unrest. We examine current strategies for improving AI fairness, assess their limitations in terms of real-world deployment, and explore potential paths forward to ensure we reap AI's benefits without causing society's collapse.

Keywords: AI fairness, AI safety, AI risks, cascading risks, biased AI models

[1] Research Associate, School of Informatics, University of Edinburgh, Edinburgh, UK; ondrej.bohdal@ed.ac.uk.

[2] Professor, School of Informatics, University of Edinburgh, Edinburgh, UK and Head of Samsung AI Center, Cambridge, UK; t.hospedales@ed.ac.uk.

[3] Professor, Department of Engineering Science, University of Oxford, Oxford, UK; philip.torr@eng.ox.ac.uk.

[4] PhD Student, School of Informatics, University of Edinburgh, Edinburgh, UK and Visiting Student, Department of Engineering Science, University of Oxford, Oxford, UK; f.barez@ed.ac.uk.

* Correspondence: ondrej.bohdal@ed.ac.uk

1. Introduction

AI approaches offer excellent performance in many practically important problems (Caruana et al., 2015; Hinton et al., 2012; Andreu-Perez et al., 2018), but they can give biased and unfair predictions (Mehrabi et al., 2019; Hardt et al., 2016; Dwork et al., 2012). AI is increasingly often deployed to high-stakes applications (Tolan et al., 2019; Larrazabal et al., 2020; Seyyed-Kalantari et al., 2021), where unfair predictions can lead to substantial disadvantage or harm to parts of the population. For example, AI has been used for deciding who to select for interviews (Cohen et al., 2020; Bogen and Rieke, 2018), who should be given a mortgage (Martinez and Kirchner, 2021) or who is more likely to repeat crime after leaving prison (Tolan et al., 2019; Dressel and Farid, 2018). Unfair decisions in such key areas can have a significant impact on one's future.

We study the long-term social implications of unfair AI, from the perspective of continuous bias amplification stemming from new AI models trained on increasingly biased data. Current AI models have been shown to be biased, especially because of being trained on biased data (Mehrabi et al., 2019). At the same time, they have been shown to make more biased decisions than present in the training data, hence amplifying the biases (Lloyd, 2018; Hall et al., 2022).

Decisions of AI models influence the real world, and information about the real world is used for training new AI models. This means that more biased decisions will lead to more biased data for training new AI models, resulting in a more biased new generation of AI models. Additionally, parts of the population can experience bias from several sources, e.g. hiring and healthcare, and these combined can also put certain groups in increasingly large disadvantage over time. Overall this represents a feedback loop where new AI models give more and more biased decisions.

It is relatively common that AI biases are uncovered only after several years since the systems have been deployed (Dressel and Farid, 2018; Obermeyer et al., 2019; Martinez and Kirchner, 2021). Once a bias becomes recognized, action should be taken to rectify it. There are two main solutions: avoid using the biased AI model, or fix the AI model so that it is no longer biased. In many circumstances, it can be difficult to stop using the AI system as the system may already be deeply embedded and alternative solutions can be costly (Viechnicki and Eggers, 2017) or unviable, for example due to shortage of employees (Ford, 2021). On the other hand, improving the fairness of the AI model may take considerable time given that it is an open research problem (Mehrabi et al., 2019; Zong et al., 2023). Perhaps one of the best ways to mitigate bias is to curate bias-free data (Jordon et al., 2022), but this has its own challenges, especially because using real already collected data is significantly easier. Most suitable solutions to rectify biases depend on the specific circumstances, but in many cases rectifying AI bias can be challenging (Hao, 2019). As a result, the system may continue to be operating, perhaps with only minor changes. It is also important to note though that we should not completely avoid using AI as using it can bring numerous benefits and valuable insights – the key is to use AI responsibly.

Nevertheless, if parts of the population are systematically marginalized because of biased AI models, they can be under severe stress and they may try to resolve the situation by protesting against the deployment of such AI systems. If the institutions find it challenging to stop using biased AI, e.g. due to lack of employees or resources more broadly, disadvantaged groups may resort to escalating the situation using violence. In this sense lack of AI fairness can act as a social stressor that can lead to social unrest if the issues are prevalent and not addressed.

Deployment of insufficiently fair AI systems would likely be only one of several social stressors. Consequently, we study the interaction with other stressors, especially climate change that has increasingly significant impact on the society. The interaction among multiple social stressors can reinforce each other and result in more extensive social unrest. Over longer time horizons, this could destabilize the situations in various countries.

In addition to studying social implications, we also investigate what approaches are being developed to improve AI fairness. We give particular focus on real-world deployment of fair AI models and identify lack of fairness generalization across data distribution shifts can be a key challenge. We discuss approaches for robust fairness, but we also discuss approaches from areas related to out-of-distribution robustness, including domain generalization and adaptation. We conclude with a broader discussion that includes suggestions that can help mitigate risks and obtain more benefits from deploying AI.

2. Definitions of Fairness

Researchers have proposed a variety of ways to define fairness (Mehrabi et al., 2019), and some of the most common include equalized odds (Hardt et al., 2016), equal opportunity (Hardt et al., 2016) and demographic parity (Dwork et al., 2012). The unifying theme of these metrics is we want to ensure the same or similar probability of the given outcomes across all considered groups, which we illustrate in Figure 1. The concept of AI fairness hence captures the notion of ensuring AI decision making is not discriminatory toward any of the groups that interact with the AI system.

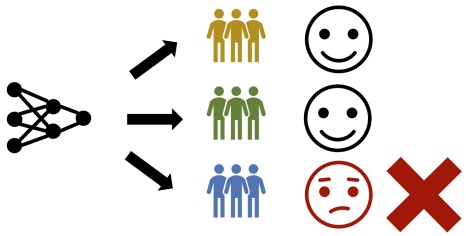


Figure 1: We focus on the topic of fairness where we want to ensure that all groups receive unbiased and equal treatment so that no groups are harmed because of using AI.

Equalized odds (Hardt et al., 2016) for predictor $\hat{Y}$, target Y and protected attribute A are defined for the binary case as:

$$Pr\{\hat{Y} = 1 | A = 0, Y = y\} = Pr\{\hat{Y} = 1 | A = 1, Y = y\}, \quad y \in \{0, 1\}.$$

The definition means $\hat{Y}$ has equal true positive rate for demographics $A = 0$ and $A = 1$ if the outcome is $y = 1$, and equal false positive rates if the outcome is $y = 0$.

Equal opportunity (Hardt et al., 2016) is a relaxed alternative of equalized odds as it only requires non-discrimination within the advantageous outcome group:

$$Pr\{\hat{Y} = 1 | A = 0, Y = 1\} = Pr\{\hat{Y} = 1 | A = 1, Y = 1\}.$$

Compared to earlier demographic parity metric (Dwork et al., 2012), the benefit of equalized odds and equal opportunity is that they do not require independence from the protected attribute (Hardt et al., 2016). More broadly when deciding which metric to use, it is key to consider the suitability for the specific application (Zong et al., 2023).

In addition to specialized fairness metrics, we can monitor the worst-case performance alongside the average performance (Zhang et al., 2021; Bohdal et al., 2022b). More specifically we can measure the performance on the most challenging group (Zhang et al., 2021) or if the notion is less clear, we can use e.g. the most challenging 10% of the examples used for evaluation (Bohdal et al., 2022b). Such way of evaluation can also be used for settings where we want to ensure fairness when deploying AI systems across different scenarios. It is related to Max-Min fairness (Lahoti et al., 2020) where a model with smaller worst-case error is seen as fairer.

We further consider a stronger notion of fairness that is important when deploying models to the real-world: fairness that is robust under data distribution shifts. Data distribution shifts are common when deploying models to the real world. They can arise for example from applying the model to populations of different characteristics than were used during training, or also from processing the data differently (e.g. due to using different cameras or sensors). A real world example is training an AI model using medical images taken with a scanner of one brand, and then deploying the model to a hospital where a scanner of different brand is used. AI models should not discriminate against any of the subgroups when deployed to real-world "in-the-wild" scenarios. We illustrate robust fairness in Figure 2, and we will also consider this notion when discussing current solutions towards AI fairness.

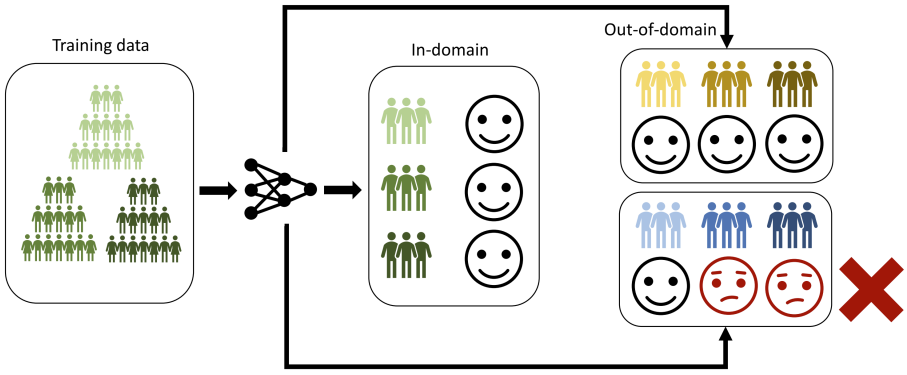


Figure 2: When deploying AI models to the real world, it is crucial to ensure the models are robust and generalize fairness also to out-of-domain ("in-the-wild") situations.

3. Social Implications of AI Fairness

We begin our analysis of social implications of AI fairness by introducing several high-stakes real-world examples where AI has already been used. We will then present a self-reinforcing feedback loop mechanism where biased AI systems lead to more biased outcomes, which then act as input data for further training of new AI systems. Over long periods of time this may lead to increasingly systemic social and economic marginalization of parts of the population. Such systemic marginalization could later become a substantial social stressor.

3.1 High-Stakes Real-World Applications

With the possibility to automate various time-consuming tasks and potentially improve upon imperfections of human decision-making, AI has been used for a number of high-stakes applications where fairness is important (Mehrabi et al., 2019). However, in many cases it has already been identified that the AI is unfair and causes harm to certain groups.

High-stakes real-world applications where fairness matters and has already been compromised include the following:

- Hiring for jobs: biased AI has been used in the context of hiring in multiple ways, including filtering of CVs (Cohen et al., 2020; Bogen and Rieke, 2018), evaluating video interviews (Kelly-Lyth, 2021) and delivering advertisements promoting jobs (Lambrecht and Tucker, 2019). Type of employment has a large impact on one's future, so it is key to ensure certain groups are not systematically disadvantaged (or given an advantage).
- Finance: AI can simplify the task of assessing if someone is likely to repay a loan or a mortgage, so such systems have already been deployed in practice. It has been shown that systems for making decisions about loans (Mukerjee et al., 2002) or mortgages (Martinez and Kirchner, 2021) can be significantly biased, for example making applicants of color 40 to 80% more likely to be denied mortgage application compared to white applicants (Martinez and Kirchner, 2021). If a group of certain characteristics is unable to get a mortgage and is forced to rent, it can have a large impact on their well-being, especially if it means they have to find new accommodation often.
- Public safety: unfair AI systems have been used in various public safety contexts, including sentencing decisions (Dressel and Farid, 2018; Tolan et al., 2019) and children welfare (Chouldechova et al., 2018). More specifically, AI has been used to predict the risk of recidivism as part of the COMPAS system (Dressel and Farid, 2018), and also as part of a tool applied to the particularly sensitive case of predicting juvenile recidivism (Tolan et al., 2019). Biased AI has also been used in the context of children welfare to perform screening of referrals for child protection (Chouldechova et al., 2018).
- Healthcare: biased AI systems have been deployed for multiple healthcare applications. For example, health-management systems (Obermeyer et al., 2019) have been shown to assign the same risk to black patients that are sicker than white patients. Biased AI has also led to underdiagnosis of under-served patient populations when applying AI to chest radiographs (Seyyed-Kalantari et al., 2021), and it also resulted in gender-biased computer-aided diagnosis (Larrazabal et al., 2020).

Additionally, it is not only the high-stakes situations where AI has the potential to discriminate and treat people unfairly. There are also situations where unfair AI can cause inconvenience. However, these can potentially act as a reminder that the person may have been treated unfairly by AI in some of the high-stakes situations, raising anger levels.

Face recognition is one of the key areas that exemplifies such scenarios and shows the need for fair and robust AI models. For instance, earlier facial processing systems from leading tech companies performed significantly worse on black women than on white men (Buolamwini and Gebru, 2018; Raji et al., 2020). Widespread use of such models could lead to frequent inconveniences, for example if face recognition technology were utilized for workplace access or for identifying criminals in public spaces.

The concept of fairness is also important in the context of generative language models such as GPT-4 (OpenAI, 2023), LaMDA (Thoppilan et al., 2022) and LLaMA (Touvron et al., 2023), because the generated content can influence people and have real-world impact. We want to ensure the generated content is not biased and does not include prejudices about any parts of the population. Further we want to ensure that safeguards in the models are robust across different languages and cultures so that we do not risk, for example, harming parts of the society in countries that speak lower-resource languages.

3.2 Self-Reinforcing Feedback Loop

Deployed AI models influence the society, and the outcomes form part of the training data for a new generation of AI models. If the initial models are biased, they produce biased outputs that will be used for training newer models. It has been shown that AI models amplify biases (Lloyd, 2018; Hall et al., 2022), which means new AI models would be biased even more due to training on increasingly biased data. We illustrate this in Figure 3, where we show the resulting feedback loop. Because many biases become evident only after wide deployment of the system, it is key to consider what long-term implications AI-amplified biases could have.

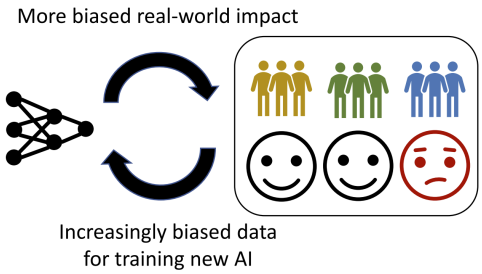


Figure 3: Biased real-world outcomes lead to increasingly biased data for training new AI models, resulting in a self-enforcing feedback loop.

Let us explain the bias amplification on examples. For example, if jobs hiring decision AI is based on past hiring decisions, then bias against subgroups in the past can reinforce to more bias in the future. If criminal sentencing AI is biased against a subgroup, that subgroup has longer prison sentences, which may make them harder to re-integrate with society after release. This may increase their likelihood of repeat crime recidivism, which will be data that increases the bias of the sentencing-decision AI the next time it is trained.

A whole ecosystem of AI models that happened to be biased against a particular subgroup could lead to persistently worse social and economic outcomes for that subgroup. More specifically, it could lead to worse jobs, worse access to finance, longer sentences for equivalent crimes, worse health due to worse medical treatment, worse educational outcomes if the education AI is biased. These biases then reinforce each other as e.g. worse health reinforces worse education and jobs. Over long periods of time that subgroup could become increasingly systemically socially and economically marginalized, which could become a substantial social stressor.

There is a risk of compounding of negative effects due to the feedback potential between the AI system decisions and real-world data, which affects the training data for the next round of AI training. The level of bias may become increasingly more difficult to tolerate and protests will arise. Moreover, as the technology becomes widespread, infrastructure will be built around it, so it will be challenging to remove it from use even as people protest against it. This would ultimately create tensions.

3.3 Fairness Risk

We present a toy model to estimate the fairness risk at time t after deployment of the first AI systems. The formula models the compounding of biases over time (similar to compounding of interest rates):

$$\text{risk} \sim \beta^t,$$

where β is the bias amplification rate of the AI models. When no bias amplification happens, $\beta = 1$, but because AI models have been shown to increase biases, typically $\beta > 1$. Over time the biases would amplify each other via repeated training of new AI systems on increasingly biased data. Such amplification of biases can lead to dissatisfaction with the AI systems and can act as a social stressor. Ideally we would like to have $\beta \leq 1$.

3.4 Challenges with Avoiding AI Automation

Budgets are typically tight (Tujula and Wolswijk, 2004), and ways to save resources are sought. AI offers ways to decrease costs (Viechnicki and Eggers, 2017; Berglind et al., 2022) and if there is a crisis, institutions may be more likely to use experimental approaches that have not been fully assessed. If the technology is likely to benefit most of the population, it may be difficult to argue against its deployment, especially if it is hard to measure potential harm it may cause. In these cases, it is important to recognize effects across different groups. One can try to mitigate them by investing more resources into making the technology fair and robust. It is also important to note that AI-based solutions may become deeply embedded in the software infrastructure and removing them once new significant biases are identified or if people start protesting, can be challenging.

In addition, automation using AI is not only about savings, it may also be inevitable due to shortages of employees to perform specific jobs (Ford, 2021). For example, shortages have been reported in areas as diverse as construction, social work, and transportation (Francis-Devine and Buchanan, 2023). Shortages may not always be resolved by increasing the budgets because some jobs require hard-to-obtain skills or cause significant amount of distress, so using AI would be desirable also because of non-monetary reasons. More generally AI can bring significant benefits and improve lives of people in various directions. As a result, it is important to find a suitable trade-off between AI use and regulation that tries to mitigate the negative impacts.

3.5 Interaction with Other Risks and Existential Implications

Unfair AI would be only one factor that would contribute to tension in the society. Another key driver of tension is likely to be climate change (Institute for Economics & Peace, 2020) and its implications (Nardulli et al., 2015), which include rising food prices or having to cover the costs of disaster responses. For example, large-scale drought in Syria is thought to have contributed to social stressors, which eventually led to an uprising in 2011 (Kelley et al., 2015). Syria is in civil war since 2011, already for more than a decade (Loft et al., 2022; Khen et al., 2020). The case of Syria shows that a combination of multiple stressors that lead to uprisings can result in a long-term civil war, which is a prime example of country's crisis. More broadly systemic unfairness, inequality, and marginalization of parts of the population has a record of leading to radicalization (van den Bos, 2020), violent uprisings and in some cases destabilization of societies. High-profile examples where these were likely to be a factor include the French revolution (Tocqueville, 1856), Indian independence movement (Chandra, 1989) and recently the Arab Spring (Haas, 2017).

A combination of multiple social stressors such as climate change and biased AI are likely to reinforce each other, which we illustrate in Figure 4. Persistent deployment of biased AI in high-stakes applications can lead to increasing levels of tension in the society and erode trust in the institutions. At a certain level it may be significant enough that in interaction with other social stressors such as climate change it escalates. It is crucial to try to mitigate the social stressors to avoid compounding effects.

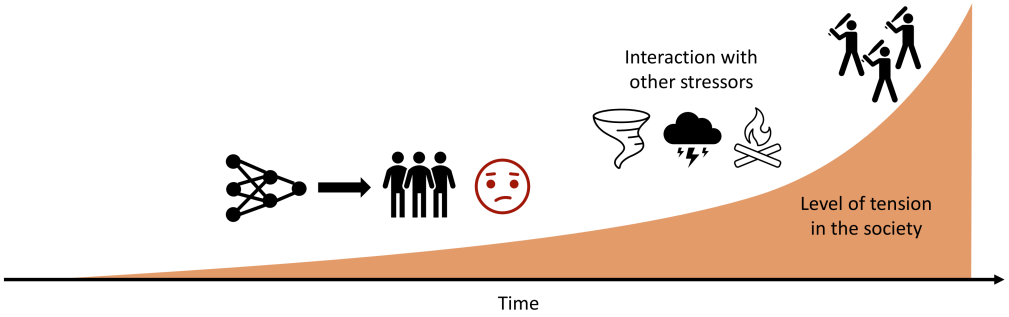


Figure 4: Biased AI outcomes in interaction with other social stressors can lead to increasing levels of tension in the society and escalate if not mitigated.

Biased AI, climate change and other social stressors have the potential to be commonplace across many nations, which brings the risk of widespread social unrest. Social unrest can turn into a civil war in severe cases (Kelley et al., 2015; Khen et al., 2020), which could potentially endanger the civilization if present in many countries. Overall this means that the risks from unfair AI can be large if it is deployed to too many critical applications without careful consideration.

4. How to Improve AI Fairness?

4.1 Approaches for Fairness

A large number of approaches for fairness has been proposed (Mehrabi et al., 2019; Zong et al., 2023), reflecting the importance of the field. Many of the recent methods try to alleviate bias as part of training the models, also known as in-processing (Mehrabi et al., 2019; Zong et al., 2023). Key in-processing families of bias mitigation methods include:

- Subgroup rebalancing (Chawla et al., 2002; Idrissi et al., 2022) that over-samples minority groups and down-samples majority groups,
- Domain independence (Wang et al., 2020; Royer and Lampert, 2015) that uses separate classifiers for different subgroups,
- Adversarial training (Madras et al., 2018; Zhao et al., 2019; Kim et al., 2019) that tries to train representations that make it difficult to identify different groups,
- Disentanglement (Tartaglione et al., 2021; Sarhan et al., 2020) methods that separate sensitive attributes and the useful attributes when constructing the representations.

Other families of methods for improving fairness also exist. Zong et al. (2023) have identified that domain generalization approaches (Sagawa et al., 2020; Cha et al., 2021; Foret et al., 2021) can be useful for improving fairness. Domain generalization methods try to learn representations that directly generalize to new out-of-domain situations without any adaptation, which relates to the goal of obtaining strong performance across different groups present in the population. Further, pre-processing methods (Khodadadian et al., 2021) try to remove bias from the dataset before training, for example by distorting the data (Khodadadian et al., 2021) as simply removing the sensitive attributes has been shown to be insufficient (Mehrabi et al., 2019). Post-processing methods (Pleiss et al., 2017) modify the predictions of an already trained model to improve fairness with respect to the sensitive attributes.

4.2 Fairness Under Data Distribution Shifts

When deploying AI models to the real-world, it is key to ensure the key properties of AI models hold also for real "in-the-wild" data. Such data are likely to come also from data distributions different from ones seen during training, so robustness against distribution shifts is crucial. For example, healthcare AI systems can be trained on data from selected prestigious US hospitals and deployed to hospitals of various quality across the US.

However, it has been shown that most existing fairness methods are only designed for in-domain settings and fail when data distribution changes (Schumann et al., 2019; Mishler and Dalmasso, 2022; Zong et al., 2023). Several approaches have been developed to tackle the challenge of fairness under distribution shift (Singh et al., 2021; Rezaei et al., 2020; Schumann et al., 2019), but these consider adaptation to a specific domain, with only (Pham et al., 2023) presenting an approach that generalizes across domains. If institutions only consider if an AI system is fair for in-domain data, such AI models may still lead to significant biases when deployed in the real-world and cause harm to parts of the population.

It has been identified (Zong et al., 2023) that domain generalization approaches (Sagawa et al., 2020; Cha et al., 2021; Foret et al., 2021) can offer competitive performance in terms of fairness, so improving fairness of domain generalization methods can be a good way forward. However, it has also been shown that domain generalization is a challenging problem on its own (Gulrajani and Lopez-Paz, 2021), with many approaches performing similarly as simple training across many domains (Vapnik, 1998) if following a fair evaluation protocol. As a result, domain adaptation approaches that adapt pre-trained models to local data distributions can be more successful in terms of maintaining fairness and strong performance. Source-free domain adaptation (Liang et al., 2020; Kundu et al., 2020; Ishii and Sugiyama, 2021; Yang et al., 2021) in particular can be practically valuable as it adapts a pre-trained model solely using unlabelled target domain data, without access to the source data. Efficient feed-forward approaches that perform adaptation without back-propagation (Bohdal et al., 2022b; Schneider et al., 2020; Bohdal et al., 2022a, 2023) can be especially useful on deployed devices.

4.3 Evaluation

Various benchmarks have been developed or repurposed for evaluating fairness. Key tabular fairness datasets include COMPAS (Dieterich et al., 2016), Adult Census (Ding et al., 2021; Kohavi and Becker, 1996) and Diabetes (Strack et al., 2014), some of which are also available within the popular Fairlearn library (Bird et al., 2020). Common computer vision fairness datasets include CIFAR-10S (Wang et al., 2020), CelebA (Liu et al., 2015) and IMDB face dataset (Rothe et al., 2015). Medical imaging datasets (Irvin et al., 2019; Johnson et al., 2019; Groh et al., 2021) have also been used extensively for evaluating fairness, with MEDFAIR (Zong et al., 2023) providing a suite of benchmarks to provide rigorous evaluation of fairness algorithms, including in-domain and out-of-domain scenarios.

Long term we believe it is key to develop new more extensive benchmarks that test both in-domain and out-of-domain scenarios, similar in scope to MEDFAIR (Zong et al., 2023) but covering various areas for which AI fairness is crucial. Because it has been observed that real-world datasets are often biased, synthetic datasets may be highly useful in the future. Synthetic data would enable us to design what unbiased outcomes look like and train models on them, improving fairness and robustness (Jordon et al., 2022). Once the model is trained with synthetic data, it can be fine-tuned using curated real-world data that do not need to be as plentiful.

4.4 Mechanistic Interpretability and Fairness

In the context of fairness, mechanistic interpretability (MI) is of significant importance. As a growing sub-field of interpretability, it aims to understand individual neurons within models and their larger circuits, playing an essential role in various applications (Olah et al., 2020b; Olah, 2022; Goh et al., 2021). The ultimate goal is to decompose a model into interpretable components, enabling more significant insights into model safety and bias detection (Hendrycks et al., 2021; Amodei et al., 2016; Vig et al., 2020). MI can be helpful in numerous fields such as autonomous vehicles (Barez et al., 2022) and Large Language Models (LLMs) (Miceli-Barone et al., 2023). However, fully achieving mechanistic interpretability in these domains remains a challenging task. One approach used in image models is feature visualization, where synthetic input images are employed to optimize the understanding of a target neuron (Olah et al., 2017). These techniques have significantly enhanced the interpretation of vision models, helping identify multimodal neurons that respond to abstract concepts (Goh et al., 2021) and cataloguing the behaviour of early neurons in Inception-v1 (Olah et al., 2020a). Examining individual neurons can allow for better and more human-readable visualizations (Foote et al., 2023), enabling testing of the role of scale, the addition of more data and parameters, and assessing the impact of removing undesired concepts (Hoelscher-Obermaier et al., 2023).

5. Discussion

5.1 Broader AI Safety

We have explored how insufficiently fair AI could contribute to the collapse of society if it is delegated too many critical systemic functions. Our approach is relatively unique in the space of existential-risk literature because we focus on discussing the danger of AI that has already become commonplace, rather than focusing on highly-intelligent AI having goals misaligned with human values. We believe both present risks and deserve significant attention. To provide a more complete view of AI safety, we briefly discuss also the other directions studying AI risks.

A variety of hazards have been associated with advanced AIs, including weaponization, proxy gaming, emergent goals, deception, and power-seeking behaviour (Hendrycks and Mazeika, 2022). It has been discussed how due to competitive pressures the most successful AIs would have by default undesirable traits, ones misaligned with human values (Hendrycks, 2023). Consequently, it can be expected that without suitable interventions, those hazards could become real when AI reaches sufficiently advanced level. A variety of high-level interventions have been proposed, including careful design of AI's intrinsic motivations, constraints on actions and cooperative institutions.

Another serious risk comes from the fact that even the current AI models have the scope for malicious use, with the ability to impact digital, physical, and also political security (Brundage et al., 2018). Current AI can be dual-use, for example with the potential to develop new chemical weapons (Urbina et al., 2022). Existing generative models also have the scope for introducing serious hazards, for example by being used for spreading misinformation on a large scale (Goldstein et al., 2023).

5.2 Recommendations

Considering the significant impact that insufficient AI fairness can have on the future of our society, there are various steps that can be taken to mitigate negative impact and make AI more beneficial. We identify three key areas that we discuss in more depth.

5.2.1 Develop the Science of Iterative Bias Amplification

Existing work has shown that AI models have the tendency to amplify biases (Lloyd, 2018; Hall et al., 2022). These biases influence the real-world, which introduces the feedback loop where the biases become stronger over time. To better analyze the impact of deploying imperfect AI models over longer time horizons, it would be useful to study this from various perspectives. More specifically, we suggest developing a science of iterative bias amplification that will help us understand how decisions made by current AI systems (which determine the training set of future AI systems) affect the evolution of AI bias and fairness in the long run. Agent-based modelling (Bonabeau, 2002) could be a useful tool for studying this area.

5.2.2 Develop Foundational Synthetic Datasets

One of the main reasons why AI models are biased is that they are trained on biased data that reflect biases present in the society (Mehrabi et al., 2019). A suitable solution could be to develop new foundational synthetic datasets that can be used for fair pre-training of AI models (Jordon et al., 2022). Constructing such large-scale datasets can be expensive, not only due to the size but also because various parties would need to be involved in the design process to ensure usefulness of the datasets. In a way the design of such datasets could benefit from consultations similar to the ones used for designing new policies. Consequently, we envision it would be primarily the governments and other large institutions that would cover the costs of these public-benefit datasets.

5.2.3 Policy Guidelines and Regulations

AI is likely to have increasingly large real-world impact, so it is crucial to ensure adequate policy guidelines and regulations regarding AI fairness are in place. This includes policy guidelines and regulations that have real-world impact on funding, algorithmic transparency, and which mandate human-in-the-loop for important applications. As the society evolves over time, it will be key to also require monitoring of deployed AI systems as they influence people in the real-world. The biases may not be present during initial evaluation, but can arise later after the system has been deployed. It is encouraging to see governments are introducing policies on fair and responsible use of AI (European Commission, 2020; Office of Science and Technology Policy, 2022; Department for Science, Innovation and Technology, 2023), but it will be key to ensure the policies are continuously adapted as the society evolves and new research is conducted.

6. Conclusion

In this paper we have investigated the long-term implications of unfair AI systems. We have identified that a feedback loop that leads to increasingly large biases can arise as biased AI models impact the population and new AI models are trained on such outcomes. Over longer time horizons, increasing levels of systemic unfairness can act as a social stressor and trigger protests, which can result in social unrest. We have discussed real-world limitations of existing AI systems designed to be fair and suggested steps that can be taken to improve the situation. Overall, we believe that thanks to the significant interest from both the ML community and institutions deploying AI systems, potential severe risks stemming from biased AI systems can be avoided, but carefulness and extensive further research will be key.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in AI safety. ArXiv.

Andreu-Perez, J., Deligianni, F., Ravi, D., and Yang, G.-Z. (2018). Artificial intelligence and robotics. Technical report, UK-RAS Network.

Barez, F., Hasanbeig, H., and Abate, A. (2022). System III: Learning with domain knowledge for safety constraints. In NeurIPS ML Safety Workshop.

Berglind, N., Fadia, A., and Isherwood, T. (2022). The potential value of AI—and how governments could look to capture it. Technical report, McKinsey & Company.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Technical report, Microsoft.

Bogen, M. and Rieke, A. (2018). Help wanted: an examination of hiring algorithms, equity, and bias. Technical report, Upturn.

Bohdal, O., Li, D., and Hospedales, T. (2022a). Feed-forward source-free domain adaptation via class prototypes. In ECCV OOD-CV Workshop.

Bohdal, O., Li, D., and Hospedales, T. (2023). Label calibration for semantic segmentation under domain shift. In ICLR Workshop on Trustworthy ML.

Bohdal, O., Li, D., Hu, S. X., and Hospedales, T. (2022b). Feed-forward source-free latent domain adaptation via cross-attention. In ICML Pre-training Workshop.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J. J., Yampolskiy, R. V., and Amodei, D. (2018). The malicious use of artificial intelligence: forecasting, prevention, and mitigation. ArXiv.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In FAccT.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare. In KDD.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., and Park, S. (2021). SWAD: Domain generalization by seeking flat minima. In NeurIPS.

Chandra, B. (1989). India's struggle for independence, 1857-1947. Penguin Books, New Delhi, India.
Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research.

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In FAccT.

Cohen, L., Lipton, Z. C., and Mansour, Y. (2020). Efficient candidate screening under multiple tests and implications for fairness. In Symposium on Foundations of Responsible Computing.

Department for Science, Innovation and Technology (2023). A pro-innovation approach to AI regulation. Technical report, UK Goverment.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. In NeurIPS.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In ITCS.

European Commission (2020). On Artificial Intelligence - A European approach to excellence and trust. Technical report, European Union.

Foote, A., Nanda, N., Kran, E., Konstas, I., Cohen, S., and Barez, F. (2023). Neuron to graph: Interpreting language model neurons at scale. In ICLR RTML Workshop.

Ford, M. (2021). Robots: stealing our jobs or solving labour shortages? Guardian.
Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. (2021). Sharpness-aware minimization for efficiently improving generalization. In ICLR.

Francis-Devine, B. and Buchanan, I. (2023). Skills and labour shortages. Technical report, House of Commons Library.
Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., and Olah, C. (2021). Multimodal neurons in artificial neural networks. Distill.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. ArXiv.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., and Badri, O. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In CVPR.

Gulrajani, I. and Lopez-Paz, D. (2021). In search of lost domain generalization. In ICLR.

Haas, M. L. (2017). The Arab Spring: The hope and reality of the uprisings. Routledge, New York, USA.

Hall, M., van der Maaten, L., Gustafson, L., and Adcock, A. B. (2022). A systematic study of bias amplification. ArXiv.

Hao, K. (2019). This is how AI bias really happens—and why it's so hard to fix. MIT Technology Review.

Hardt, M., , Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In NIPS.

Hendrycks, D. (2023). Natural selection favors AIs over humans. ArXiv.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. (2021). Unsolved problems in ML safety. ArXiv.

Hendrycks, D. and Mazeika, M. (2022). X-risk analysis for AI research. ArXiv.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Processing Magazine.

Hoelscher-Obermaier, J., Persson, J., Kran, E., Konstas, I., and Barez, F. (2023). Detecting edit failures in large language models: An improved specificity benchmark. In ACL Findings.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. (2022). Simple data balancing achieves competitive worst-group-accuracy. In Conference on Causal Learning and Reasoning.

Institute for Economics & Peace (2020). Ecological threat register. Technical report.

Irvin, J. A., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K. S., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C., Patel, B. N., Lungren, M. P., and Ng, A. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In AAAI.

Ishii, M. and Sugiyama, M. (2021). Source-free domain adaptation via distributional alignment by matching batch normalization statistics. ArXiv.

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., ying Deng, C., Mark, R. G., and Horng, S. (2019). MIMIC-CXR: A large publicly available database of labeled chest radiographs. Scientific Data.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., and Weller, A. (2022). Synthetic data - what, why and how? ArXiv.

Kelley, C. P., Mohtadi, S., Cane, M. A., Seager, R., and Kushnir, Y. (2015). Climate change in the fertile crescent and implications of the recent Syrian drought. Proceedings of the National Academy of Sciences.

Kelly-Lyth, A. (2021). Challenging biased hiring algorithms. Oxford Journal of Legal Studies.

Khen, H. M.-E., Boms, N. T., and Ashraph, S. (2020). Introduction: An overview of stakeholders and interests. In The Syrian war: Between justice and political reality, pages 1–8. Cambridge University Press.

Khodadadian, S., Ghassami, A., and Kiyavash, N. (2021). Impact of data processing on fairness in supervised learning. In International Symposium on Information Theory (ISIT).

Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. (2019). Learning not to learn: Training deep neural networks with biased data. In CVPR.

Kohavi, R. and Becker, B. (1996). Adult data set. UCI Machine Learning Repository.

Kundu, J. N., Venkat, N., M, R., and Babu, R. V. (2020). Universal source-free domain adaptation. In CVPR.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. H. (2020). Fairness without demographics through adversarially reweighted learning. In NIPS.

Lambrecht, A. and Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Management Science.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences.

Liang, J., Hu, D., and Feng, J. (2020). Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In ICML.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In ICCV.

Lloyd, K. (2018). Bias amplification in artificial intelligence systems. ArXiv.

Loft, P., Sturge, G., and Kirk-Wade, E. (2022). The Syrian civil war: Timeline and statistics. Technical report, House of Commons Library.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations. In ICML.

Martinez, E. and Kirchner, L. (2021). The secret bias hidden in mortgage-approval algorithms. The Markup.

Mehrabi, N., Morstatter, F., Saxena, N. A., Lerman, K., and Galstyan, A. G. (2019). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR).

Miceli-Barone, A. V., Barez, F., Konstas, I., and Cohen, S. B. (2023). The larger they are, the harder they fail: Language models do not recognize identifier swaps in python. In ACL.

Mishler, A. and Dalmasso, N. (2022). Fair when trained, unfair when deployed: Observable fairness measures are unstable in performative prediction settings. In NeurIPS Algorithmic Fairness through the Lens of Causality and Privacy Workshop.

Mukerjee, A., Biswas, R., Deb, K., and Mathur, A. P. (2002). Multi-objective evolutionary algorithms for the risk-return trade-off in bank loan management. International Transactions in Operational Research.

Nardulli, P. F., Peyton, B., and Bajjalieh, J. (2015). Climate change and civil unrest: The impact of rapid-onset disasters. The Journal of Conflict Resolution.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science.

Office of Science and Technology Policy (2022). Blueprint for an AI Bill of Rights: Making automated systems work for the American people. Technical report, The White House.

Olah, C. (2022). Mechanistic interpretability, variables, and the importance of interpretable bases. Transformer Circuits Thread.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. (2020a). An overview of early vision in inceptionv1. Distill.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. (2020b). Zoom in: An introduction to circuits. Distill.

Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. Distill.

OpenAI (2023). GPT-4 technical report. ArXiv.

Pham, T.-H., Zhang, X., and Zhang, P. (2023). Fairness and accuracy under domain generalization. In ICLR.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In NIPS.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In AIES.

Rezaei, A., Liu, A., Memarrast, O., and Ziebart, B. D. (2020). Robust fairness under covariate shift. In AAAI.

Rothe, R., Timofte, R., and Van Gool, L. (2015). DEX: Deep expectation of apparent age from a single image. In ICCV Workshops.

Royer, A. and Lampert, C. H. (2015). Classifier adaptation at prediction time. In CVPR.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2020). Distributionally robust neural networks. In ICLR.

Sarhan, M. H., Navab, N., Eslami, A., and Albarqouni, S. (2020). Fairness by learning orthogonal disentangled representations. In ECCV.

Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., and Bethge, M. (2020). Improving robustness against common corruptions by covariate shift adaptation. In NeurIPS.

Schumann, C., Wang, X., Beutel, A., Chen, J., Qian, H., and Chi, E. H. (2019). Transfer of machine learning fairness across domains. In NeurIPS AI for Social Good Workshop.

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., and Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. Nature medicine.

Singh, H., Singh, R., Mhasawade, V., and Chunara, R. (2021). Fairness violations and mitigation under covariate shift. In FAccT.

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., and Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. BioMed Research International.

Tartaglione, E., Barbano, C. A., and Grangetto, M. (2021). EnD: Entangling and disentangling deep representations for bias correction. In CVPR.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhao, V., Zhou, Y., Chang, C.-C., Krivokon, I., Rusch, W., Pickett, M., Srinivasan, P., Man, L., Meier-Hellstern, K., Morris, M. R., Doshi, T., Santos, R. D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E., and Le, Q. (2022). LaMDA: Language models for dialog applications. ArXiv.

Tocqueville, A. (1856). The old regime and the revolution. Harper and Brothers, New York, USA.

Tolan, S., Miron, M., Gómez, E., and Castillo, C. (2019). Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in Catalonia. In International Conference on Artificial Intelligence and Law.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and efficient foundation language models. ArXiv.

Tujula, M. and Wolswijk, G. (2004). What determines fiscal balances? An empirical investigation in determinants of changes in OECD budget balances. SSRN Electronic Journal.

Urbina, F., Lentzos, F., Invernizzi, C., and Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence.

van den Bos, K. (2020). Unfairness and radicalization. Annual Review of Psychology.

Vapnik, V. (1998). Statistical learning theory.

Viechnicki, P. and Eggers, W. D. (2017). How much time and money can AI save government? Technical report, Deloitte.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. In NeurIPS.

Wang, Z., Qinami, K., Karakozis, I., Genova, K., Nair, P., Hata, K., and Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. In CVPR.

Yang, S., Wang, Y., van de Weijer, J., Herranz, L., and Jui, S. (2021). Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In NeurIPS.

Zhang, M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., and Finn, C. (2021). Adaptive risk minimization: learning to adapt to domain shift. In NeurIPS.

Zhao, H., Coston, A., Adel, T., and Gordon, G. J. (2019). Conditional learning of fair representations. In ICLR.

Zong, Y., Yang, Y., and Hospedales, T. (2023). MEDFAIR: Benchmarking fairness for medical imaging. In ICLR.

Stanford
Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

The Looming Nuclear War

Jean-Pierre Dupuy [1]*

Citation: Dupuy, Jean-Pierre. The Looming Nuclear War. *Proceedings of the Stanford Existential Risks Conference* 2023, 187-192. https://doi.org/10.25740/ks918pb4169

Academic Editor: Steve Luby, Dan Zimmer

Funding: N/A

Conflict of Interest Statement: N/A

Informed Consent Statement: N/A

Acknowledgments: N/A

Author Contributions: N/A

Abstract: Despite the fact that the sole purpose of nuclear weapons in contemporary international security and defense policy relates to deterrence there is a looming nuclear war simply because it is always possible that these weapons would be used at some point. This analysis goes beneath geopolitics to look at nuclear weapons as the formidable and indeterminate tools they are but also offers criticism of the USA and Russia for leaving the Intermediate-Range Nuclear Forces treaty (INF) in 2009, a treaty that Reagan and Gorbachev had signed in 1987.

Keywords: nuclear war, deterrence, pre-emption, Ukraine, tactical weapons

[1] Professor by courtesy of Political Science, Stanford University; Professor Emeritus of Philosophy, École Polytechnique, Paris; jpdupuy@stanford.edu

* Correspondence: jpdupuy@stanford.edu

1. The Beautiful Legend of Nuclear Deterrence

Nuclear weapons are by design weapons of non-use. It is their disproportionate power that excludes anyone from ever thinking of detonating them over a civilian population. So, what is their purpose, as President Trump asked when he received the nuclear code? They have only one purpose: to deter other nuclear powers from using theirs; and, incidentally, to dissuade states or terrorist groups that do not have them from acquiring them. They are weapons of deterrence.

They are certainly not weapons of conquest, by means of which one could gain the upper hand over an adversary by robbing him of an object that he possesses and that one desires to possess, in an attack perspective. Nor, in a perspective of defense, are they weapons by which one would prevent an enemy from seizing an object that one possesses and that the other wants to possess. By "object", I mean a territory (Afghanistan, Ukraine), a zone of influence (Syria), the prestige linked to the size of one's arsenal. No, there is no longer an object or a desire for an object. Violence has reached its peak, where it is only concerned with itself. But that is the good news! It is by their mere existence that nuclear weapons are instruments of deterrence. So, there is no need to worry. Nuclear war will not occur, because it is impossible.

This is what the French "experts", most of whom work directly or indirectly for the French nuclear force, keep telling us about the war that is currently taking place on the battlefield that has become Ukraine, but which is in fact a latent war between Russia and NATO, or, if one prefers, between Putin's Russia and the United States of America. Deterrence is working as well as it did in the past, we are told, and it is very unlikely, if not impossible, that it will lead to nuclear war.

This verbiage is irresponsible and unacceptable. A nuclear war in Europe is unlikely, to be sure, but it is possible. When the stakes are enormous, prudence may lead us to act *as if* the possible were doomed to become real.[1]

2. Nuclear Weapons in the Service of War

Since the beginning of February 2022, even before his troops entered Ukraine on 24 February, Vladimir Putin has constantly warned NATO and the United States of the risk of an escalation that would lead to a nuclear conflict. French leaders got tired of what they took to be a veiled threat, but one that was so lacking in credibility that they did not take it seriously. As a result, no attention was paid to a statement made by the head of the Kremlin to the press on December 9 of the same year announcing a possible change in nuclear doctrine. Until then, the official doctrine was that Russia would only use its arsenal if it was the target of a nuclear attack, or if the survival of the nation was in danger. Putin feigned to discover that the American strategic spectrum included a component called "pre-emption" and that Russia would do well to learn from it in turn.

Preemption is the name of the attack in the nuclear field. To appreciate its singularity, it is useful to return to the logic of deterrence. It has two phases: first, the threat to resort to immeasurable reprisals if the enemy power crosses a certain red line, which is not specified; and, if deterrence has failed, the decision to carry out this threat. France fails to consider this possibility, on the grounds that deterrence cannot fail. Yet this is where the stumbling block to nuclear deterrence lies, namely the non-credibility of the threat of retaliation that underpins it. If deterrence fails, will the attacked power take the risk of triggering, as promised, an escalation leading ultimately to mutual destruction, and

[1] I have laid out the philosophical conditions that justify this move from the possible to the necessary in my book *How to Think About Catastrophe: Toward a Theory of Enlightened Doomsaying* (Dupuy, 2023).

therefore to suicide? Do you have to be crazy or pretend to be crazy to be credible? The answer to this question determines the strength of the deterrence edifice.

Preemption does not bother with this pitfall. It acts as if the first phase were assured: the enemy has crossed the line, or if he has not, he is about to do so. The second phase is therefore justified. What is really a first attack is presented as a reprisal. It is "anticipatory retaliation."[2] Whatever the stated nuclear doctrines, it can be argued that both Soviet and then Russian and American leaders never excluded the decision to strike first from their repertoires of actions (Ellsberg, 2018). However, convincing the enemy that one is ready to do so is no more self-evident than playing the game of deterrence. There is a credibility problem here as well. A first strike will not be sufficient to neutralize the adversary, and the latter will retain the capacity to retaliate: it must therefore be demonstrated that one will be able to ride out its reprisal and limit the damage, and therefore that one will remain fully capable of retaliating to the retaliation. This can be a major challenge.

The United States and Russia have been and continue to be ambivalent about an element of nuclear doctrine that has been given the convoluted and misleading name of "escalate to de-escalate." Their hesitation and vagueness in this regard illustrate the dilemma between deterrence and preemption that both nuclear superpowers face.

The idea of escalation with a view to de-escalation can already be found in Thomas Schelling's seminal book, *The Strategy of Conflict*, and it has influenced several generations of strategists (1960). The doctrine of graduated response advocated by Robert McNamara from the 1960s onwards, the concept of limited nuclear warfare, that of "escalation control", etc. are all variations on the same idea. The simplest way to present it is to compare it to the logic of an auction. Prices are raised until the others can no longer keep up. One gradually increases the intensity of the fighting with conventional forces until the moment when the step to a nuclear strike seems inevitable to end the conflict while winning it: that is what "de-escalation" is.

Both American and Russian strategists recite the credo of nuclear deterrence: one does not deter a limited attack by making a threat of limited retaliation highly credible. One deters it by keeping the probability of mutual annihilation at a moderate level. In practice, however, escalation to de-escalation continues to tempt military planners. This idea is especially present among Russian strategists in their unofficial discussions. According to Alexei Arbatov, a senior national security advisor, "Conventional precision weapons should be capable of inflicting sufficient losses on attacking NATO forces and bases to induce NATO either to stop its aggression, or to escalate it to the level of massive conventional warfare, including a ground offensive. This *would then justify* Russia's first use of tactical nuclear weapons" (Arbatov, 2008).[3]

3. Deterrence Out of the Picture

When asked why no atomic bomb has been dropped on civilian populations since August 9, 1945, the immediate answer is to say that this proves that deterrence has worked. Robert McNamara, the former Secretary of Defense under Presidents Kennedy and Johnson, dismissed the question with "We lucked out. It was luck, just luck, that prevented nuclear war. Dozens of times during the Cold War and ever since, we have come within a hair's breadth of unleashing the horror." Yet there is a more radical way to absolve deterrence of any responsibility for the absence of nuclear war during these eighty-odd years. It is to show that it has very rarely been applied. In the absence of a demonstration, the following episode, which took place at the end of the Cuban missile crisis, is sufficiently suggestive.

[2] The nickname of preemption is "striking second first."
[3] Emphasis mine.

On Saturday 27 October 1962, a Soviet submarine cruising in the Sargasso Sea, north-east of Cuba, was spotted and surrounded by the American aircraft carrier USS Randolph accompanied by several destroyers. The submarine was commanded by Lieutenant Savitsky, flanked by the political officer Maslennikov. The American ships had started to send the signal agreed with the Soviet General Staff to order the enemy submarine to surface. Savitsky was simply not informed of this agreement. As the signal consisted in exploding depth charges near the hull, he believed that he was really attacked by the Americans: a first false alarm or communication error in this story which was to include several, increasingly tragic ones. The circumstances on board were truly hellish. The temperature had reached 50 to 60 degrees Celsius and the men were falling one after the other. To make matters worse, communications with the headquarters in Moscow were cut off.

Savitsky didn't even know if the war had started or not. Exhausted, at the end of his tether, he was about to give the order to launch a few torpedoes with nuclear warheads mounted on them. Yes, the Soviet submarines cruising off Cuba were equipped with atomic bombs. But the Americans did not know this. They learned it forty years later. Savitsky came to his senses and remembered that he needed the approval of his political officer to make such a fatal decision. He agreed.

As luck would have it, Captain Vasily Alexandrovich Arkhipov was on board that day. Although he was of the same rank as Savitsky, he was under Savitsky's command. But he was also the chief of staff of the entire submarine fleet. Savitsky felt it was his duty to get Arkhipov's opinion. Arkhipov disagreed, arguing that Moscow had not given its permission. The firing order was not given and the submarine surfaced.

Hearing such a story, we wonder what would have happened if... A chain of counterfactual propositions immediately comes to mind. If Arkhipov had not been in that troubled submarine but in another submarine, it is highly probable that Savitsky would have given the firing order. The aircraft carrier USS Randolph and its destroyers would have been engulfed in a terrifying nuclear explosion. The American command, convinced that there were no atomic charges in the Soviet submarines, would have inferred that the attack came from Cuba. President Kennedy had already announced on October 22 that if such a thing happened, America would launch an all-out nuclear attack on the Soviet Union. It is easy to imagine what happened next. In the real world, the crisis was resolved peacefully the next day.

Each of the links in this sequence of inferences relates to a contingent event or state of affairs: it might not have happened or it might have been different. But the most fragile part of this story, the most shocking, is that the American command did not know that the Soviet submarines were equipped with nuclear torpedoes. Not that American intelligence was faulty. It was, that's obvious. But what is surprising is that the Soviets did not inform the Americans. If the atomic weapon was really a deterrent, it would have been the least we could do to let the enemy know that we had it and that we were ready to use it. No doubt the aircraft carrier USS Randolph would have been more careful in its approach to the Soviet submarine.

The failure to communicate such crucial information immediately brings to mind Stanley Kubrick's 1964 film *Dr. Strangelove*, in which the concept of the "doomsday machine" appears. The idea is simple, at least on paper. The best way to make credible the threat of immeasurable retaliation that is the basis of deterrence is to make its execution automatic. Gone are the ethical and strategic dilemmas that have plagued heads of state from John Kennedy to Valéry Giscard d'Estaing. In a way, it is the one who shoots first who is responsible for the ensuing holocaust, since the response is not human. In Kubrick's film,

the Soviets invented a machine that would immediately destroy all human life on Earth in response to an American first strike. The problem is that they have not (yet) informed the Americans of its existence when the story begins. However, an enlightened colonel has already launched a B52 armed with H-bombs in the direction of Siberia without authorization and without any possible return. Far from being a parody, this film is a documentary, a member of the Stanford Strategic Research Center, CISAC, recently said.

4. The Crucial Role of Tactical Weapons

Isn't the unprecedented power of the atomic bomb reason enough to dissuade anyone from even thinking about using it? Who would have an interest in triggering an escalation from which all would be defeated? These ideas, with which we began this text, have always been present since 1945 and they retain an undeniable power of conviction. In fact, attempts have been made to reduce both the power of weapons and the range of the missiles that deliver them, in the hope of bringing the devastation produced by a nuclear conflict closer to that of a traditional war, before it was understood that it is, on the contrary, these weapons and missiles, which are said to be "tactical", that should be banished. Their relatively low power[4] encourages us to use them on the battlefield, as we would with a conventional weapon, which amounts to putting our foot in the nuclear gear, which we can show, by a priori reasoning, is destined to go to extremes, that is to say to mutual annihilation. Now here are these tactical weapons more than ever back with the war in Ukraine.

On February 1 and 2, 2019, a double event occurred, unnoticed by public opinion, in France at least, from which the current events are largely derived. The heads of state Trump, first, Putin the next day, announced that they were going to disengage from a treaty, signed in 1987 in Washington by their predecessors Ronald Reagan and Mikhail Gorbachev, by which the two signatories eliminated from their respective arsenals all cruise and ballistic missiles launched from the ground and having a range between 500 and 5,500 km. This treaty was misleadingly named INF (for "Intermediate-Range Nuclear Forces"). It was seriously misleading, as we shall see, because it did not restrict nuclear weapons, but rather a certain type of missile, whether or not it had a nuclear warhead. The American withdrawal became official on August 2, 2019.

With the end of the Cold War in 1989, there was a dramatic reversal in the balance of power between Washington and Moscow in terms of the division between nuclear and conventional weapons. Prior to 1989, the Soviet Union's superiority in conventional weapons was clear, and the United States sought to compensate for its backwardness by developing its nuclear arsenal. After the collapse of the USSR, the Pentagon, proud of the victory of the "free world", i.e. of liberal democracies and market economies, turned its attention to other things, for example to regional conflicts in which conventional weapons proved more effective than atomic bombs. At the same time, Putin in Russia was developing his nuclear arsenal.

It is not nuclear weapons in general that America has relatively neglected, it is mainly tactical nuclear weapons. The doctrine was: conventional weapons on regional battlefields and, if "escalation to de-escalation" so required, the use of strategic nuclear weapons carried by their ICBMs. In 2022, America has only about 100 tactical nuclear warheads in Europe, spread over five countries: Germany, the Netherlands, Belgium, Italy and Turkey. Russia has perhaps twenty times that number.

[4] It can go as far as seven times the explosive power of *Little Boy*, the bomb that destroyed Hiroshima.

With this background, how did the two nuclear superpowers react in 2019 to their mutual rejection of the INF treaty? I recall that this treaty placed a constraint on missiles, whether they carry nuclear warheads or not. The United States and NATO immediately saw the opportunity they now had to place low- and medium-range, non-nuclear-tipped missiles in Europe. This was without counting on the Russian response. This was repeated several times, with Putin calling on the United States and NATO to impose a moratorium on the deployment of such nuclear-armed missiles in Europe. This request went unheeded.

A technical point of considerable importance here is that it is impossible to determine before it reaches its target whether a ballistic missile carries a nuclear warhead or not. Faced with this indeterminacy, Russia has chosen to treat any missile that approaches its territory as a nuclear attack. This is, according to its stated doctrine, sufficient reason for it to launch its own nuclear missiles even before the enemy's missiles reach its soil. This can only make America think twice, as it thought it had a free hand to deploy its missiles in Europe again, both conventional and nuclear. I recall that all this was happening just before Putin decided to invade Ukraine.

This analysis has almost entirely overlooked the geopolitical dimension of the issue. Far be it from me to minimize its importance. I simply wanted to show the decisive power of the tool, in this case the tool of destruction, the atomic weapon. The tool is not neutral, it does not do good or evil according to the intentions of those who wield it. If a nuclear war were to break out in Europe, which none of the actors involved wants, the agent ultimately responsible would not be Putin, Zelensky, Biden or NATO, but the atomic weapon itself and its excessive power. This is what the protagonists of the drama that is being played out feel confusedly, as shown by the extreme caution with which they are advancing their pawns, not without contradictions and a good dose of bad faith. These pretenses and this collective self-deception are undoubtedly necessary to avoid the catastrophe. No, NATO is not at war with Russia, it is simply providing Ukraine with the weapons without which Russia would have crushed it long ago. Can this fool's game go on indefinitely? A clumsy gesture by one or the other can be enough to tip the fiction into the horror of reality.

References

Arbatov, Alexei (2008). “Reducing the Role of Nuclear Weapons.” Paper presented to the International Conference on Nuclear Disarmament, Oslo, Norway 26–27 February 2008.

Dupuy, Jean-Pierre. (2023) *How to Think About Catastrophe. Toward a Theory of Enlightened Doomsaying*. Ann Arbor: Michigan State University Press.

Ellsberg, Daniel. (2018) *The Doomsday Machine*. New York: Bloomsbury.

Schelling, Thomas. (1960) *The Strategy of Conflict.* Cambridge, MA: Harvard University Press.

Section IV

Governance, Policy Infrastructure, and Scenarios

Stanford | Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Collective Intelligence as Infrastructure for Reducing Broad Global Catastrophic Risks

Vicky Chuqiao Yang [1]*, Anders Sandberg [2]

Citation: Yang, VC & Sandberg, A. Collective Intelligence as Infrastructure for Reducing Broad Global Catastrophic Risks. *Proceedings of the Stanford Existential Risks Conference* 2023, 194-206. https://doi.org/10.25740/mf606ht6373

Academic Editor: Paul Edwards, Trond Undheim, Dan Zimmer

Funding: V.C.Y. was partially supported by NSF Grant 2117564

Conflict of Interest Statement: The authors report no conflict of interest

Informed Consent Statement: No human subjects were involved in this study

Acknowledgments: The authors thank Rory Greig for helpful discussions and feedback on an earlier version of the manuscript.

Author Contributions: V.C.Y. led the design of the research and the writing of the manuscript, while A.S significantly contributed to both efforts.

Abstract: Academic and philanthropic communities have grown increasingly concerned with global catastrophic risks (GCRs), including artificial intelligence safety, pandemics, biosecurity, and nuclear war. Outcomes of many, if not all, risk situations hinge on the performance of human groups, such as whether governments or scientific communities can work effectively. We propose to think about these issues as Collective Intelligence (CI) problems—of how to process distributed information effectively. CI is a transdisciplinary research area, whose application involves human and animal groups, markets, robotic swarms, collections of neurons, and other distributed systems. In this article, we argue that improving CI in human groups can improve general resilience against a wide variety of risks. We summarize findings from the CI literature on conditions that improve human group performance, and discuss ways existing CI findings may be applied to GCR mitigation. We also suggest several directions for future research at the exciting intersection of these two emerging fields.

Keywords: global catastrophic risks, collective intelligence, collective behavior, risk intersections

[1] Assistant Professor, MIT Sloan School of Management, Massachusetts Institute of Technology, 100 Main Street, Cambridge MA, USA; vcyang@mit.edu

[2] Senior Research Fellow, Future of Humanity Institute, University of Oxford, Oxford, UK, anders.sandberg@philosophy.ox.ac.uk

* Correspondence: vcyang@mit.edu

We've got to be as clear-headed about human beings as possible, because we are still each other's only hope. — James Baldwin

1. Global Catastrophic Risks

Recent years have witnessed increasing concerns for humanity's survival and prosperity in the long-term future. Scholars are increasingly concerned about global catastrophic events. Examples of such events include nuclear wars, climate catastrophes leading to large-scale failures in agriculture, pandemics, and powerful artificial intelligence (AI) not aligned with human values (see Ord, 2020 for a summary). Other scholars are also concerned with future technological innovations that enable a single individual to have great destructive power, such as DIY biohacking tools that can kill millions (Bostrom, 2019). Scholars studying these potentially catastrophic events and how to mitigate them have formed a transdisciplinary field called global catastrophic risks (GCR), and a subset focused on events that can lead to human extinction is referred to as existential risks. An associated, more practice-focused field, effective altruism, investigates how an individual can do the most good to help others. These areas are not only of concern to academics but have also attracted considerable attention and action in philanthropic efforts.

Discussion in GCR has been centered around which risks are likely to occur and how to circumvent each one. A critical component crosscuts many, if not all, risks but remains less explored—GCR reduction hinges on the better collective performance of human groups. For example, mitigating global pandemic risks includes coordinating individual lifestyle changes, the scientific community's research efforts, and nation-states' policies. We consider factors helping to mitigate many risk scenarios "infrastructures" of the system. Prior research investigated several other such factors, including several risks having the common denominator of disrupting agriculture (Denkenberger et al., 2021); the pace of regulatory innovation being far slower than that of technological innovation (Marchant, 2011); and dangerous technology being accessible to a small number of actors (Bostrom, 2019). The vulnerability of having a weak infrastructure, such as in collective decision-making and collective action, reduces the general stability of the system. Such a system can be compared to a tightrope walker—many forces can lead them to fall, be it a gust of wind or a shake of the body. Compared to focusing on which forces will tip the

Figure 1. *Tightrope walker Margret Zimmermann over Köln in 1946 (Heukeshoven, 1946). The balance pole used by tightrope walkers serves as a device for improving the ability to adaptively balance and adjust to perturbations. The role of Collective Intelligence in Global Catastrophic Risk mitigation is similar to that of the balance pole to a tightrope walker—but performed by a multitude of actors.*

tightrope walker over and prevent these forces from happening, it is more useful to identify tools that can help improve the general stability of the system. For example, adding a balancing pole (Fig 1), which increases the toques needed that lead to a fall and expands the walker's ability to adjust their center of gravity through small pole movements. Similarly, in GCR mitigation, it is essential to consider building general tools, such as the capability for collective decision-making, that improve a system's ability to respond to change. This view is echoed in a perspective piece for collective behavior to be considered a "crisis discipline" (Bak-Coleman et al., 2021). The idea of a system's stability to any perturbation also goes beyond metaphors---it has been rigorously developed in the mathematical theory of dynamical systems, and there has been promising progress (albeit with limitations) in inferring the closeness of a system to its tipping point from time-series data (Pananos et al., 2017; Bury et al., 2022).

While most GCR mitigation efforts center on resolving technological challenges, a significant minority of the GCR community is concerned with human collective decision and action (Liu et al., 2018) and typically frame these issues as coordination problems that can be approached in a game-theoretic framework such as in prisoner's dilemma. While informative in many applications, it can be too narrow and restricting to think about these complex phenomena of collective decisions and action exclusively as problems of coordination. Firstly, this framework typically considers a small number or type of agents with a fixed set of explicit rules. While many hard problems we face are among a large number or type of agents, and the mechanisms of interaction may be implicit or change over time. This distinction is at the core of some most important human endeavors---humans created systems for unrelated individuals to cooperate at a large scale and in a wide range of ways (Melis & Semmann, 2010). This is unique among animals. For example, chimpanzees and other mammals can cooperate in flexible ways but in smaller groups; social insects, such as bees and ants, can work in large numbers but with more rigid roles. Second, besides the consideration of group size and flexible ways of interaction, another issue with viewing the human collective through the lens of coordination is that it tends to dwell on the negative outcomes of poor coordination, like the tragedy of the commons. These are important to avoid; however, making it the sole focus risks neglecting the upsides of the human collective, where the group is more capable than the sum of the individuals within it. The focus on avoiding bad coordination may lead to neglecting how to achieve that. Instead of thinking about these human-collective issues as an issue of coordination, we would advocate for thinking about them more broadly as an issue of Collective Intelligence, one of how to process distributed information effectively in order to reach common goals.

2. Collective Intelligence

Broadly speaking, Collective Intelligence (CI) is concerned with a group's ability to perform a wide variety of tasks or solve a wide variety of problems (Woolley et al., 2010). Since what are relevant problems are subjective to the observer, collective intelligence is referred to by some as groups of individuals acting collectively in ways that *seem* intelligent (Malone et al., 2009). This ability is typically associated with group synergy, that the group outperforms the capability of its individual members (Kurvers et al., 2015).

Examples of CI include democratic elections, prediction markets, and juries. It also goes beyond human groups to include animal groups, such as in a flock of birds deciding the direction for migration; robotic swarms, such as designing rules for individual robots such that the collective can perform certain tasks; and neurons, such as the brain making coherent sense of the world while each neuron responds to different, and sometimes conflicting stimuli (see Fig. 2 for illustrations of these examples). In all these diverse applications, a shared problem is when information is distributed across the individuals

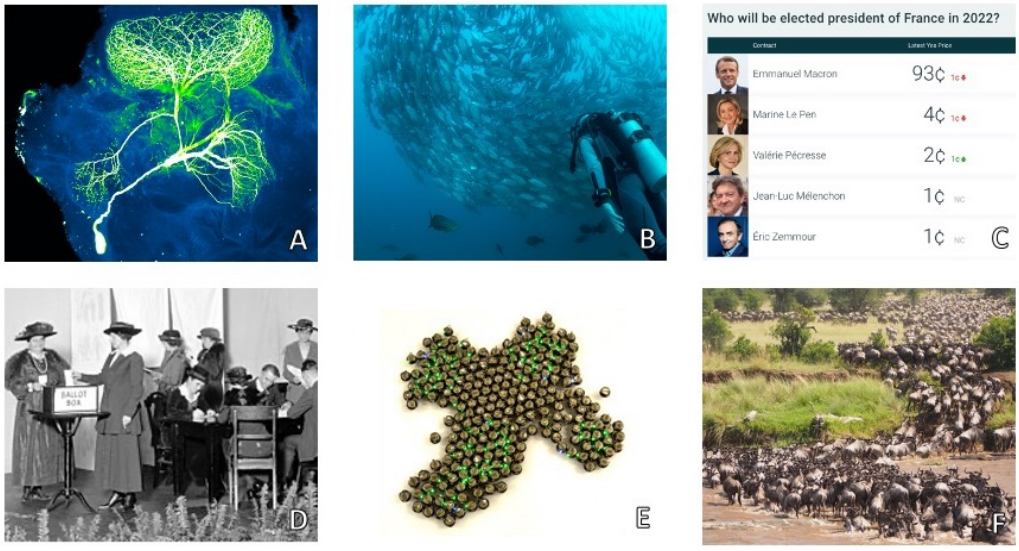


Figure 2. *Examples of collective intelligence systems, which process and aggregate distributed information. (A) Neurons in an insect brain (NICHD, 2015). (B) Fish school respond to the presence of a diver (van de Vendel, 2017). (C) A prediction market for who will be elected president of France in 2022 (Predictit, 2022). (D) Elections. Picture shows women voting in the United States in 1920 (Voting, 1920). (E) A swarm of robots (Carrillo-Zapata, 2019). (F) Great wildebeest migration at Serengeti National Park, Tanzania (Tung, 2019).*

in the system, how to process distributed information effectively and aggregate it together. This distributed information processing can take a wide variety of forms spanning a wide range of complexity. Some examples of more explicit aggregation mechanisms include majority vote in political elections, and quorum voting in African buffalos when deciding on new grazing locations (one signals preference by standing up and facing the favored direction) (Prins, 1996). The more implicit mechanisms involve the schooling of fish, where each fish's preferred direction of movement interacts with that of its neighbors to generate movements of the school. As a result, the school can collectively sense the temperature gradient, while no individual fish is able to sense such information (Puckett et al., 2018). Another example of more implicit aggregation is prediction markets, where individuals bet on the likelihood of an outcome, and the information on the odds is embedded in the prices. Researchers have come together to study these phenomena, crosscutting application domains, and formed a transdisciplinary field around CI. We refer to this field as transdisciplinary instead of interdisciplinary, because it studies a shared phenomenon that manifests in many disciplines, spanning computer science, neuroscience, robotics, animal behavior, and many social sciences. It transcends the disciplinary boundaries, while interdisciplinary often refers to combining two fields to discover new things in either one.

With many rigorous research efforts in the past two decades in Collective Intelligence, researchers have found many factors that help or hurt the performance of human groups in solving problems and making collective decisions, predictions, or estimations. While research does not yet offer recipes for constructing a good human collective, as collective performance is the outcome of complex interactions of many variables. These findings point to a few general directions for improving the collective performance of human groups, and we offer a brief and non-exhaustive summary of relevant findings below.

2.1 Summary of Human Collective Intelligence Findings

Presence of Collective Intelligence. Similar to individual general intelligence, that one individual can excel at a wide range of tasks, such as math and music, a wide range of experiments found a similar property for human groups (Wooley et al., 2010; Wooley et

al., 2015; Riedl et al., 2021), referred to as Collective Intelligence. When asking small groups to perform a wide range of tasks, including creative brainstorming, negotiation, verbal, mathematical, and moral reasoning, some groups outperform others on a wide variety of tasks. Factor analysis shows a single dominant factor explaining 43% of the variance in performance, similar to the explanatory power of individual intelligence on individual tasks (Wooley et al., 2015). These findings suggest the equivalent of general intelligence for groups.

Group social processes more important than individual skills. Researchers have dedicated extensive work into what makes some groups more effective than others. The consensus is that the group's social interaction processes are more important than the skill of individual members (Riedl et al., 2021). Research has found that the group member's social perceptiveness, the ability of individuals to identify social cues, improves group social processes and, consequently, group performance. This is manifested in group behaviors such as even conversational turn-taking—each group member speaks roughly equal amounts. Consequently, groups with a higher proportion of women tend to have higher collective intelligence, as being female is correlated with greater social perceptiveness. The IQ of individual members plays a much less, and some argue, negligible role.

Diversity. Besides the social perceptiveness of group members, research finds benefit in having a diverse group of individuals. The diversity referred to here is diversity in knowledge and cognitive models. Abundant research has found, in lab experiments, theoretical and computational models, and real-world problem-solving scenarios, the phenomenon of a "diversity bonus"—that a diverse group performs better than a homogenous group (Page, 2019; Aminpour et al., 2021). Some even find a diverse group of non-experts can outperform a homogeneous group of experts (Hong & Page, 2004). For a group of diverse agents to work together, an important aspect is cognitive alignment, such as commitment to group goals and shared beliefs (Krafft, 2019). Another line of work finds that maintaining diversity in the group is hard—conformity and traditional market forces are against it. Mann & Helbing (2016) propose alternative incentives to reward accurate minority predictions for maintaining diversity in the group—especially rewarding the minority which is right when the majority is wrong.

Committed minorities. There has been much evidence that the presence of committed minorities, a small fraction of individuals who are little affected by the opinions of others, can lead to substantial changes in the collective behavior of a group. Most notably, Centola et al. (2018) find through human subject experiments that a critical mass of committed individuals (around 25% in their experiments) can tip social conventions towards the direction of these committed individuals. This critical transition is also predicted by an abundance of theoretical studies (such as Xie et al., 2011). It would require further investigation to understand the precise critical mass needed in different scenarios; however, the powerful effects of committed minorities on groups can be a fruitful direction in thinking about how to elicit (or prevent) social change.

Social influence. An area under debate is whether and how individuals should communicate with each other in a collective. Experimental studies have found that social influence can lead to both positive and negative effects on collective performance (see Jayles et al., 2017 for an example of positive effect, and Lorenz et al., 2011 for an example of negative effect). On the one hand, letting individuals exchange information can lead to the loss of independent information, and worse collective performance (also referred to as groupthink). On the other hand, communication may help individuals deliberate and discover better answers. The conflicting findings are likely an outcome of the effect of social influence depending on several other variables. A study predicts it depends on the

proportion of individuals using social information for their decisions, and whether committed minorities are present (Yang et al., 2021). Others find they also depend on the social network structure and adaptability (Almaatouq et al., 2020; Becker et al., 2017). The bottom line is that most people blindly following others does not lead to good outcomes in most scenarios.

Better sensors. One important component for improving group performance is to let individuals gather better information. This is especially important in forecasting and prediction tasks, such as election forecasts. This can be achieved by asking better survey questions. For example, in a method called "surprisingly popular" (Prelec et al., 2017), instead of asking people what they think, researchers also ask people what they think the majority thinks. This helps uncover surprising outcomes whose signals are suppressed by social norms, such as Trump winning the 2016 US election. Another example is to use individuals' social circles as better sensors (Galesic et al., 2021)—instead of asking individuals whom they will vote for, ask whom their friends will vote for.

The literature in this field is vast, and here we highlight a small subset of findings relevant to the performance of human groups. One theme that appears in these scientific studies is that CI requires certain conditions to appear---individuals' skills matter less than how the individuals interact, and individuals being highly correlated tend to hurt group performance.

Most CI research on human groups, such as those summarized above, focuses on improving estimation accuracies, the effectiveness of teamwork, and collective decision speed. Many aspects of these investigations have implications for GCR mitigation. Below we outline examples of current GCR mitigation efforts and discuss how CI research can be applied to mitigating GCR.

3. Collective Intelligence for Global Catastrophic Risk Mitigation

CI has been increasingly considered a useful part of crisis response (Vivacqua & Borges, 2010; Büscher & Thomas, 2014), and can be useful in GCR mitigation. Mitigation of GCR can naturally be divided into prevention (prevent risks from occurring), response (react before or during the hazard to limit the damage), and resilience (survive and rebuild), with a key role of adaptive collective responses to implement these mitigation measures (Cotton-Barratt et al., 2020). CI can be applied in each of these layers, in particular by enhancing predictive and adaptive ability (See Fig. 3 for a conceptual illustration of the relationships among these concepts).

3.1 Prediction

For prevention, identification of a possible risk and action to reduce its probability (or at least severity) are needed. The identification step is where committed minorities have historically played a clear role, often starting in the scientific community. For example, Clair Cameron Patterson built up a network concerned about the spread of lead in the environment, and the atmosphere scientists of the Crutzen-Rowland-Molina group discovered and documented ozone depletion. Their persistent research, documentation,

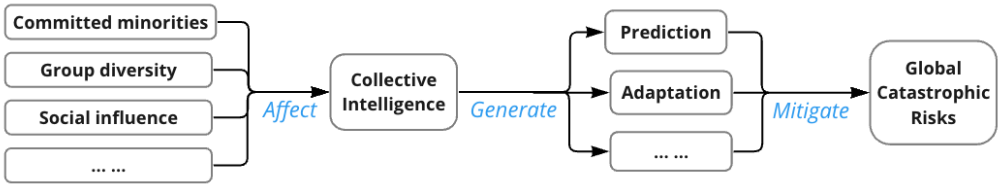


Figure 3. *A conceptual illustration of the concepts in this article and their connections. Collective intelligence as infrastructure for mitigating global catastrophic risks through avenues such as prediction and adaptation.*

and outreach provided the signal needed for scientific societies and later international organizations to react to the risk. Currently, the planetary defense community working on the impact risks of asteroids and the AI safety community are examples of committed minorities having the specialized concern needed to determine the risk and propose preparedness actions. These are examples of how the scientific and policy community can act with CI, by having specialized networks examining potential threats and—if the evidence becomes compelling—amplifying their signal through a social influence mechanism.

Predicting GCRs ahead of time is challenging because they represent unprecedented problems, or occur in domains with very little data. There has been significant interest in estimating the likelihood of different GCRs, especially due to the need for prioritization and making trade-offs among risks. This has been done using a variety of methods that aggregate expert or public opinion (Beard et al., 2020). At the simplest, these methods consist of taking the median estimate, while more sophisticated approaches give weightings based on consistency (Prelec, 2004; Frank et al., 2017) or past predictive performance, involve structured deliberative processes to build a consensus, or are prediction markets where (updateable) bets are made.

One mechanism for prediction discussed at length in the CI literature is prediction markets, where participants bet on future outcomes, and the prices of bets indicate the joint belief in their likelihoods (Wolfers & Zitzewitz, 2006). Here, distributed and diverse knowledge is aggregated through a price mechanism. In theory, attempts to manipulate the price can even increase the accuracy by injecting capital exploitable by informed participants (Hanson & Opera, 2009). Existing prediction markets contain bets on various natural or anthropogenic disasters. In an experiment, a prediction market was applied to forecasting epidemic disease rates and performed more accurately than predictions from extrapolating historical trends (Li et al., 2016). However, prediction markets have so far been limited by anti-gambling laws in many countries, lack of liquidity/participants, and the time preferences of participants.

Similar to research on prediction markets, work on forecasting tournaments, which is not limited by anti-gambling laws, has produced methodologies for eliciting short- and medium-term forecasts that are consistently better than individual expert forecasts, demonstrating CI. These methodologies involve weighted aggregation of many individual estimates, identifying top performers and grouping them together into teams, and supporting learning, updating, and debiasing (Tetlock & Gardner, 2016; Tetlock et al., 2017). Inspired by the success of forecasting tournaments and prediction markets, metaculus.com is a community for group forecasting using many of the same aggregation mechanisms, plus others like gamified scoring and reputation to encourage higher accuracy. The site covers a vast variety of real-world events, including disasters all the way up to existential risk. The site also did well compared to both the public and experts in forecasting during the COVID-19 pandemic (Recchia et al., 2021).

Often preparation, rapid detection, and response are more feasible than predicting the precise moment a GCR will happen. Rapid detection can draw on the distributed and sensitive nature of CI to find early signals of a hazard, amplify the warning, and trigger pre-planned responses. One experiment demonstrated how rapid social mobilization using social media and an incentive mechanism to solve an "impossible" search problem (finding ten weather balloons located across the US) was able to succeed within 9 hours. Here the winning team CI was enabled by an incentive mechanism that rewarded not just direct search but recruiting likely candidates, plus information sharing across the participants. Other teams recruited suitable pre-existing communities (Pickard et al., 2011). Similar experiments have demonstrated fast finding of people (<12 hours) globally (Rutherford et al., 2013). These results suggest that rapid, CI-mediated crisis response such as finding pathogen carriers or dangerous objects is possible.

3.2 Adaptation

After the initial forecast, adaptivity may become essential. Decision-making and governance relevant to global catastrophic risk are complex, and naturally take place inside a collective cognitive system that is part of the greater system: complexity, reflexivity, and uncertainty cannot be avoided (Fisher & Sandberg, 2022). Responding to GCR requires adaptive governance. In evolving disaster situations, flexible solutions and the ability to innovate on the fly can be crucial. The responses need to be more adaptive than the hazard, which typically implies a great need for generating diverse solutions, rapid spread of successful solutions, and forming tight feedback loops to help evaluate and improve the solutions. CI has a clear advantage here over centralized control systems that tend to act slowly and with less information. Such top-down approaches may be more suited for pre- or post-disaster preparation and systemic recovery where timescales are long, optimized global solutions perhaps possible and desirable, and especially when a severe GCR needs to be avoided at all costs. Adaptive governance has demonstrated success in managing common-pool resources in a decentralized fashion (Ostrom, 1990). Safety can be thought of as a special common pool resource that can be managed in this way. GCR raises the stakes by making the threat global and hence requires global adaptive governance, a challenge for CI.

One component that can improve adaptive governance in a GCR situation is effective data-sharing, a lesson learned especially from the COVID-19 pandemic. Data aggregation sites such as ourworldindata.org played a key role in providing various agencies and networks with vital information while they struggled to get data from governments in useful formats. To enable effective CI, constructing better data-sharing infrastructure may be necessary.

Another component that can improve adaptive governance is resiliency. Resiliency can be viewed as the capacity of a system exposed to a shock or stress to adapt and survive by changing its non-essential attributes and rebuilding itself (Downes, 2013). Resiliency, both during and after a catastrophe, is often aided by collective memory: adaptations to past disasters remain distributed among community members (and sometimes, stigmergically (Marsh & Onof, 2008), in the environment like Japanese tsunami stones reminding of long-term danger). Collective memory may have a limited range (Fanta et al., 2019) but can both exceed individual lifespans and allow distributed reactions. It also interacts with other collective attitudes such as risk-taking and trust (Viglione et al., 2014). Another form of collective risk cognition affecting resiliency is via insurance pricing, where insurers set premiums based on aggregated and estimated risk, creating incentives. For example, in response to tsunami risks, insurance incentives can incentivize homeowners and builders to build where the risks are reduced.

Many forms of adaptations to risk in human society take the form of institutions, whether formal (such as fire services, disaster agencies, the Intergovernmental Panel on Climate Change) or informal (such as cautionary tales, safety practices). They can be seen as integrated systems of rules that structure social interactions (Hodgson, 2015): hence a way of structuring or creating CI that (when done well) promotes resilience. This institution-forming often occurs in a reactive way when disaster strikes, which may not be acceptable for the most serious risks.

CI is often described as spontaneously emergent, but this is rare. Usually, it is enabled by the existence of certain structures—suitable abilities, platforms, and incentives that make forming a shared cognitive system easier and allow it to have strong effects. The CI literature suggests such structures need to give preference to cognitively diverse and socially attuned individuals, and egalitarian group social processes. Individual influences need to be homogeneously distributed, and there is not too much correlation in the system, such as a few individuals having an overwhelming influence on the group. Both sets of ideal properties generating CI are counter to elements in the current Western society. In companies, individual employees are often evaluated based on individual-level output, rather than how well they facilitate teamwork. In online social networks, a small number of actors can have an overwhelming influence on the overall network. The CI literature suggests these conditions could hinder collective intelligence. Thus, there is a need to use CI research to design and improve these structures—such as the types of institutions and market incentives—for adaptive governance.

4. Discussion

Systems can be stable by resisting shocks directly ("toughness"), by returning to their equilibria through internal feedback (homeostasis, such as how mammals regulate body temperature), by dynamic stability (constant nudging the system to stay in an intrinsically unstable equilibrium, for example, riding a bicycle), to adaptive stability (the system changes structure to respond to the shocks, for example, buildings add safety stairs and smoke alarms to improve resilience to fires). In human systems facing disaster, sufficient toughness can rarely be achieved but feedback and adaptation are both possible and common. CI represents the distributed, bottom-up approach to achieving these without central control, which can either work on its own or in combination with top-down governance.

A central issue in GCR mitigation is how to reconcile individuals with conflicting goals and interests. This is especially manifested in concerns derived in game-theoretic analyses, such as prisoners' dilemma and tragedy of the commons. Collective intelligence offers a broad and, we think, hopeful, perspective on this issue. These differences may first appear as an obstacle to coordination, while in the collective intelligence perspective, these differences, if harnessed with the right aggregation mechanism, are a source of strength. Take the example of a flock of fish. Each fish senses their local environment—food, temperature, potential predator, etc., and has a different preferred direction of where to go next. Nevertheless, the collective effectively aggregate information from all fish through local interactions in movement (Lopez et al., 2012). The collective's movement responds to a much wider environment, more than the range of any individual fish. Thus, collective intelligence can be generated from different individuals acting on different information and wanting different outcomes. The differences should be seen as a resource, and the key question should be how to harness and aggregate the individuals' differences in productive ways.

It can be useful to take advantage of the transdisciplinary nature of CI and learn from other disciplines. Biological systems, through evolution, have devised efficient solutions

to complex CI problems. Besides the example of fish, each neuron in the human brain receives different, and sometimes conflicting information, but the brain harnesses the inconsistent signals into a cohesive understanding of the world. What systems in nature face similar challenges in processing distributed information as we do in collective action and decision-making? Following the biomimicry concept in engineering, social system researchers could gain valuable lessons from biological systems' approaches to CI, informing solutions to societal challenges.

5. Conclusions

The emerging field of GCR mitigation presents significant challenges with substantial safety implications. Insights from CI research can provide valuable guidelines to aid policymakers in their decision-making processes, particularly in incorporating diverse and socially-aware groups and prioritizing egalitarian interactions. CI research highlights the role of social processes within a group's capacity for collective decision-making. Future studies should examine how shifts in social interaction and media distribution affect large-scale decisions, such as those made in democratic elections. While most CI studies occur in routine contexts, such as brain-storming or medical diagnosis, future research should examine whether the factors improving group performance in these situations also apply to high-stake, urgent situations like addressing GCR mitigation, affecting a vast population.

CI research also suggests new strategies for GCR mitigation, like leveraging committed minorities for prediction and utilizing collective memory for adaptation. Beyond these areas, one promising direction of future research is collective intelligence in human-AI groups, where the different kinds of cognition may complement each other (Guszcza et al., 2022). AI methods can obviously outsource some human cognitive tasks, but could also perform tasks not suited to human thinking (e.g., watching for high dimensional patterns). However, how to design human-AI collaborations well remains under-explored. Another important topic worth further analysis is how the learnings from CI research can be used in the existing disaster management frameworks such as the Sendai framework and the EU Civil Protection Mechanism. In many ways, the frameworks represent a form of institution-based CI already applied to risk management, but it is plausible that other CI methods may help improve their performance if they can be integrated with existing structures.

We encourage CI researchers to think more about using CI in practice, especially in GCR scenarios representing perhaps the most important societal challenges. We encourage GCR researchers to expand their approach, which is centered on studying the physical sciences behind each risk scenario, to further engage with the behavioral sciences, especially the study of collective intelligence, to formulate useful theoretical frameworks for the reduction of risks across domains. We encourage general behavioral researchers to become aware and take advantage of the transformative opportunities of making an impact on the most important societal challenges through engaging with these two transdisciplinary efforts.

References

Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaid, M., & Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, *117*(21), 11379-11386.

Aminpour, P., Gray, S. A., Singer, A., Scyphers, S. B., Jetter, A. J., Jordan, R., & Grabowski, J. H. (2021). The diversity bonus in pooling local knowledge about complex problems. *Proceedings of the National Academy of Sciences*, *118*(5).

Bak-Coleman, J.B. et al. (2021). Stewardship of global collective behavior, *Proceedings of the National Academy of Sciences*, *118*(27).

Beard, S., Rowe, T., & Fox, J. (2020). An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards. *Futures*, 115, 102469.

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 114(26), E5070-E5076.

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, 10(4), 455--476.

Bury, T. M., Sujith, R. I., Pavithran, I., Scheffer, M., Lenton, T. M., Anand, M., & Bauch, C. T. (2021). Deep learning for early warning signals of tipping points. *Proceedings of the National Academy of Sciences*, 118(39), e2106140118.

Büscher, M., Liegl, M., & Thomas, V. (2014). Collective intelligence in crises. In *Social Collective Intelligence* (pp. 243-265). Springer, Cham.

Carrillo-Zapata, D., Sharpe, J., Winfield, A. F. T., Giuggioli, L., & Hauert, S. (2019). Toward controllable morphogenesis in large robot swarms. *IEEE Robotics and Automation Letters*, 4(4), 3386-3393.

Centola, D., Becker, J., Brackbill, D., & Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, *360*(6393), 1116-1119.

Cotton, Barratt, O., Daniel, M., & Sandberg, A. (2020). Defence in depth against human extinction: Prevention, response, resilience, and why they all matter. *Global Policy*, 11(3), 271-282.

Denkenberger, D., Sandberg, A., Tieman, R., & Pearce, J. M. (2021). Long Term Cost-Effectiveness of Resilient Foods for Global Catastrophes Compared to Artificial General Intelligence Safety (No. vrmpf). *Center for Open Science*.

Downes, B. J., Miller, F., Barnett, J., Glaister, A., & Ellemor, H. (2013). How do we know about resilience? An analysis of empirical research on resilience, and implications for interdisciplinary praxis. *Environmental Research Letters*, 8(1), 014041.

Fanta, V., Šálek, M., & Sklenicka, P. (2019). How long do floods throughout the millennium remain in the collective memory? *Nature Communications*, 10(1), 1-9.

Fisher, L. & Sandberg, A. (2022). A Safe Governance Space for Humanity: Necessary Conditions for the Governance of Global Catastrophic Risks. *Global Policy*.

Frank, M. R., Cebrian, M., Pickard, G., & Rahwan, I. (2017). Validating Bayesian truth serum in large-scale online human experiments. *PloS ONE*, 12(5), e0177385.

Galesic, M., de Bruin, W. B., Dalege, J., Feld, S. L., Kreuter, F., Olsson, H., & van der Does, T. (2021). Human social sensing is an untapped resource for computational social science. *Nature*, 1-9.

Guszcza, J., Danks, D., Fox, C. R., Hammond, K. J., Ho, D. E., Imas, A., ... & Woolley, A. W. (2022). Hybrid Intelligence: A Paradigm for More Responsible Practice. *Available at SSRN*.

Hanson, R., & Oprea, R. (2009). A manipulator can aid prediction market accuracy. *Economica*, 76(302), 304-314.

Hodgson, G. M. (2015). On defining institutions: rules versus equilibria. *Journal of Institutional Economics*, 11(3), 497--505.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385-16389.

Heukeshoven, D. (Photographer) (1946). Tightrope walker Margret Zimmermann over Köln in 1946 [Photograph]. Wikimedia Commons. Retrieved April 5, 2022, from https://commons.wikimedia.org/wiki/File:Seilt%C3%A4nzerin_Rosanna_%C3%BCber_K%C3%B6ln_in_1946_(aus_K%C3%B6ln_und_der_Krieg._Leben,_Kultur,_Stadt._1940_-1950).jpg

Jayles, B., Kim, H. R., Escobedo, R., Cezera, S., Blanchet, A., Kameda, T., ... & Theraulaz, G. (2017). How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences*, *114*(47), 12620-12625.

Li, E. Y., Tung, C. Y., & Chang, S. H. (2016). The wisdom of crowds in action: Forecasting epidemic diseases with a web-based prediction market system. *International Journal of Medical Informatics*, 92, 35-43.

Liu, H. Y., Lauta, K. C., & Maas, M. M. (2018). Governing Boring Apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. *Futures*, *102*, 6-19.

Lopez, U., Gautrais, J., Couzin, I. D., & Theraulaz, G. (2012). From behavioural analyses to models of collective motion in fish schools. *Interface Focus*, *2*(6), 693-707.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020-9025.

Mann, R. P., & Helbing, D. (2017). Optimal incentives for collective intelligence. *Proceedings of the National Academy of Sciences*, *114*(20), 5077-5082.

Marsh, L., & Onof, C. (2008). Stigmergic epistemology, stigmergic cognition. *Cognitive Systems Research*, 9(1-2), 136-149.

Marchant, G. E. (2011). The growing gap between emerging technologies and the law. In *The Growing Gap Between Emerging Technologies and Legal-ethical Oversight* (pp. 19-33). Springer, Dordrecht.

Melis, A. P., & Semmann, D. (2010). How is human cooperation different? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1553), 2663-2674.

National Institute of Child Health and Human Development (NICHD) (2015), A neuron in an insect brain [Photograph]. Wikimedia Commons. Retrieved April 5, 2022, from https://commons.wikimedia.org/wiki/File:Insect_neuron.png

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, *2*(2), 126-132.

Krafft, P. M. (2019). A simple computational theory of general collective intelligence. *Topics in Cognitive Science*, *11*(2), 374-392.

Kurvers, R. H., Krause, J., Argenziano, G., Zalaudek, I., and Wolf, M. (2015a). Detection accuracy of collective intelligence assessments for skin cancer diagnosis. *JAMA Dermatol*. 151, 1346–1353. doi: 10.1001/jamadermatol.2015.3149

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

Page, S. (2019). *The Diversity Bonus*. Princeton University Press.

Pananos, A. D., Bury, T. M., Wang, C., Schonfeld, J., Mohanty, S. P., Nyhan, B., ... & Bauch, C. T. (2017). Critical dynamics in population vaccinating behavior. *Proceedings of the National Academy of Sciences*, 114(52), 13762-13767.

Pickard, G., Pan, W., Rahwan, I., Cebrian, M., Crane, R., Madan, A., & Pentland, A. (2011). Time-critical social mobilization. *Science*, 334(6055), 509-512.

Predictit (2022). Who will be elected president of France in 2022? Retrieved March 28, 2022, from https://www.predictit.org/markets/detail/7360/Who-will-be-elected-president-of-France-in-2022.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, 306(5695), 462-466.

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532-535.

Prins, H. (1996). *Ecology and Behaviour of the African Buffalo: Social Inequality and Decision Making* (Vol. 1). Springer Science & Business Media.

Puckett, J. G., Pokhrel, A. R., & Giannini, J. A. (2018). Collective gradient sensing in fish schools. *Scientific Reports*, 8(1), 1-11.

Recchia, G., Freeman, A. L., & Spiegelhalter, D. (2021). How well did experts and laypeople forecast the size of the COVID-19 pandemic? *PloS ONE*, 16(5), e0250935.

Riedl, C., Kim, Y. J., Gupta, P., Malone, T. W., & Woolley, A. W. (2021). Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences*, *118*(21).

Rutherford, A., Cebrian, M., Rahwan, I., Dsouza, S., McInerney, J., Naroditskiy, V., ... & Miller, S. U. (2013). Targeted social mobilization in a global manhunt. *PloS ONE*, 8(9), e74628.

Surowiecki J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The Art and Science of Prediction*. Random House.

Tetlock, P. E., Mellers, B. A., & Scoblic, J. P. (2017). Bringing probability judgments into policy debates via forecasting tournaments. *Science*, 355(6324), 481-483.

Tung, J. (Photographer) (2019), Great wildebeest migration crossing Mara river at Serengeti National Park - Tanzania [Photograph]. Retrieved April 5, 2022, from https://unsplash.com/photos/1pZJqQlgpsY

van de Vendel, P. (Photographer) (2018). Jackfish Tornado [Photograph]. Retrieved April 5, 2022, from https://unsplash.com/photos/gcG_b9ijyqU.

Viglione, A., Di Baldassarre, G., Brandimarte, L., Kuil, L., Carr, G., Salinas, J. L., ... & Blöschl, G. (2014). Insights from socio-hydrology modelling on dealing with flood risk–roles of collective memory, risk-taking attitude and trust. *Journal of Hydrology*, 518, 71-82.

Vivacqua, A. S., & Borges, M. R. (2010, April). Collective intelligence for the design of emergency response. In *The 2010 14th International Conference on Computer Supported Cooperative Work in Design* (pp. 623-628). IEEE.

Voting (1920) Women learn to vote at NCR in Dayton on Oct. 27, 1920. Wikimedia Commons. Retrieved April 5, 2022, from https://commons.wikimedia.org/wiki/File:Women_practice_voting_in_Dayton_Oct._27,_1920.jpg.

Wolfers, J., & Zitzewitz, E. (2006). *Interpreting Prediction Market Prices as Probabilities*. National Bureau of Economic Research.

Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015). Collective intelligence and group performance. *Current Directions in Psychological Science*, 24(6), 420-424.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004), 686-688.

Xie, J., Sreenivasan, S., Korniss, G., Zhang, W., Lim, C., & Szymanski, B. K. (2011). Social consensus through the influence of committed minorities. *Physical Review E*, *84*(1), 011130.

Yang, V. C., Galesic, M., McGuinness, H., & Harutyunyan, A. (2021). Dynamical system model predicts when social learners impair collective performance. *Proceedings of the National Academy of Sciences*, 118(35).

Stanford | Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Convergence on Existential Risk Policy

Philip Arthur [1*]

Citation: Arthur, Philip. Convergence on Existential Risk Policy. *Proceedings of the Stanford Existential Risks Conference* 2023, 207-219. https://doi.org/10.25740/wc416vz8616

Academic Editor: Dan Zimmer, Trond Undheim

Funding: N/A

Conflict of Interest Statement: N/A

Informed Consent Statement: N/A

Acknowledgments: Special thanks to Dan Zimmer and Tyler DesRoches for discussions and commentary surrounding this essay.

Author Contributions: N/A

Abstract: Nick Bostrom, Toby Ord, and Will MacAskill argue that humans face rapidly increasing existential risks this century. Given these authors focus on technology as both the villain and the hero of our existence, Zoe Cremer and Luke Kemp categorize them as taking a Techno Utopia Approach (TUA) towards existential risk. While recognizing merits to this approach, they also argue that a more democratic methodology and diverse set of approaches is needed. Joshua Schuster, Derek Woods, and Emile Torres have also taken a critical view of the (TUA), arguing that it fails to consider the interests of marginalized populations and other species. Given the more diverse scope of existential threats identified by these authors, and the shorter-term motivation of their concerns, I categorize them as taking a Shorter-term Pluralist Approach (SPA).

This paper aims to explore the following question: given the different values of the (TUA) and the (SPA), where might their support of policies related to existential risk converge? Answering this question is important, because while theoretical discussions have independent intellectual value for many, policies associated with existential risks have potential consequences for everyone. This discussion can help scholars and policymakers form alliances, better understand their rivals, and more effectively allocate resources to practical areas of research and debate.

Keywords: existential risk, policy, value, TUA

[1] PhD Student, School of Sustainability, Arizona State University, parthur@asu.edu

* Correspondence: parthur@asu.edu

1. Introduction

Nick Bostrom, Toby Ord, and Will MacAskill argue humans face rapidly increasing existential risks this century (Ord, 2020; Bostrom, 2013; MacAskill, 2022). This follows primarily because while rapidly evolving technologies such as gene editing and artificial intelligence create immense potential for human advancement, they also create exponentially more power to destroy—increasing the odds that a rogue individual, group, or machine could intentionally or unintentionally use these technologies to wipe out humanity. These thinker's define existential risk and existential catastrophe as follows: "An *existential risk* is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development" (Bostrom, 2013); "An *existential catastrophe* is the destruction of humanity's long-term potential" (Ord, 2020). Important words in these definitions are "desirable" and "potential." Existential threats are not just those that would eliminate all lives, but ones that would eliminate desirable lives (Bostrom, 2002); moreover, we should not simply preserve the status quo and settle for good lives, but protect human potential and the ability to improve the quality of life. This optimism involves speculation about drastically higher levels of human well-being in the future from Ord (2020) and an explicit endorsement of human enhancement and transhumanism from Bostrom (2008). Because of this optimism, these scholars believe future lives are just as valuable and perhaps even more valuable than current lives; therefore, protecting such lives should play an important role in our current ethical reasoning and policy formation.

Given Ord, Bostrom, and MacAskill's focus on technology as both the villain and the hero, Carla Cremer and Luke Kemp (2021) have labeled this school of thought the Techno Utopia Approach (TUA). And while not denying useful aspects of this approach, they argue that a more pluralistic and comprehensive approach is needed to effectively mitigate existential risks. By "Democratizing Risk," Cremer and Kemp say we can avoid biases in identifying and weighing risks and develop more effective policies to mitigate such risks. Joshua Schuster and Derek Woods (2021) also take a critical view of the TUA in *Calamity Theory: Three Critique of Existential Risk*, arguing it advances a problematic form of probabilistic epistemology, fails to engage with traditional existentialist literature, and fails to consider marginalized populations and a diverse set of attitudes towards extinction. Emile Torres (2021) advances the overarching concern with the TUA that by focusing attention on potential human lives in the far future, we might neglect important concerns of human lives today. Torres (2021) says,

> Why do I think this ideology is so dangerous? The short answer is that elevating the fulfilment of humanity's supposed potential above all else could nontrivially increase the probability that actual people – those alive today and in the near future – suffer extreme harms, even death. Consider that, as I noted elsewhere, the Longtermist ideology inclines its adherents to take an insouciant attitude towards climate change. Why? Because even if climate change causes island nations to disappear, triggers mass migrations and kills millions of people, it probably isn't going to compromise our longterm potential over the coming trillions of years.

Given a broader and more diverse scope of existential risks identified by Schuster and Woods, Cramer and Kemp, and Torres, and the shorter-term motivation of their concerns, I categorize these thinkers as taking a Shorter-term Pluralist Approach (SPA).

This paper examines the TUA alongside the SPA, with the aim to answer the following question: given the different values and motivations embraced by the TUA and the SPA, where might their attitudes towards policies related to existential risk converge and diverge? Answering this question is important, because while discussions of existential

risk have independent intellectual value for many in academia, policies associated with existential risk have consequences for almost everyone. Accordingly, I hope to elucidate areas where different values and attitudes towards existential risk may result in different policy stances, areas where different attitudes are likely to converge towards similar policy goals, and areas where there is little clarity. This discussion can help scholars and policymakers better form alliances, better understand their rivals, and more effectively allocate resources to areas of research that might be most practically and morally relevant.

2. Norton's Convergence Hypothesis

In *Toward Unity Among Environmentalists*, Bryan Norton (1994) develops a *convergence* hypothesis, positing that while different environmentalist groups might embrace different core values, their interests will likely converge towards similar policy goals. For example, a non-anthropocentrically motivated group of Deep Ecologists might petition to preserve a forest to maintain species' right to exist without human interference; on the other hand, an anthropocentrically motivated group of outdoor enthusiasts or economists might advocate preserving the same forest because of its beauty and various ecosystem services. Regardless of whether the forest has intrinsic or instrumental value, these groups are likely to favor policies aimed at forest preservation.

However, in the current existential risk literature, rather than highlighting areas of aligned interest and policy convergence, scholars with different values often focus on hypothetical scenarios that suggest their interests are in competition (that we might have to decide between mitigating climate change today and trillions of happy lives in the far future). From a policy perspective we need evidence of how theoretical conflicts might become actual ones and how these causal chains between the short and long term play out in the real world. For example, regardless of whether one categorizes climate change happening this century as an existential threat, climate change will cause major environmental, economic, and social disruption for those living today; moreover, it will also impact the nature and likelihood of other existential risks and be a major contributor to the threat of cascading risks (Kemp et al., 2022; Undheim, Forthcoming).

Accordingly, even if TUA scholars are most worried about engineered pandemics, unaligned AGI, and nuclear warfare as existential threats, it might be the case that one of those most effective ways to mitigate such threats is by mitigating climate change. In the same vein, while the SPA rejects a focus on the far distant future and techno utopian daydreaming as a guide to policy, it is not obvious that their shorter term and pluralistic focus is not the most promising method to one day arrive at such a utopia (a diverse group of people may one day decide to become posthumans). It might be the case that minimizing human suffering, democratizing decision making, and advancing technology with egalitarian concerns creates a relatively stable and cooperative world that makes such a utopian future most likely. Moreover, it might also be the case that thinking about the far future can spur action towards shorter-term humanitarian goals.

However, while such convergence among existential risk thinkers might be auspicious, unlike Norton my analysis is not committed to the idea that TUA and SPA will almost always converge on important policy decisions surrounding existential risk. Rather than arguing that moral differences will collapse towards similar policy outcomes, this paper aims to better understand how different value systems affect policy, exploring areas of both convergence and divergence. I will begin exploring a variety of topics related to policy surrounding existential risk including pandemics and public health, climate change, non-anthropocentric extinction, and technology and artificial intelligence. This discussion is not meant to be comprehensive or exhaustive, but exploratory and provocative.

3.1 Policy and Values: Pandemics and Public Health

The COVID-19 pandemic has elevated the importance of viruses and public health in the minds of researchers and the public. SPA advocates have voiced concerns about the TUA on public health outcomes. Schuster and Woods quote Slavoj Zizek (2011) saying,

> You want to raise taxes a little bit for the rich; they tell you that's impossible because we lose competitivity. You want more money for health care, they tell you, 'impossible, this means a totalitarian state; there's something wrong in the world, where you are promised to be immortal but cannot spend a little bit on healthcare.

Specifically, Zizek targets a faction of TUA donors who also fund longevity and human enhancement research; he sees the confluence of these projects as problematic to the public health interests of poorer and more vulnerable populations. Accordingly, is the TUA approach likely to crowd out spending on shorter-term focused health care outcomes, especially among the most vulnerable populations? The answer seems unlikely.

Well before the COVID-19 pandemic, TUA scholars Bostrom and Ord highlighted naturally occurring and engineered pandemics as important existential threats, with engineered pandemics being number two in Ord's overall rankings of existential risks (Ord, 2020). Complementing this research, Robert Reid (2019) has studied engineered pandemics and advocates for a variety of policy solutions that involve expanding access to health care worldwide. This follows because one of the best ways to limit the existential threat of a pandemic is early detection, and early detection is much easier with improved and universal health care access. Accordingly, wealthy countries and individuals have an interest in providing better health care access to the world's most vulnerable populations as part of a comprehensive existential risk mitigation plan.

Zizek's and Schuster and Wood's criticism also seems practically misguided from a policy perspective because many donors that fund TUA research are aligned with the political left and the Effective Altruism charity movement (Shleffer and Molla, 2020). Effective Altruism gives a large percentage of their donations to medical care for the least advantaged people currently alive (Matthews, 2020). Accordingly, there is theoretical rationale and empirical evidence to suggest that TUA thinkers are likely to endorse policies that lead to improved health care outcomes for the least advantaged in the present day, both because such efforts help mitigate existential risk and because such efforts have value independent of preventing existential risk.

More broadly, while the TUA advocates do endorse increased spending and research on existential threats with a longer-term focus, there is no evidence they believe these funds should come at the opportunity cost of programs that fund social programs for the least advantaged currently. Toby Ord suggests we should begin by spending as much on existential risk research and mitigation as we spend on ice cream, which in 2021 was $79 billion worldwide (Grand View, 2022; Ord, 2020). Health care spending in the U.S. alone in 2022 amounted to $4.3 trillion (CMS, 2023).

Accordingly, from a spending and access perspective, The TUA and SPA are likely to converge on advocating more accessible and comprehensive health care worldwide, regardless of how they define existential risk and regardless of whether they are motivated by short-term or long-term concerns or utilitarian or egalitarian motives. Again, while sacrificing current welfare to promote future welfare is theoretically acceptable according to the TUA, there is no evidence that this is practically the best

mechanism or heuristic for doing so. Relating this to climate change, Tyler DesRoches (Forthcoming) posits the idea of *Sustainability Without Sacrifice* and argues that many individual behavioral shifts that help combat climate change and solve other sustainability problems also lead to increases in individual welfare. Creutzig et al. (2022) also argue that climate change mitigation efforts, despite economic costs, are likely to have significant short and long-term positive well-being effects. Moreover, according to Nordhouse and Schellenberger (2007), bringing the poorest people in Brazil out of poverty might be the most practical solution to deterring Amazon rainforest destruction and mitigating climate change. From a policy perspective, the long-termism criticism from the SPA only holds significant weight when there is a negative causal relationship between levels of welfare of current lives and future lives. Accordingly, while the SPA is right that we should be vigilant of Pascal's mugging to justify policies that cause short term harm, from a policy perspective we need evidence to show that such concerns are relevant in specific cases.

However, despite ostensible agreement on increased access and spending on health care, the TUA and SPA might diverge with respect to policies related to pandemics and health care monitoring and vaccine requirements. Given that (Bostrom, 2019) advocates for increased surveillance in general to monitor existential risks, the TUA might be more likely to require mandatory population wide health care information sharing or vaccination. However, there is also reason to doubt this policy divergence given evidence from the recent COVID-19 pandemic. Many of those with egalitarian concerns were the biggest proponents of lockdowns and vaccination requirements during the COVID-19 pandemic (Schmeltz and Bowles, 2022). While a value pluralist would seemingly steer away from health care requirements and limitations on individual freedoms and privacy, it's unclear how SPA thinkers approach value pluralism in the face of existential threats where individual freedoms threaten to harm others. SPA thinkers criticize the TUA for not considering "existential threats" and "existentialism" more broadly, because they see it as ignorant or disrespectful to marginalized populations and cultures that face threats to their current existence; however, it is not clear that even if considered, such concerns would affect policies aimed to mitigate catastrophic and existential threats that affect nearly everyone (Shuster and Woods, 2021). Accordingly, it seems we need more clear guidelines from the TUA and SPA on how to weigh individual liberties and value pluralism against global harm reduction in the case of vaccine mandates, surveillance, and other risk mitigation strategies related to pandemics and synthetic biology.

Practically, given the current political landscape in Western democracies, it might be the case that because TUA and SPA thinkers are both comprised of academics and people who generally support Leftist candidates, that policy convergence between the groups is inevitable. While perhaps true in the current political environment, given the dynamic nature of politics, this political heuristic may be relatively fragile or short-lived. This might be especially true if events associated with global catastrophe tend to upend existing political structures. Moreover, given the potentially rising influence of independent institutions and individuals in existential risk research and mitigation, political policy convergence at the nation state level alone may not lead to convergence with respect to many existential risk mitigation efforts. For example, The Centre of the Study of Existential Risk and Future of Humanity Institute have received multi-million-dollar donations from billionaires Elon Musk and Peter Theil; such institutes and individuals in the future may bring about state like actions through decentralized autonomous organizations or DAOs (Tegmark, 2015; Hassan and de Fillippi, 2021). More generally, technologies such as blockchain and Web3 may limit some of the power nation states have traditionally held, clouding what counts as public policy in existential risk and other arenas (Srinivasan, 2023).

However, while the SPA critics have not proved that sacrificing short term welfare to mitigate existential risk is likely to obtain in the case of health care and pandemics, it also seems inevitable that in some cases it will. A trolly problem involving throwing a group of current lives off a bridge to save many more future lives is inevitable at some point. The theoretical framework for the TUA to handle such tradeoffs seems relatively clear given their utilitarian leanings, but it is not clear how the SPA deals with such an intergenerational trolly problem. If the SPA is to diverge with TUA policies based on a long-termism critique, we must know more about how the SPA handles disputes when the survival of certain values, cultures, and lives are mutually exclusive. We know that condoning a genocide to give the wealthy a chance to colonize space might be off the table, but what trade-offs would be acceptable?

3.2 Policy and Values: Non-Human Extinction and Anthropocentrism

The TUA has been accused of an anthropocentric agenda, and not considering other species' existence as central to risk mitigation (Schuster and Woods, 2021). Will this result in policies that greatly differ from the SPA with respect to the protection of non-human species and ecosystems? In theory, Bostrom's post-human ideals suggest that he does not see humans as an ideal species. His interest in becoming post-human is tied to a utilitarian framework about maximizing well-being for sentient creatures, which ostensibly includes humans and non-humans. He thinks humans should consider becoming post-humans because they may have higher levels of well-being and increased capacity to mitigate existential threats (Bostrom, 2008). Accordingly, in terms of policies encouraging human enhancements, the TUA and SPA seem likely to diverge. However, it is less clear how the TUA would consider the value of other sentient creatures *relative* to humans and post-humans in policy decisions. In his 2014 presentation on "Crucial Considerations and Wise Philanthropy," Bostrom says,

> To just pick an example: *insects*. If you are a classical utilitarian, this consideration arises within the more mundane—we're setting aside the cosmological commons and just thinking about here on Earth. If insects are sentient then maybe the amount of sentience in insects is very large because there are so very, very many of them. So that maybe the effect of our policies on insect well-being trump the effect of our policies on human well-being or animals in factories and stuff like that. I'm not saying it does, but it's a question that is non-obvious and that could have a big impact. (Bostrom, 2014)

Bostrom sees humans transcending beyond humanity as possible and preferable, but this does not entail that all sentient beings can or should transcend, or that humans or posthumans necessarily take ethical priority in the utilitarian tally. On the one hand, the SPA criticizes the TUA for being too anthropocentric, but then by attacking their ideal of transcending humanity, they accuse them of not being anthropocentric enough. Accordingly, with respect to policies that value the existence of certain species over others, it's not clear exactly where either the TUA or SPA stand.

It might be the case that through education, cultural evolution, or more artificial cognitive enhancement, the chasm between eco-centric and human centric concerns of extinction disappears. In her chapter "Looking into Extinction" from *Wild Dog Dreaming,* Rose (2011) says, "we humans emerged in dynamic relationships with animals and plants, with them we share our dependence on water and air . . . understanding how we fit into the community of life and death is not an optional extra." Rose blurs or eliminates the line between anthropocentric and non-anthropocentric thinking just like she blurs the line between the self and non-self; we might care about the extinction events of animals because they have intrinsic value, but also because relationships with these animals are

essential to humanity and human well-being. This analysis seems to mirror Bryan Norton's analysis on what causes convergence among environmentalists. Norton says, "The convergence hypothesis rests firmly on the central insight of ecology—that all things in nature are interrelated. If humans damage the larger context shared by both humans and other species there will, eventually, be negative impacts on all species." (Norton, 1994). It might be the case that having certain knowledge and adopting an attitude of care entails actions and polices that in the aggregate protect a diverse set of species in a diverse set of time periods.

However, at smaller scales, conflict between human and non-human interests will happen, and because extinctions are inevitable, it is not clear to what lengths humans might go to prevent other species extinction. As Norton discusses in his convergence hypothesis, a symbiotic relationship at a larger scale is consistent with inevitable conflict between species at smaller scales (Norton, 1994). Looking at issues surrounding endangered species and species protection, it is unclear how the TUA and SPA might differ in policy prescriptions.

Animal lab testing policy seems to be an area where the TUA and SPA could meaningfully diverge. Under the utilitarian framework, the suffering and death of animals, while important, could be trumped by maximizing long term utility for humans or other animals. This seemingly would involve a cost benefit analysis unique to each case of testing, and so it is difficult for a TUA utilitarian to universally apply "a test or do not test" maxim. The SPA seems more conducive to a "do no harm" principle regarding animal testing, based on their concerns for current human suffering and unwillingness to rationalize such suffering for future welfare gains (Rose, 2011; Schuster and Woods, 2021).

It does seem likely that both the TUA and SPA would converge to promote policies to eliminate many modern factory farming practices. Deborah Rose (2011) sees factory farming as a form of extinction, while the TUA would likely object on utilitarian welfare grounds. Moreover, eliminating many factory farming practices would likely reduce the risk of novel virus outbreaks (Greger, 2021).

3.3: Policy and Values: Artificial Intelligence and Technological Trajectory

Given technological progress is the main cause of increasing existential risk according to both TUA and SPA, policies surrounding technological development might represent the most important area of policy debate. Cremer and Kemp (2021) say,

> If the lion's share of extinction risk stems from emerging technologies, why do we rarely ask how to stop dangerous developments? This option is usually considered infeasible or outright impossible. This may be due to the exogenous threat model and technological determinism of the TUA. Since halting the technological juggernaut is considered impossible (by them), an approach of differential technological development is advocated.

"Differential technological development" is the idea that although the TUA might not see stopping technological progress in general as desirable or feasible, there is an opportunity to speed up or slow down certain specific technological developments. While it's not clear whether the SPA would disagree on policies to speed up or slow down a specific technology, the assumption that technological progress *must* occur to achieve human potential is hostile to the SPA. The idea of increasing the level of existential risk for everyone to give TUA advocates a better chance at achieving their utopian goals is undemocratic, insensitive, and imprudent to the SPA. This may spur heavy policy debate and divergence especially in the arenas of biological human enhancements and artificial

intelligence. Whereas the TUA will favor various safety mechanisms to slow these advances, the SPA may favor all out bans. Like debates on potentially harmful drugs and whether banning or regulating is likely to produce the best outcomes, debates on technologies are likely to be dynamic based on perceived and actual risks of any specific technology.

Moreover, technological determinism has both normative and descriptive components. While there might be good theoretical and normative reasons to slow or ban certain technological developments, there might be practical reasons why humans are unable to do so. With respect to policy prescription, the TUA and SPA need to decide to what extent realism and counterfactuals factor into their recommendations. If it is practically impossible to ban certain technologies, even if desirable to ban them, how would the SPA and TUA react? For example, one might not have originally supported the U.S. developing nuclear weapons, but then upon realizing that Japan or Germany was close to developing nuclear weapons, supported developing a nuclear arsenal to pursue a deterrence strategy. In today's world, some have argued that the U.S. cannot slow down AI research, even if it would make sense in isolation, because of the threat of AI from China. Such counterfactual analysis is uncertain and subjective, however, given the uncertainty and dynamic nature of technologies like AI, engaging in such counterfactual analysis will be helpful for both the TUA and SPA if their theories are to be comprehensive and effectively contribute towards policymaking.

Despite confusion and potential divergence on banning certain technologies, the shared concerns linking the TUA and SPA likely lead both to support enhanced regulatory bodies surrounding technology. Just as we have randomly controlled trials and approval processes for access to drugs and medical treatments, a more explicit approval process for new technologies should be espoused by both TUA and SPA advocates. Exactly which technologies will fall under this umbrella and how will such decisions be made? The recent introduction of GPT-4 makes answering this question all the timelier. It seems the SPA will advocate transparent and democratic decision making on such issues, while the TUA—given their consequentialist framework—may be less concerned with the intrinsic value of any specific decision-making process and be open to different governance structures (Holm, 2019).

4. Converging at the Policy Table

This paper has argued that, although TUA and SPA diverge in their theoretical orientations, this does not preclude the possibility that both may converge in the practical trenches of policy work. What might this reconciliation look like? Let us close by considering a few hypothetical examples, beginning with climate change.

For the TUA most climate change scenarios do not represent existential threats. The existential risk from climate change comes from extreme scenarios of runaway global warming which involve high degrees of uncertainty (Weitzman, 2009; Wagner and Weitzman, 2016). In a recent interview Toby Ord says, "unfortunately, we may have substantially more than six degrees of warming, and I've never seen reports on that. So, trying to understand these types of tail-risk possibilities, I think, is the most important way that we could help." (EA, 2020). SPA advocate Emile Torres criticizes the TUA focus on extreme scenarios as being callous, summarizing their position on climate change saying, "it'll probably cause tens of millions of deaths, hundreds of millions of people being displaced, maybe billions of people having to move. And it'll be a catastrophe, but really, when you zoom out and take the cosmic vantage point, that's just a little hiccup." (Torres, 2023)

Let us assume the SPA primarily values the billions of current people that may be displaced by even 1+ degrees of warming, whereas the TUA is most concerned with avoiding catastrophic scenarios of 6+ degrees warming that could, if not cause total human extinction, then at least cause the kind of civilizational collapse that would halt or reverse the technological developments required to generate trillions of intelligent lives in the far future. Despite these distinct values and priorities, both should support policies of funding research on understanding 6+ degree scenarios and interventions that aim to keep warming under 2 degrees (Kemp et al., 2022). While the SPA might think 1+ degrees is sufficiently bad to command our attention, 6+ degrees will be far worse, and understanding how to both avoid and deal with such a scenario can help minimize suffering for billions of people alive today. As Kemp et al. (2022) highlight, the IPCC has neglected rigorous study of extreme climate change scenarios, and a focus on such scenarios could further galvanize climate change mitigation efforts. Similarly, TUA advocates should join the SPA and support interventions that aim to keep warming under 2 degrees in part to reduce the risk of a global civilizational collapse that forecloses the possibility of reaching techno utopia. Even if TUA advocates are only worried about 6+ degree scenarios, growing evidence for climate 'tipping points' and positive feedback cycles suggests that keeping warming to below 2 degrees may in fact be required to prevent runaway warming (Lenton et al., 2019). Moreover, both schools of thought should recognize the threat of cascading risks and how lower levels of warming could increase the likelihood of threats like nuclear war and bioterrorism (Undheim, Forthcoming).

Accordingly, regardless of what makes global warming problematic, how problematic it is, and how it compares to other existential threats, when sitting down at the policy table, both TUA and SPA advocates should support novel climate change research and interventions aimed to keep warming under 2 degrees. While the best types of intervention will be debated (carbon tax, cap and trade, and mandates etc.), this should not detract from policy convergence that more interventions are necessary. The case of climate change shows that the suffering of billions of current people is not a price we should be willing to pay to preserve trillions of future happy lives, but a price we should not be willing to pay because it puts the likelihood of future happy lives in jeopardy.

Also promising is the TUA and SPA converging on policy related to synthetic biology. In their essay, "Existential Risk and Cost-Effective Biosecurity," Piers Millet and Andrew Snyder-Beattie (2017) analyze the cost effectiveness of mitigating low probability and high impact existential threats from the proliferation of synthetic biology, including pandemics, biowarfare, and bioterrorism. They analyze the cost effectiveness of mitigating extreme and existential risks in comparison to traditional health care spending, first considering current lives and possible lives in the far future (TUA) relatively equally, and then by applying higher discount rates that effectively negate the value of lives in the far future (SPA). They say, "most policymakers will be responsible primarily for the interests of a more limited constituency compromising only the current generation and near future. It is therefore instructive to evaluate how well the cost-effectiveness results hold up when we largely ignore the benefits to future generations." Even when heavily discounting the value of future lives, their results show that focusing on extreme negative events from synthetic biology is more cost-effective than traditional health care spending in scenarios where the chance of an existential threat is anything above .0001 events per year. As far as how the extreme risk mitigation money is spent, Millet and Snyder-Beattie follow World Bank recommendations to raise global health infrastructure spending by $2.5 billion per year to bring all countries up to minimal international health standards, noting that in 2014 only 33% of countries reported meeting these standards (Millet and Beattie, 2017).

For the TUA advocate, the analysis and policy implications importantly show that existential risk focused spending is cost effective compared to traditional health care spending, especially when looking at the cost per live saved over very long time periods. For the SPA advocate, the policy prescription focuses spending on health infrastructure in the world poorest nations, leveling the playing field between developed and developing countries alongside a cost-effective intervention that promotes saving lives for those living today. Overall, in the case of synthetic biology, the kind of basic public health care provisioning that alleviates immediate suffering as endorsed by the SPA advocate also increases monitoring capabilities for novel pathogens and improves population resilience against novel disease by increasing overall health.

At the broader level, it is important to note that in the policy recommendations just discussed on climate change and synthetic biology, there is no tension between shorter and longer-term thinking (assuming by shorter term we mean the next 50 years), but instead the approaches are complimentary. In the case of global warming, thinking about extreme warming scenarios and the near-term civilizational bottlenecks required to achieve the infinite long-term potential of humanity gives us more reason to act towards mitigating climate change now. Moreover, thinking about the importance of billions of current lives helps us avoid neglecting the potential for threats that would be 'merely' catastrophic when considered in isolation, but when combined may cascade humans into extinction. Similarly with threats from synthetic biology, thinking about the far future provides exponentially stronger reasons to act towards improving the health of current generations than thinking about current generations in isolation. Acting towards helping the global poor today gives us a better chance at having trillions of happy lives in the far future.

5. Objections

First, while the TUA and SPA might have reason to converge on policies that marginally promote their distinct values, this could mean abandoning policies that *optimally* promote their values. For example, while supporting strict national mandates on emissions raise the likelihood for a techno utopia in the distant future, it might be the case that a global carbon tax does so more effectively. Such conflict and compromise on policy details is inevitable and might be seen as a feature rather than a bug. Discussing and researching the nuances of policy can expose different schools of thought to novel information and promote rigorous research and analysis that benefits all. That neither the TUA nor the SPA leave the policy table satisfied is not evidence that such reconciliation is not useful.

Second, policies that promote shared values might also promote conflicting values. For example, supporting a carbon tax that effectively mitigates climate change might be seen as an unacceptable endorsement of capitalism by a socialist leaning member of the SPA. Such conflict also seems inevitable, but in the crafting the nuances of policy might be mitigated. For example, the revenue from a carbon tax might be used to mitigate welfare inequalities. Moreover, given the uncertainty and high stakes involved in catastrophic and existential risks, the schools of thought should be encouraged to weigh and prioritize their values (to what extent should a socialist SPA member be willing to compromise on market-based policies that mitigate existential threats?); embracing counterfactuals and deciding when to compromise certain values will be part of both good theory and policy.

Third, it might be the case that using the TUA and SPA nomenclature is not particularly useful and encourages a type of political polarization that harms the debate; scholars attributed to these schools have unique attitudes on specific issues, and many people are unaware of these schools of thought and will have no interest in aligning with a particular group. The practical and theoretical nature of this debate means that engagement from

academics, policymakers, and the public is all important. Moreover, given the importance of coordination and cooperation in existential risk mitigation, encouraging open minded and inclusive attitudes is also important.

The answer to this objection is that the use of these labels is meant to be provocative and descriptive, rather than decisive and normative. Going forward, getting rid of such labels and being open to new schools of thought should be considered.

6. Conclusion

Thinking about existential risk brings questions of what we should value and how we should prioritize values to the fore, and TUA and SPA scholars offer unique ways of thinking about and answering these questions. It would be a mistake to assume that existential risk prevention is a purely technocratic, and apolitical, value neutral enterprise. However, alongside theoretical discussions it is important to analyze the role policy plays in promoting these values, and how policies promoting one set of values might compete with or complement a different set of values. Accordingly, as Norton demonstrates from the environmentalism movement, it would also be a mistake to overlook the way in which striving to develop technical policy solutions to existential risk can practically reconcile theoretically opposed value systems in some cases. While philosophical "would you rather games" can be entertaining and useful, moving many of these discussions to practical policy debates will be important in effectively mitigating existential risks going forward.

References

Bostrom, N. (2002). Existential risks. *Journal of Evolution and technology*, *9*(1), 1-31.

Bostrom, N. (2008). Why I want to be a posthuman when I grow up. In *Medical enhancement and posthumanity* (pp. 107-136). Springer, Dordrecht.

Bostrom, N. (2014) Crucial considerations and wise philanthropy. http://www.stafforini.com/blog/bostrom/

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, *10*(4), 455-476.

CMS. (2023). https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nationalhealthaccountshistorical

Cremer, C. Z., & Kemp, L. (2021). Democratising risk: in search of a methodology to study existential risk. *arXiv preprint arXiv:2201.11214*.

Creutzig, F., Niamir, L., Bai, X., Callaghan, M., Cullen, J., Díaz-José, J., ... & Ürge-Vorsatz, D. (2022). Demand-side solutions to climate change mitigation consistent with high levels of well-being. *Nature Climate Change*, *12*(1), 36-46.

DesRoches, Tyler. (Forthcoming). Sustainability without sacrifice.

Grandview Research. (2022). Ice cream market size, share and trend analysis. https://www.grandviewresearch.com/industry-analysis/ice-creammarket#:~:text=The%20global%20ice%20cream%20market,4.2%25%20from%202022%20to%202030.

Greger, M. (2021). Primary pandemic prevention. *American Journal of Lifestyle Medicine*, *15*(5), 498-505.

Hassan, S., & De Filippi, P. (2021). Decentralized autonomous organization. *Internet Policy Review*, *10*(2), 1-10.

Holm, S. (2019). Precaution, threshold risk and public deliberation. *Bioethics*, *33*(2), 254-260.

Hull, A. D., Liew, J. K. S., Palaoro, K. T., Grzegorzewski, M., Klipstein, M., Breuer, P., & Spencer, M. (2022). Why the United States must win the artificial intelligence (AI) Race. *The Cyber Defense Review*, *7*(4), 143-158.

Lenton, T. M., Rockström, J., Gaffney, O., Rahmstorf, S., Richardson, K., Steffen, W., & Schellnhuber, H. J. (2019). Climate tipping points—too risky to bet against. *Nature*, *575*(7784), 592-595.

MacAskill, W. (2022). *What we owe the future*. Basic books.

MacAskill, W., Mogensen, A., & Ord, T. (2018). Giving isn't demanding. *The Ethics of Giving: Philosophers' Perspectives on Philanthropy*, 178-203.

Millett, P., & Snyder-Beattie, A. (2017). Existential risk and cost-effective biosecurity. *Health security*, *15*(4), 373-383.

Nordhaus, T., & Shellenberger, M. (2007). *Break through: From the death of environmentalism to the politics of possibility*. Houghton Mifflin Harcourt.

Norton, B. G. (1994). *Toward unity among environmentalists*. Oxford University Press

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hachette Books.

Ord, Toby. (2020). Fireside chat with effective altruism. Effective Altruism. https://www.effectivealtruism.org/articles/fireside-chat-q-and-a-with-toby-ord

Reid,R. (2019). How synthetic biology could wipe out humanity. TED. https://www.ted.com/talks/rob_reid_how_synthetic_biology_could_wipe_out_humanity_and_how_we_can_stop_it?language=en

Rose, D. B. (2011). *Wild dog dreaming: Love and extinction*. University of Virginia Press.

Schmelz, K., & Bowles, S. (2022). Opposition to voluntary and mandated COVID-19 vaccination as a dynamic process: Evidence and policy implications of changing beliefs. *Proceedings of the National Academy of Sciences*, *119*(13), e2118721119.

Schuster, J., & Woods, D. (2021). *Calamity theory: Three critiques of existential risk*. U of Minnesota Press.

Shleffer,T and Molla, R. (2020). Silicon Valley is spending millions more for Joe Biden than it did Hillary Clinton. Vox. https://www.vox.com/recode/2020/10/30/21540616/silicon-valley-fundraising-donald-trump-joe-biden-analysis

Simchon, A., Turkin, C., Svoray, T., Kloog, I., Dorman, M., & Gilead, M. (2021). Beyond doubt in a dangerous world: The effect of existential threats on the certitude of societal discourse. *Journal of Experimental Social Psychology*, *97*, 104221.

Srinavasan, B. (2023). *The Network State*.

Tegmark, M. (2015). Elon Musk donates 10m to keep AI beneficial. Future of Life. https://futureoflife.org/2015/10/12/elon-musk-donates-10m-to-keep-ai-beneficial/

Torres, P. Emile. (2021). The dangerous ideas of long-termism. Current Affairs. https://www.currentaffairs.org/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk

Torres, P. Emile. (2023). Longtermism: an odd and peculiar ideology. Netzpolitik. https://netzpolitik.org/2023/longtermism-an-odd-and-peculiar-ideology/

Weitzman, M. L. (2009). The extreme uncertainty of extreme climate change: An overview and some implications. *Harvard University, Boston*.

Wagner, G., & Weitzman, M. L. (2016). Climate shock. In *Climate Shock*. Princeton University Press.

Zizek,S. (2011). Don't fall in love with yourselves. *Occupy!: scenes from occupied America*. Verso Books.

Stanford | Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Governing and Anticipating Anthropogenic Existential Risks: Envisioning Some New Approaches

Mariana Todorova [1*]

Citation: Todorova, Mariana. Governing and Anticipating Anthropogenic Existential Risks: Envisioning Some New Approaches. *Proceedings of the Stanford Existential Risks Conference* 2023, 220-232. https://doi.org/10.25740/vp972pt5387

Academic Editor: Trond Undheim, Dan Zimmer

Funding: N/A

Conflict of Interest Statement: N/A

Informed Consent Statement: N/A

Acknowledgments: N/A

Author Contributions: N/A

Abstract: In my role as a futurist, I delve into the paradigms that pertain to our reality characterized by escalating complexity, the acceleration of social time, and the proliferation and cascading of existential risks. By tracing the origins of the "risk society" concept and elucidating the fundamental attributes of existential risks and their multifaceted nature, I emphasize their inherent distinctions. The contemporary societal landscape, characterized by the "thickening" and compression of the present, reaches a threshold where reflexivity becomes arduous. This necessitates novel approaches and methodologies to address existential risks. The study primarily focuses on anthropogenic existential risks, acknowledging the interplay of subjective and objective factors contributing to their emergence. Subjectively, these risks are influenced by human bounded rationality, knowledge gaps leading to irrationality (cognitive deficits, heuristics, and biases), the employment or neglect of critical and deliberate thinking, and the challenge of comprehending complex systems amidst our cognitive limitations. The article also introduces a novel forecasting methodology that I have developed utilizing the resource of counterfactual analysis. Counterfactuals effectively capture the intricate and fluid essence of our contemporary reality, as elucidated through some illustrative case studies.

Keywords: risk society, anthropogenic existential risks, subjective factors, counterfactuals

[1] Associate Professor, Bulgarian Academy of Sciences; Sofia, Bulgaria

[*] Correspondence: marianafuturestudies@gmail.com

1. Introduction

Risk society places significant emphasis on the recognition and comprehension of risks across diverse spheres of contemporary society, encompassing environmental, technological, and societal domains. It underscores the imperative of proactive risk governance, well-informed decision-making, and collective accountability. By engaging in meticulous analysis and evaluation of risks within the framework of risk society, societies can devise strategies and policies to alleviate potential detriments and bolster their resilience. Conversely, existential risks denote perils capable of imperiling humanity's very existence or inflicting irreparable global harm. These risks typically transcend the confines of individual societies and necessitate global cooperation and coordination for effective resolution. While the concept of risk society can contribute to risk comprehension and the formulation of risk management strategies, addressing existential risks usually demands a broader perspective and interdisciplinary approaches. Such an undertaking entails deliberation not only of immediate and tangible risks but also of low-probability, high-consequence events with far-reaching ramifications.

Studies of risk in various social spheres outline a diverse range of approaches, concepts, and criticisms, which, however, can be organized most generally into four main paradigms (Mythen, 2004). The first is inspired by the work of M. Douglas (Douglas, 1985) and can be called anthropological, focusing on understandings and perceptions of risk among different types of groups. The second paradigm emerged within social psychology and consists of developing psychometric methods to assess which risks are perceived as more threatening and harmful. The focus is on individual perceptions of risk. This is close to Nasim Taleb's notion of concept of "skin in the game". The third broad area regarding risks is governmentality studies. The authors working in this direction develop and apply a series of concepts first introduced by Michel Foucault in his work on the disciplinary effects of various discourses. Risk is understood as one of the heterogeneous governmental strategies of disciplinary ruling and controlled power. The conclusion of this rationalized discourse is that risk could be controlled as long as expert knowledge can be properly utilized (Foucault, 1977). The fourth paradigm is presented by Ulrich Beck (Beck, 1992) and Anthony Giddens (Giddens, 1992), who focus on a set of risks produced by human activity. The destructive consequences of these "manufactured risks" threaten the entire planet and cause fundamental changes in social structure, politics, and cultural experience.

Mitigating existential risks may entail interdisciplinary research, collaboration among scientific, technological, and policy communities, and the establishment of global governance mechanisms. It mandates comprehension of intricate interdependencies and feedback loops among diverse systems, including technological advancements, environmental transformations, and socio-economic factors. Scot Lash (Lash, 2000) suggests replacing the notion of 'risk society' with an idea of 'risk culture'. A robust risk culture that integrates the principles of risk society can significantly contribute to the management of existential risks by cultivating risk awareness, fostering transparent communication, enabling proactive risk governance, promoting collaborative endeavors, integrating ethical considerations, and cultivating a culture of continual learning and adaptation. By embracing these principles, individuals, organizations, and societies can augment their capacity to identify, comprehend, and mitigate existential risks, thereby safeguarding the well-being and future of humanity.

In the article, I aim to delineate several subjective human factors that can serve as significant components of existential risks, amplifying and deepening their consequences. I believe that by employing an approach that dissects the complexity of these phenomena into conditional and abstract "building blocks," we could potentially enhance our understanding and management of these risks.

2. Anthropogenic Subjective Existential Risk Factors

2.1 Longstanding Unresolved Problems

The multifaceted nature of existential risks arises from deep-rooted and unresolved problems that have accumulated over time, giving rise to novel and unfamiliar phenomena. The challenge lies in the instability and inadequacy of existing frameworks designed to address these issues in different context in past times. A prominent example is the reemergence of nuclear weapons as one of the most threatening existential risks, as the once-established conventions of mutual military restraint face new complexities, further exacerbated by the potential influence of artificial intelligence (AI) in accessing and controlling such weaponry.

2.2 Cognitive Deficit

One key factor that can shape, amplify, or intensify existing existential risks is what I refer to as the "pandemic of cognitive deficit." This phenomenon represents a novel dimension of cognitive decline, not in the medical sense but more in terms of epigenetics. It encompasses the way information is sought, consumed hastily and without careful consideration, and the limited capacity for critical reflection. The limiting of the brain to only dopamine pleasure (dopamine is a molecule in the brain and body that is closely linked to our sense of motivation**)**, the emergence of ChatGPT 3.5 and GPT-4, which provide directions, synthesize information, replace multiple work tasks and activities that can transform slow thinking (according to Kahneman's definition and metacognition into heuristics (shortcuts and prescriptions), actually lead to cognitive decline. The enormous amount of unchecked information, fake news, and conspiracy theories, “echo chambers” in which opposing and distrustful groups of people self-isolate and mutually believe in each other’s views, also contribute to these processes.

2.3 Grey Areas Between Lack of Knowledge and Expertise

Novel and unexplained ambiguous realms are emerging, occupying the space betwixt knowledge scarcity and expertise. The process of automation and robotization imparts a dehumanized character to various operations, thereby engendering hitherto unforeseen phenomena such as AI-assisted social engineering. It is already observed that military personnel in control and command centres (C2) or judges feel hesitant to challenge an AI decision. A similar phenomenon is people giving human attributes to AI or personalizing the information from it. This will greatly affect decision-making in the future.

Another example that may enhance collective irresponsibility and thus transforms the nature of existential risks is the loss of expert knowledge. We can illustrate such a process with mega-acquisitions and mergers of companies. Here are some cybersecurity breaches: If a large tech company acquires a smaller company with specialized expertise in cybersecurity, and then eliminates that expertise because of optimization, it could increase

the vulnerabilities to cyber-attacks, which could lead to massive data breaches or even a collapse of huge systems.

2.4 Specific and Non-Specific Lack of Knowledge

Lack of knowledge sometimes is also a component in decision making. Despite constant efforts to find measures for existential risk, it always builds upon a multi-component basis in which the key one is the construction of meaning on the unknown. In this case, the distinction between specific and non-specific lack of knowledge is of paramount importance. The specific lack of knowledge is in a particular area and becomes a motive for research and production of a new knowledge. For example, to explore the factors for sun flairs and to think what would be possible to be done. They can interfere with the Earth's magnetic field, which can induce electric currents in power lines and cause blackouts or damage to electrical infrastructure. Generally, it can also be part of the bounded rationality, irrationality (heuristic and biases and the most popular of it – the Dunning-Kruger effect). On the other hand, non-specific knowledge deficits manifest as a state of profound disinterest on the part of social actors or the presence of unknown unknowns phenomena. The question arises: How can we harness social support to overcome a specific problem if individuals simply do not exhibit concern or investment?

2.5 Polarization

Polarization, as a salient political phenomenon, experiences amplification through the advent of new technologies, exacerbated inequality, eroded trust, and the erosion of truth. Its repercussions extend directly to existential risks. We find ourselves in increasingly fragmented societies, impeding effective political governance due to this disintegration. Consequently, it becomes exceedingly challenging for political representatives to legitimately advocate for their constituents to other societies and states. Political processes and decisions are confronted with instability, perpetually in a state of flux. To illustrate, while the European Union has prohibited the fusion of autonomous and semi-automatic weaponry with AI, such restrictions are absent in the United States. Conversely, China has implemented a facial recognition-based social credit system, an endeavor deemed wholly unacceptable within democratic nations. Thus, there exists a genuine risk of fragmenting the world into disparate jurisdictions concerning AI, a development that inherently harbors the potential emergence of existential risks.

Zygmunt Bauman accepts that the future is radically undetermined and uncertain. For him, it is precisely the postmodern situation that is both undetermined and un-determining: the influence of the past is weakened, and the faith and striving to colonize the future are dying out (Bauman, 1992). In a way, this trend is good because we already suffer from the ideologization and colonization of the future, which leave no room for anything else and, for example, extend the polarization of politics into the future.

2.6 De-Ideologization of the Future

The ideologization and subsequent colonization of the future often leave little room for alternative perspectives. This colonization permeates various aspects of society and extends undesirable practices and political polarization into the future. Here, I refer to strategically crafted political documents that may be contradictory or possess morally hazardous implications. What measures can be taken to rectify this situation? Decolonization, in this context, involves a critical examination of the methodologies

employed in shaping our visions of the future, challenging the definitions we have constructed, including those concerning time and "the future" itself. Perhaps we must embark on the task of rebuilding and constructing new narratives for the future that are non-catastrophic, non-alarmist, and contribute to a mindset that does not perceive annihilation as an inevitable outcome.

In the contemporary era, there is a pressing need to embark upon a comprehensive process of de-ideologization in order to mitigate the influence of rigid ideologies and attain a more objective future. The de-ideologization of our collective destiny necessitates a concerted endeavor encompassing individuals, communities, and governments, wherein primacy is accorded to factual information, empirical evidence, and rationality, transcending the sway of political convictions and ideological predispositions. To this end, a range of measures can be undertaken. De-ideologizing the future in terms of certain trends that have the potential to become self-fulfilling prophecies can lead to new approaches to dealing with existential risks. An example of this could be overcoming the polarization of the debate about climate change.

The de-ideologization of the future to some extent also involves depolarization of the present. Depolarization involves preventing cognitive decline, limiting fake news, conspiracy theories, echo chambers, overcoming widening social inequality, and perhaps combating the fragmentation of political ideas and the breakdown-anomie of society, as well as returning to big ideological transformational ideas. In this sense, de-ideologizing the future and its "wrong" colonization passes through several trends: on the one hand, "consolidating" democracy through big transformational ideas and narratives that are ideologies directed towards the future, and on the other hand, "deconstructing" (in Derrida, 2000) contemporary complexity of existential risks into smaller understandable and manageable processes. The appropriate approach in this case, if there is a lack of expertise and knowledge, is to make only temporary measures.

3. Specificity of Existential Risks

In the works of various authors (Yudkowsky, Rees, Bostrom, Ord, Tegmark), we can find different definitions and categorizations of existential risks. They threaten the very existence of humanity or other advanced forms of life. Here are some of the most commonly identified existential risks:

1. Global catastrophic events: These include natural disasters such as supervolcanic eruptions, asteroid impacts, or solar flares, as well as human-caused disasters such as nuclear war, bioterrorism, or artificial intelligence (AI) gone rogue.
2. Climate change: The gradual increase in global temperatures caused by human activity could lead to widespread ecological collapse and societal disruption, with potentially catastrophic consequences for humanity.
3. Pandemics: A highly contagious and lethal pandemic could lead to mass mortality and societal megacrisis, especially if there is no effective medical treatment or vaccine.
4. AI safety: As AI systems become more advanced and powerful, there is a risk that they could become uncontrollable or turn against their human creators, posing a threat to humanity's existence.
5. Global systemic collapse: The interconnectedness of global systems such as finance, energy, and food production means that a collapse in one area could trigger a cascade of failures, leading to a global systemic collapse.

6. Unforeseen events: There is always the possibility of unexpected and unprecedented events that could threaten the existence of humanity, such as a Black swan event that could destabilize society or cause a widespread catastrophe.

It's worth noting that different experts and organizations may prioritize these existential risks differently, and there may be other risks that are not included in this list. However, these are generally considered to be among the most pressing existential risks facing humanity today. Those popular definitions help to describe all possible, probable, and plausible complexities and cascading events so that proactive action strategies can be developed. Clearly defined protocols for the first five factors could also help as an algorithm for reacting to "unknown unknowns".

3.1 Narratives as constructs of existential risks

An illustration of how narratives that can influence the present and future and also have the potential to become an existential risk. As an example, we can give the comparison of the state of the Western world now with the process of decline of the Roman Empire.
As Immanuel Wallerstein points out, we are currently in the third stage of development of any system - its end. During this stage, it moves further and further away from equilibrium, and fluctuations become stronger. We cannot even say whether we are at the beginning of a fundamental change because it has not yet happened, and in a similar phase of system development - the bifurcation phase - prediction is a real challenge (Wallerstein, 2000).

The American writer and journalist, Colin Murphy, wrote a commentary in The Atlantic in 2021, which revises some of the ideas in his book: "Are We Rome?: The Fall of an Empire and the Fate of America". As he explains, the comparisons that come to mind now are not only about realities on the ground but about unrealities in our heads. The debasement of truth, the cruelty and moral squalor of many leaders, the corruption of basic institutions—signs of rot were proliferating well before January 6, and they remain, though the horde has been repelled.

The decay of the Roman Empire is a complex historical phenomenon that is attributed to a combination of political, economic, military, and social factors. Therefore, many analysts see a coincidence of factors and look for similarities between then and now. Political instability that the Roman Empire experienced with a series of weak and ineffective emperors, as well as a period of civil war that weakened the central government and undermined its ability to govern effectively. The Roman Empire suffered from a long period of economic stagnation, marked by high inflation, a declining tax base, and a shrinking population. This made it difficult for the government to finance its military campaigns and maintain its vast infrastructure. Military weakness was due to that the Roman Empire was unable to defend its borders effectively against invading barbarian tribes, leading to a series of military defeats that weakened the Empire's power and prestige. Social decay was also quite visible because the Roman Empire experienced a decline in social cohesion, with increasing social and economic inequality, widespread corruption, and declining moral values. This weakened the Empire's ability to maintain social order and stability.

It would be difficult to predict some kind of "end" to our Western civilization, or at least not a classical one, but rather a transformational one - of eclecticism and mutual intertwining and coexistence of multiple (contradictory) trends and tendencies. Therefore, existential risks today are contradictory and generate multiple and different uncertainties.

But this can lead to certain strategies of action, which can add additional elements to the existential threats. For example, Saudi Arabia and other countries consider accepting yuan instead of dollars for Chinese oil sales. This has the potential to upset the energy balance and geopolitical influence.

3.2 Existential Risks: Supra-Personal, Supra-Institutional and Supra-National

How can we turn the subjective stake as motivation and guarantee for dealing with existential risks? Nassim Nicholas Taleb's book "Skin in the Game: Hidden Asymmetries in Daily Life" argues that people who have "skin in the game" - that is, who have something to lose - are more likely to make sound decisions and take responsibility for the outcomes of their actions. (Taleb, 2018). Taleb uses the concept of "skin in the game" to discuss the idea of risk and how it relates to decision-making. He argues that individuals or organizations that are insulated from the risks of their decisions (i.e., those who do not have "skin in the game") are more likely to make reckless or irresponsible decisions, as they do not bear the full consequences of those decisions. There are two important aspects here. The first is that it is almost impossible to translate the abstractness and scale of existential risks into "skin in the game" for each individual. Even if they are fully aware of the totality of the threat, the abstractness remains regarding the belief in the possibility of influence from a personal position, and the situation happening "somewhere in the future" deprives the feeling of "skin in the game". Catastrophic scenarios are often used in future studies to raise awareness of these issues. A popular approach is storytelling, which can sometimes harm the historical narrative or other types of ideological discourses, because sometimes it may erose the chronology in history. Nassim Nicholas Taleb has also written extensively about uncertainty in his various books, including “The Black Swan” and “Antifragile”. In “The Black Swan”, Taleb introduces the concept for events, which are unpredictable, high-impact events that have major consequences but are often rationalized or explained away after the fact. He suggests that traditional approaches to risk management, which rely on statistical models and historical data, are inadequate in dealing with these types of events, and that we need to embrace a more antifragile" approach that allows us to adapt to and even benefit from uncertainty. The “unknown unknowns” phenomenon requires an attempt to rationalize the irrational and the unknown, or the gray areas between them. In addition, humanity needs us to think in nuances, not polarities. Realizing existential risks as personal, trying to anticipate Black swan contingencies, and building antifragilities are important components of mitigating existential risks.

4. Counterfactuals

As a futurist, I have proposed a novel forecasting method, which integrates counterfactuals with scenario building (Todorova, 2015; Todorova & Gordon, 2017, 2019; Illieva & Todorova, 2023). This methodology aims to explore potential elements of existential risks. To begin, I will delineate the distinction between counterfactuals and counterfacts. Counterfactuals are conditional statements, often employed in psychology, historiography, and everyday thinking, to mentally trace alternative paths of development (Illieva & Todorova, 2023). These counterfactual conditionals retrospectively analyze the mechanisms by which events unfold, aiming to assess the extent to which a particular event (counter-fact) could have influenced the course and direction of development. They align with Fogel's analysis of railways and familiar speculations like "What if Napoleon was not born..." (Illieva & Todorova, 2023). On the

other hand, counterfacts function as non-facts but are treated as facts with the potential to alter the course of events (Illieva & Todorova, 2023).

By employing this methodology, I aim to apply counterfactual analysis to anticipate and understand various facets of existential risks (Todorova, 2015; Todorova & Gordon, 2017, 2019; Illieva & Todorova, 2023). I have classified several types of counterfacts (Todorova, 2015). The first one is what I termed **dormant facts**. Dormant facts denote existing situations or realities whose potential impact has remained unrecognized or unfulfilled. These latent factors have lain dormant in the past, only to be activated and manifest their consequences if and when there are substantial changes in the social, economic, political, or religious context. Such transformations often give rise to dramatic and frequently violent shifts. A prime example of a dormant fact is the concept of frozen conflicts, as seen in the ongoing war in Ukraine, which is the result of a dormant conflict (Todorova, 2015). Similarly, other dormant facts such as unresolved frozen conflicts or gaps in knowledge (such as the weakening of the Earth's magnetic shield) can serve as elements contributing to existential risks. By categorizing and analyzing these dormant facts, we can endeavor to navigate the realm of "unknown unknowns."

Counterfacts may also be subject to reinterpretation or even reinvention and assume new meaning or content when political, social, religious, or economic developments take a significant turn. The act of reinventing a fact is frequently employed for propaganda purposes and ideological manipulation. Instances of employing counterfacts for these purposes can be witnessed in endeavors to rewrite history or redefine identity, often with the intention of mobilizing public support. Anticipating the reinvention of history and its subsequent reinterpretation can potentially generate military tensions and even instigate new conflicts. This holds particularly true for regions of heightened tension, such as former Soviet republics, the China-Taiwan-Japan and US dynamics, and the Middle East. These instances also represent critical elements contributing to existential risks.

A third category of counterfacts consists of rumors and hypotheses, which, though not yet universally accepted as facts, nevertheless can have immediate and potentially lasting impacts on reality equally to or even stronger than established facts. We can provide a myriad of examples here from divine theory, through Brexit, to bank bankruptcies. Even political elites sometimes rely on rumors, fake news, or conspiracy theories in promoting their goals. The volatility in the value of cryptocurrencies (which can be influenced by public statements by Elon Musk), the bankruptcy of many crypto exchanges, and recently, of Silicon Valley Bank, demonstrate how strong this factor is. In addition to that we can hereby consider some subtypes such as:

Fake news, operating under the guise of truth, presents itself as factual information with the primary purpose of influencing reality in two distinct ways: either by promoting and imposing specific circumstances or by challenging existing facts. In essence, fake news can be seen as a contemporary manifestation of propaganda. While it falls within the category of counterfacts, it is essential to note a significant distinction: unlike rumors and hypotheses, which can be exposed as false, fake news proclaims its authenticity. The structure of fake news is carefully crafted, incorporating elements of truth to lend credibility while simultaneously introducing distortions that manipulate desired perceptions and interpretations. The rise of artificial intelligence has introduced new dimensions to this phenomenon, with the emergence of deep fake technology and the proliferation of programs such as Dall-E and Midjourney, which enable sophisticated imitations. This technological advancement further facilitates the dissemination of "false

flag" manipulations, potential acts of sabotage, and the spread of deep fakes, among other deceptive tactics.

Intentional and unintentional self-fulfilling and self-denying prophecies, also referred to as normative forecasts, encompass a range of constructed counterfacts that serve distinct purposes. Intentional self-fulfilling and self-denying prophecies, characterized as purposeful interventions, are strategically shaped to mobilize support or provoke rejection. These counterfacts are deliberately designed to emulate the driving forces behind a particular process, often taking the form of influential trends or trend-setters. An illustrative example of an unintentional self-fulfilling prophecy is the placebo effect observed in the administration of mock medicines. On the other hand, unintentional self-denying prophecies operate through a fascinating mechanism. In certain cases, goals and policies are publicly announced, planned, and programmatically pursued with the aim of realization. However, over extended periods and years, these envisioned goals and policies may go unrealized, transforming into unintentional self-denying forecasts. A notable and frequently cited example is the European Union's Lisbon Strategy, which sought to position the Union as the world's most dynamic and competitive economy by 2010. Similarly, within the current decade, the "Green Deal" initiative, in light of global energy, economic, and food crises triggered by the war in Ukraine, could potentially become an unintentional self-denying prophecy as it may necessitate self-correction aligned with the prevailing circumstances. Such an outcome could pose significant existential threats and risks to the entire planet.

In contrast to **unintentional self-denying prophecies, intentional self-denying prophecies** can be strategically employed as political instruments to redirect social energy and public attention in desired directions. Frequently, politicians address people's fears by presenting dreadful scenarios as probable events, aiming to galvanize collective efforts in averting the predicted outcomes. Conversely, intentional self-denying prophecies can serve as counterbalances to catastrophic scenarios that have become increasingly ineffective, causing societies to stagnate in a state of horror and inaction.

The distinction between counterfactuals and counter-facts lies in their inherent value. Counterfactuals primarily focus on the process and delve into the exploration of alternative dynamics, challenging established norms and revealing potential mechanisms for social change. On the other hand, counter-facts place greater emphasis on events—non-factual occurrences that have the potential to redirect the course of processes and offer insights into potential points of intervention for shaping the future. The combination of these two elements proves highly valuable, as it enables a reflexive counterfactual analysis that challenges and reevaluates existing cognitive and epistemic frameworks.
By incorporating counterfactuals and counterfacts, anticipatory activities can be infused with reflexivity. This enriches futures research by expanding the analytical space and shedding light on alternative mechanisms of social change, moving beyond a predominant focus on trend projections (extrapolation). Counterfactual reflexivity holds the potential to offer a vast range of reflective feedback that can inform future political actions. Furthermore, through the exploration and evaluation of various manifestations of counter-facts within the realm of counterfactualities, an entirely new vision of the world can be constructed—a world brimming with untapped potentials and unfulfilled possibilities. It represents a realm of potentiality on par with the reality we currently inhabit. Through the identification of different types of counterfacts, we can uncover seeds of the future, potential discontinuities, Black swans, or prospective anchor points that

have the capacity to unravel the complex nature of existential risks and facilitate their deconstruction.

5. Counterfactual and Scenario Methods

When we combine counterfactuals with the scenario-based forecasting method, we have the potential to establish a solid foundation for delineating the boundaries of possible existential risks and strategies for mitigating them. Finding a probable cause for a wide range of events in the future is a cognitive function of a combination of looking back at the past and expectations about the future. Kahneman and Miller (Kahneman, Miller, 1986, p.136) summarize this thesis in their statement: "thought processes flow not only forward, arising from expectations and hypotheses for confirmation or revision, but also backward to our previous experience."

The search for counterfactual alternatives is provoked when the event deviates from knowledge gained from past experience and expectations. When the pursued goal moves away from the norms, counterfactual alternatives are created. Scenario building uses the same methodological approach to enhance understanding of the future (MacKay, McKiernan, 2004, p.173). Stories about the future, such as scenarios, are also a kind of historical balance sheet, but seen from the perspective of the future. They explain the state of the world in the future through a causal chain of events going back to the known present. In the counterfactual analysis of the past, by contrast, we do the same thing, but by changing the antecedent (which is an already clear event that took place). Past and future counterfactuals offer related types of analysis and scenario method forecasting that can potentially enrich one another.

It is important when making strategic decisions, especially how to mitigate or overcome existential risks, to always do historical analysis unless we are faced with a precedent. Strategic thinking and planning concerning the future considers the causes and effects of past and present events, as well as predictions and forecasts. The authors Lindgren and Bandhold (Lindgren, Bandhold, 2003) explore the relationship between the future and strategy, which allows the use of scenario planning. It receives further development at the strategic level as a kind of map for strategic "routes" and a choice between them according to different "relief" of the future and the environment. Based on this approach, a strict definition of the relationship between scenario and strategy is possible (Mason, Herman, 2003, p. 25-26).

Counterfactuals, the scenario-based forecasting method, and the strategies built upon this foundation can serve as effective tools in finding solutions to address existential risks. However, the question of working with individual consciousness and engaging institutions remains equally important.

6. Individual Approach to Existential Risks (Back to Taleb's Concept)

If we have to summarize, Taleb's idea of "skin in the game" is a critique of systems that allow decision-makers to escape accountability for their actions, and a call for individuals and organizations to take responsibility for the risks they take. How to deal with this when the risk is global, comprehensive, and in some cases long-term, beyond the deadlines associated with leaders and institutions? Karen Barad's work (Barad, 2012) may offer partial insights. One of her key contributions is the idea of "intra-action," which she uses to describe the ways in which different entities are entangled and co-constituted through

their interactions. According to Barad, entities do not exist in isolation, but are always already part of a larger network of relationships and interactions that shape their identities and actions. From this perspective, risk can be understood as a phenomenon that emerges from the complex interactions between different entities and systems. Risk is not simply a property of a given entity or system, but rather an emergent property that arises from the relationships between them. Barad also emphasizes the importance of recognizing the ethical implications of our interactions with others and with the environment. She argues that our actions and decisions always have consequences for others, and that we must take responsibility for these consequences and work to mitigate harm. Barad tries to reform the concept of rationality via a new kind of empiricism. Overall, Barad's work offers a nuanced and holistic perspective on risk, emphasizing the ways in which it is shaped by the complex interactions between different entities and systems, and calling for an ethical approach to understanding and managing risk.

Implicitly, this is also a new form of ethics that must be a priori placed before interaction with others and the environment, as it is a future component of risk or its prevention. We must be able to deconstruct at any moment these even sometimes invisible complex connections and relationships, so that they can be understandable, manageable, and subject to control, which may require corresponding responsibility. All authors to whom I refer call for a new ethics that is embedded in decision-making and risk management. Among them are David R. Boyd, Naomi Klein, Greta Thunberg.

7. Conclusion

The article aims to explore the relationship between the concept of risk society and the nature of existential risks, with a focus on the anthropocentric nature of some of these risks. “Risk society” is a way for a different approach to existential risks. It goes beyond the categories of "good" and "bad" and requires mobilization of actions to avoid one-dimensional catastrophic and alarming scenarios that condemn people to anxiety (mental problems), passivity, and anticipation of destruction.

The elements of anthropogenic existential risks are connected to numerous psychological and neurological characteristics of contemporary society, such as bounded rationality and irrationality (heuristics and biases), which involve a complex decision-making process, avoidance of responsibility, or the dilution of responsibility. This necessitates the development of new concepts for our era, new approaches to forecasting that have an internal resource for deconstructing the cascading and complex nature of risk. Therefore, I propose a new perspective - through the prism of counterfactual analysis and scenario building. However, it is of utmost importance that existential risks are recognized and reimagined at all levels - personal, institutional, national, and supranational.

References

Allbrow, M. (1992). *The Global Age*. Stanford University Press.

Barad, K. (2010). Quantum entanglements and hauntological relations of inheritance: Dis/continuities, spacetime enfoldings, and justice-to-come. *Derrida Today.*

Baum, S. (2013). Existential Risk and Cost-Effective Biosecurity. In *Health Security*. Johns Hopkins University Press.

Baumann, Z. (1992). *Institutions of Postmodernity*. London: Routledge.

Beck, U. (1992). *Risk Society: Towards a New Modernity*. Sage.

Beck, U. (2006). Living in the World Risk Society. Public Lecture at London School of Economics and Political Science, February 15th. Economy and Society, 35(3), 329-345.Beck, U. (2006a). *The Cosmopolitan Vision*. Polity Press.

Bloch, E. (1995). *The Principle of Hope*. MIT Press.

Bora, A. (2006). Risk, risk society, risk behavior, and social problems. In Ritzer, G. (Ed.), *The Blackwell Encyclopedia of Sociology* (Vol. 8, pp. 3926–3932). Oxford, UK: Blackwell.

Boyd, D. (2012). *The Environmental Rights Revolution: A Global Study of Constitution, Human Rights and the Environment*. UBC Press.

Douglas, M., & Wildavsky, A. (1983). *Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers*. University of California Press.

Douglas, M. (1992). *Risk and Blame: Essays in Cultural Theory*. London: Routledge.

Foucault, M. (1995). *Discipline and Punish* (A. Sheridan, Trans.). Vintage Books.

Giddens, A. (1992). *Risk Society: The Context of British Politics*. Polity Press.

Gordon, T., & Todorova, M. (2019). *Future studies and counterfactual analysis: Seeds of the future*. Palgrave Macmillan.

Habermas, J. (1996). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press.

Heidegger, M. (1962). *Being and Time* (J. Macquarrie & E. Robinson, Trans.). Harper and Row.

Holton, G. (2004). Defining Risk. *Financial Analysts Journal*, 60(6), 19-28.

Ilieva, B., & Todorova, M. (2023). On Desired Remote Possibilities of the Future: Could Counterfactual Analysis Challenge Prognostic Reflexes? *Journal of Future Studies*, 2023, ISSN: 1027-6084.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kahneman, D., & Miller, D. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93, 136-153.

Klein, N. (2007). *The Shock Doctrine: The Rise of Disaster Capitalism*. Knopf Canada.

Knight, F. (2006). *Risk, Uncertainty and Profit.* Dover Publications.

Lash, S. (2000). *The Risk Society and Beyond: Critical Issues for Social Theory*. Chapter DOI: https://doi.org/10.4135/9781446219539.

Latour, B. (2003). Is Re-Modernization Occurring—And If So, How to Prove It? *Theory, Culture & Society*, 20(2), 35-48.

Lindgren, M., & Bandhold, H. (2003). *Scenario planning: The link between future and strategy*.

Luhmann, N. (1995). *Social Systems*. Stanford University Press.

Mason, D., J. Hermann. 2002. Scenarios and strategies: making the scenario about the business, in *Strategy & Leadership*, 31(1), pp. 23-31.

Maturana, H., & Varela, F. (1980*). Autopoiesis and Cognition*. Dordrecht: D. Reidel.

MacKay, P., P. McKiernan. 2004. The role of hindsight in foresight: refining strategic reasoning, in *Futures,* 36, pp. 161–179.

Mythen, G. (2004). *Ulrich Beck: A Critical Introduction to the Risk Society*. Pluto Press.

Peterson, G., Cumming, G., & Carpenter, S. (2003). Scenario Planning: A Tool for Conservation in an Uncertain World. *Conservation Biology*, 17(2), 358-366.

Reith, G. (2004). Uncertain Times: The Notion of "Risk" and the Development of Modernity. *Time & Society*, 13(2-3), 383-402.

Rees, M. (2003). *Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future in This Century—On Earth and Beyond*. Basic Books.

Sinek, S. (2019). *The Infinite Game*. Penguin.

Todorova, M. (2015). Counterfactual Construction of the Future (Building a New Methodology for Forecasting). *World Future Review*, 7(1), 30-38.

Todorova, M., & Gordon, T. (2017). Report on a Study of Counterfacts as a Futures Research Technique for Forecasting Future Developments. *World Future Review,* 9(2), 93-105

Stanford | Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

The International Panel on Global Catastrophic Risks (IPGCR)

R. Daniel Bressler [1]*, Jeff Alstott [2]

Citation: Bressler, R. Daniel and Jeff Alstott. The International Panel on Global Catastrophic Risks (IPGCR). *Proceedings of the Stanford Existential Risks Conference* 2023, 233-247. https://doi.org/10.25740/tb895xv0451

Academic Editor: Paul Edwards, Trond Undheim, Dan Zimmer

Funding: R.D.B: Funding from the Open Philanthropy Project and Columbia University is gratefully acknowledged. J.A: Funding from the MIT Media lab is gratefully acknowledged.

Conflict of Interest Statement: The authors declare no conflict of interest. The opinions represented in this article are the personal opinions of the authors and do not necessarily reflect the opinions of their institutions.

Informed Consent Statement: N/A

Acknowledgments: N/A

Author Contributions: R.D.B. and J.A. wrote the initial draft of the paper in 2017. R.D.B. made final edits and revisions in 2023.

Abstract: This article motivates and describes a potential International Panel on Global Catastrophic Risks (IPGCR). The IPGCR will focus only on GCRs: risks that could cause a global collapse of human civilization or human extinction. The IPGCR seeks to fit an important and currently unoccupied niche: an international expert organization whose only purview is to produce expert reports and summaries for the international community on risks that could cause a global collapse of human civilization or human extinction. The IPGCR will produce reports across scientific and technical domains, and it will focus on the ways in which risks may intersect and interact. This will aid policymakers in constructing policy that coordinates and prioritizes responses to different threats, and minimizes the chance that any GCR occurs, regardless of its origin. The IPGCR will work in some areas where there is more consensus among experts and some areas where there is less consensus. Unlike consensus-seeking organizations like the Intergovernmental Panel on Climate Change (IPCC), the IPGCR will not necessarily seek consensus. Instead, it will seek to accurately convey areas of consensus, disagreement, and uncertainty among experts. The IPGCR will draw on leadership and expertise from around the world and across levels of economic development to ensure that it promotes the interests of all humanity in helping to avoid and mitigate potential global catastrophes.

Keywords: global catastrophic risks, global governance, international organization, research synthesis

[1] Ph.D. Candidate, Columbia University School of International and Public Affairs, 420 W 118th St, New York, NY 10027; rdb2148@columbia.edu.

[2] The Massachusetts Institute of Technology, Cambridge, MA 02139..

* Correspondence: email address: rdb2148@columbia.edu

1. Introduction and Rationale

Global catastrophic risks (GCRs) are *risks that could cause a global collapse of human civilization or human extinction* (Bostrom 2013, Bostrom & Cirkovic 2011, Posner 2004). Addressing these risks requires good policy, which requires a good understanding of the risks and options for mitigating them. However, primary research is not enough: policymakers must be informed by objective summaries of the existing scholarship and expert-assessed policy options.

We propose the creation of the Intergovernmental Panel on Global Catastrophic Risks (IPGCR). The IPGCR is an international organization that synthesizes scientific understanding and makes policy recommendations related to global catastrophic risks. The IPGCR will report on the scientific, technological, and socioeconomic bases of GCRs, the potential impacts of GCRs, and options for the avoidance and mitigation of GCRs.

The IPGCR will synthesize previously published research into reports that summarize the state of relevant knowledge. It will sit under the auspices of the United Nations, and its reports will include explicit policy recommendations aimed at informing decision-making by the UN and other bodies. To draw an analogy, the IPGCR does not put out forest fires; it surveys the forest, and it advises precautionary measures to minimize the chance of a forest fire occurring.

The IPGCR's reports will aim to be done in a comprehensive, objective, open, and transparent manner, including fully communicating uncertainty or incomplete consensus around the findings. The mechanisms for how this will be accomplished are described throughout this document.

The IPGCR draws on best practices from other international organizations and adopts those that best fit within the IPGCR's purview. Like the US National Academy of Sciences, the UK Royal Society, and the Intergovernmental Panel on Climate Change, the IPGCR will primarily operate through expert volunteers from academia, industry, and government, who will write and review the reports**.** In contrast to these other institutions, the IPGCR will be explicitly focused only on how potential risks could destroy civilization globally or cause human extinction. For instance, an IPGCR report on climate change would consider climate change as a GCR itself, as well as how climate change could affect other GCRs, both in terms of their probability of occurrence and the magnitude of their effects. This contrasts with, e.g., IPCC and World Health Organization (WHO) publications, which are meant to provide broad and comprehensive assessments of global health[1] and climate change respectively (IPCC 2014, Hulme 2022). While the IPGCR will likely draw from some of the same experts and knowledge base as these organizations, it will have a sharper focus on global catastrophes. Finally, the IPGCR is constructed to assess GCRs across domains, unlike the domain-specific WHO and IPCC. This will allow the IPGCR to compare potential risks on a like-for-like basis, and to understand how they may intersect and interact. This will aid policymakers in constructing policy that coordinates and prioritizes responses to different threats, and minimizes the chance that any GCR occurs, regardless of its origin.

[1] Key WHO publications cover a broad range of global health topics including: *The World Health Report*, *World Health Statistics*, *International Travel and Health*, *International Health Regulations*, *The International Classification of Diseases*, and *International Pharmacopoeia*. See: "Publications: Key Publications." World Health Organization. Accessed 23 Sep. 2017. http://www.who.int/publications/en/

The IPGCR's structure mandates that leaders have diverse expertise, but also that they come from countries across different regions and levels of economic development. The IPGCR will draw on voices across geographies and levels of economic development. This minimizes real and perceived favoring of particular groups, elevates these risks as truly global concerns, and ensures that the IPGCR promotes the interests of all humanity by helping to avoid and mitigate potential global catastrophes.

The IPGCR will sit under the auspices of the United Nations, and it will be accountable to all participating countries through the IPGCR Delegates Assembly. Each participating country gets a single Delegate in the Assembly. Delegates are both political and technical representatives of their countries. They elect the IPGCR's management, sign off on changes in strategic direction, and endorse findings and recommendations. The Delegates' oversight tools are shaped so that they do not unnecessarily impede the IPGCR's flexibility or speed. For example, Delegates can approve the IPGCR's actions remotely (without meeting in person), reports' scientific findings have a lower bar for endorsement than their policy recommendations, and only 90% of Delegates are necessary to endorse a policy recommendation. The governance structure thus gives the IPGCR's actions broad legitimacy while staying nimble.

In short, the IPGCR process starts with review of a diffuse array of scholarship and creates a path for analysis and recommendations to ratchet up to higher and higher levels of buy-in, agreement, and authority. The IPGCR would improve existing global governance by making it more effective in noticing, understanding, and combating global catastrophic risks.

The rest of the article is structured as follows: section 2 provides a brief background on global catastrophic risks. Section 3 describes the content produced by the IPGCR. Section 4 describes the structure and operations of the IPGCR. Section 5 describes the process for the adoption and implementation of the IPGCR.

2. Background

For the purposes of the IPGCR, *global catastrophic risks are defined as risks that could cause a global collapse of human civilization or lead to human extinction*. Risks that have the potential to pose a catastrophe of this scale and severity will fall under the purview of the IPGCR's reports. Risks that are not global in scope or are not severe enough to cause global destruction of civilization, while important, are not covered by the IPGCR. Potential GCRs covered by the IPGCR must have a *direct mechanism* by which they contribute toward a global catastrophe. Broad social issues such as inequality, poverty, education, and health are critically important, but will not be directly covered by the IPGCR. These issues may contribute towards instability that makes a global catastrophic risk more likely and can be discussed in this context, but these issues will not be the focal point of the IPGCR's work.

While there have been many catastrophic events in human history, GCRs have little precedent. World War II caused approximately 60 million deaths[2] or ~3% of world population. The 1918 Influenza Pandemic caused 20-40 million deaths[3] or 1-2% of world population. The 14th Century Black Death caused approximately 50 million deaths

[2] *Deaths by Country*. Available at http://enroll.nationalww2museum.org/learn/education/for-students/ww2-history/ww2-by-the-numbers/world-wide-deaths.html. (Accessed: 25 Nov. 2016).
[3] *The Influenza Pandemic of 1918*. Available at https://virus.stanford.edu/uda/. (Accessed: 25 Nov. 2016).

(Benedictow 2005) or ~13% of world population (Roser & Ortiz-Ospina 2017). There have been collapses of civilization on a regional scale, including the Roman Empire in the 5th century AD, the Classical Maya in the 9th century, and the Rapa Nui of Easter Island in the 17th century (Diamond 2005). Although tragic, these events are not included under the purview of GCRs as defined here because they did not lead to human extinction or a collapse of civilization globally.

One historical event which may have risen to the level of a GCR was the Toba super-volcano eruption in Sumatra ~74,000 years ago, which led to a volcanic winter (Rampino & Self 1992). Because it occurred in prehistory, there is significant uncertainty about the consequences of this event, but some studies suggest that there may have only been a few thousand humans worldwide that survived the event (Ambrose 1998, Oppenheimer 2002).

GCR-relevant research is currently being done in numerous fields, including environmental science, biology, physics, geology, economics, statistics, psychology, philosophy, political science, computer science, engineering, and more. Most of this research is being carried out as part of the existing research agenda for each of these fields, not necessarily within the specific context of GCRs. However, several scholarly organizations specifically designed to study GCRs and related topics have been created recently. Examples of these institutions include the Oxford Future of Humanity Institute,[4] the Cambridge Centre for the Study of Existential Risk,[5] and the Stanford Existential Risk Initiative.[6] The IPGCR is not intended to replace or duplicate the research of any of these institutions, but to collect and summarize scholarly research relevant to GCRs (regardless of whether that research was originally intended to have GCR implications), and to produce reports that can inform policy.

3. IPGCR Content

This section describes the two key types of documents produced by the IPGCR: IPGCR Reports (3.2) and IPGCR Working Group Summaries (3.3). Each Working Group creates a series of reports throughout the 5-year cycle, and each Working Group creates a Working Group Summary at the end of the 5-year cycle. This section focuses on the content of these two types of documents; the processes and operations by which these documents are created are described in section 4.

3.1 Defining Scope and Working Groups

The IPGCR will produce reports on different GCR topics. The report-writing process will be administered by Working Groups, organized around different sub-topics of Global Catastrophic Risks. In section 4, we describe mechanisms by which Working Groups and reports are created.

While the Working Groups will ultimately be decided by the IPGCR's governance bodies (described in section 4), for purposes of illustration and clarity, we will assume that the IPGCR's governing bodies will create two Working Groups throughout the rest of this document: Object-level Risks (WG1) and Meta-Level Topics (WG2). The Object-level

[4] https://www.fhi.ox.ac.uk/
[5] http://cser.org
[6] https://seri.stanford.edu/

Working Group will produce reports on specific risks, while the Meta-Level Working Group will produce reports on topics that are not specific risks themselves, but are relevant to understanding and assessing GCRs. Below is *an illustration* of what the Working Groups and their report topics might look like:

Working Group 1 (WG1): Object-Level Risks
Assessments of specific risks that could be globally catastrophic. Potential Task Forces:

- Nuclear Weapons
- Biosecurity and Pandemic Preparedness
- Artificial Intelligence
- Climate Change
- Biodiversity Loss
- Interplanetary Object Impacts
- Super Volcanoes
- Cybersecurity Risks to Critical Infrastructure

Working Group 2 (WG2): Meta-Level Topics
Topics relevant to understanding, assessing, avoiding, and mitigating GCRs. Potential Task Forces:

- Forecasting methods, uncertainty, and high-impact statistical tails
- Systemic risk mechanisms, complex systems, and contagion effects
- Human psychology and decision-making as it relates to understanding and assessing GCRs
- Benefit-Cost Analysis of Catastrophes including topics in Ethics, discount rates, and measuring social welfare across time
- International Cooperation and collective action
- Creating resilient critical infrastructure
- Legal and regulatory methods to address GCRs
- Horizon scanning and preparing for yet unknown risks

3.2 Reports

The purpose of IPGCR reports is to assess all relevant scholarly information and then communicate the current state of the knowledge to policymakers. As such, unlike the IPCC, IPGCR reports will not necessarily seek consensus, but will identify and communicate the level of uncertainty around the report's findings. There may indeed be certain parts of a report topic that have wide consensus, while other parts may not have consensus and have high degrees of uncertainty. A major role for IPGCR reports is to present 'long tail' considerations - ideas that exist outside the range of consensus of probable futures, but are nevertheless worth consideration from policymakers. IPGCR reports will treat uncertainty explicitly so that readers will be able to assess and compare uncertainties across different reports.

Each IPGCR report will be written by a Task Force of expert volunteers, selected through a governance process described in section 4. IPGCR reports will use literature review, not original research. Priority will be given to peer-reviewed scientific, technical, and social-economic literature. Other literature, such as reports from governments and industry, can be used to expand the depth and breadth of the assessment. However, each Task Force must ensure the quality and validity of cited sources and information.

Regardless of the report topic or Working Group, each IPGCR report will have 3 main sections:

1. Report Findings
2. Report Technical Summary
3. Report Summary for Policymakers

3.2.1 Report Findings

The report findings section is the core of the report that provides a detailed description and overview of the state of current knowledge as it relates to the report's topic.

Working Group 1 (Object-Level Risks)
To make WG1 reports as useful as possible, the report findings section will have a standardized high-level structure. This will help to ensure continuity between the different Object Level Risks covered by WG1, and to ensure that reports on risks ranging from nuclear weapons to artificial intelligence to climate change will not be wildly different in the scope of their content. There will be a degree of authority given to Task Forces to make decisions that make sense for each specific report, but this will be within the standardized high-level structure outlined below:

- **Physical and Socioeconomic Bases of the Risk** – A literature review of the current state of the natural, social, engineering, and other relevant sciences to describe the risk.
- **Risk Assessment** – Qualitatively and quantitatively assess the risk in terms of:
    1. Direct Impacts and the Severity of Direct Impacts
    2. Indirect Impacts and the Severity of Indirect Impacts
    3. The Scope of Direct and Indirect Impacts
    4. Timing
    5. Likelihood of Occurrence

  A major part of the assessment will be determining the amount of uncertainty involved with the report's assertions.
- **Risk Avoidance and Mitigation** – A detailed discussion of different potential strategies for avoiding or mitigating the risk, including corresponding uncertainties. This section will not provide specific recommendations for policymakers, which are handled in the report's Summary for Policymakers section (discussed in **section 3.2.3**).

Working Group 2 (Meta-Level Topics)
Unlike WG1, WG2 covers a range of topics that are useful in understanding GCRs, but does not provide assessments of specific GCRs themselves. As such, WG2's report findings section will not have a systematized structure, but will be at the discretion of the Task Force, with oversight from the WG2 Board leadership.

3.2.2 Report Technical Summary

This section provides a summary of the report findings section using technical/scientific terms.

3.2.3 Report Summary for Policymakers

This section provides an overview of the report findings in lay terms. Importantly, this section also introduces policy recommendations. These recommendations will be informed by the review of existing scholarship, but the Task Force will have the freedom to consider risk avoidance and mitigation strategies that have not been previously implemented or tested. Still, whenever possible they should anchor these strategies in prior research.

The reports of Working Group 1 (Object-Level Risks) will include recommendations for avoiding and mitigating the risk covered by that report. Working Group 2 (Meta-Level Topics), however, may not necessarily include recommendations, since these reports are focused on providing a synthesis of topics relevant to GCRs, and not on assessing specific GCRs themselves. However, Working Group 2 reports can include recommendations if the Task Force feels that it has found concrete policy recommendations as part of its work.

3.3 Working Group Summaries

Every 5 years, each Working Group will create a summary of all new reports (process described in detail in section 4). This summary will be based on each report's individual Technical and Policymaker Summaries. The final Working Group Summaries will include:

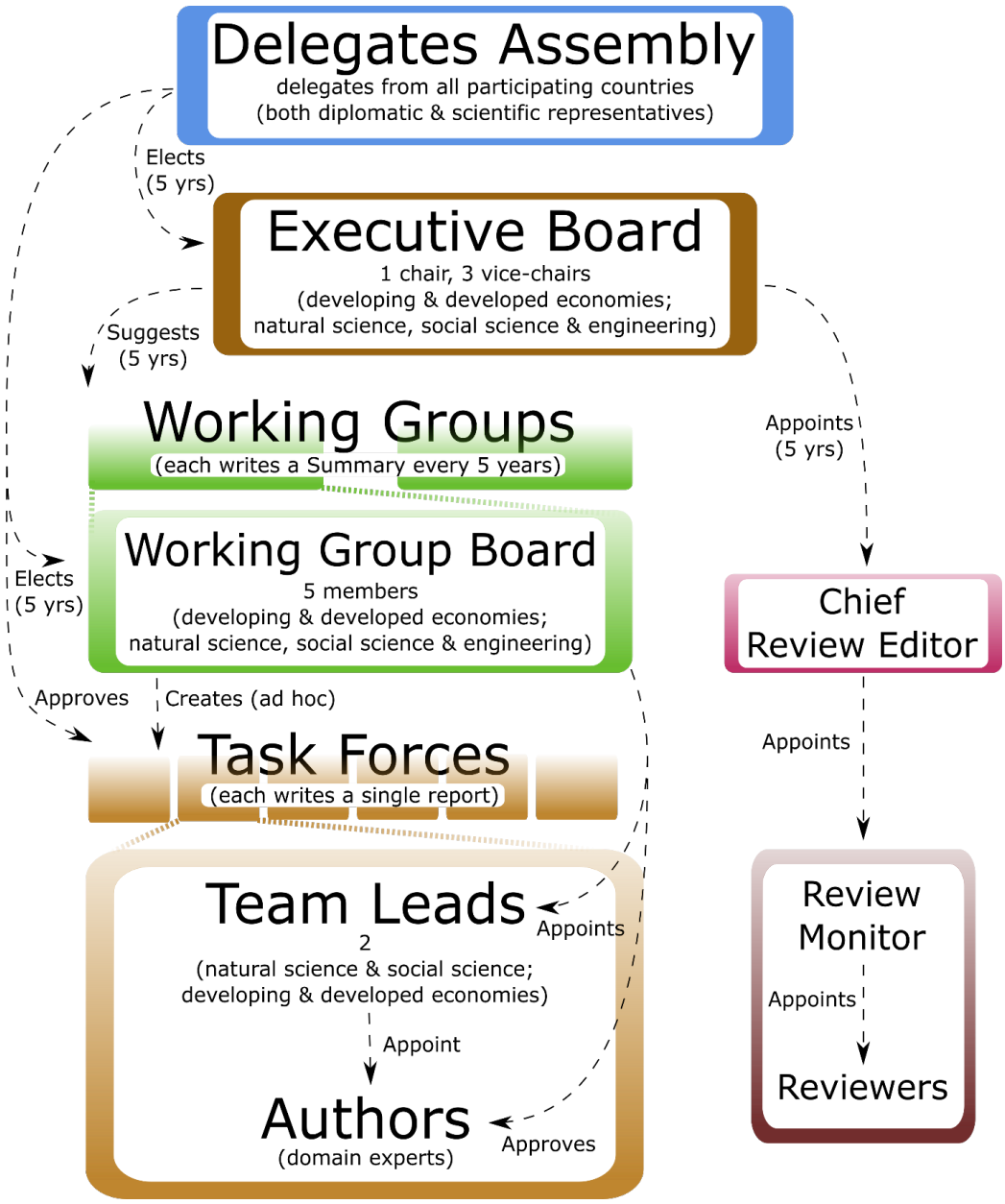
- **Working Group Technical Summary** - Provides a more detailed technical summary of the Working Group's reports.
- **Working Group Summary for Policymakers** - Provides an overview of the Working Group's reports in layman's terms, including recommendations.

Because these Working Group Summaries will summarize 5 years of the Working Group's efforts, they will likely be the IPGCR documents most widely read by both policymakers and the general public.

4. Structure & Operations

At a high level, the IPGCR will run on a 5-year cycle. Every 5 years, the Delegates Assembly (the IPGCR's representative body, described in more detail below) meets as a large group to discuss and vote on key aspects of the IPGCR's plan, policies, and procedures, including what Working Groups should exist. Working Groups will then act somewhat autonomously, with oversight from the Delegates Assembly, to produce reports as they see fit over that 5-year period. This structure will allow the IPGCR to create up-to-date reports that can address the often-changing state of knowledge in a timely manner. This structure will allow the IPGCR to be nimbler and responsive, and distinguishes it from a body such as the IPCC, which produces a single large report every 5-8 years. In the IPGCR, each Working Group will aggregate its reports every 5 years, creating a Working Group Summary. Working Group Summaries will likely be the most widely read IPGCR document. All IPGCR content will require endorsements from the Delegates Assembly in a process described throughout this section.

Figure 1: ***IPGCR governance structure****. The IPGCR is designed as a nimble structure capable of frequently synthesizing varied research while obtaining broad buy-in and legitimacy from participating countries.*



4.1 Governance

4.1.1 Delegates Assembly

The Delegates Assembly is made of Delegates from all participating countries. Delegates will be both diplomatic and scientific representatives of their countries, and will approve IPGCR reports. The Delegates Assembly will have a large in-person meeting once every 5 years to discuss and vote on key aspects of the IPGCR, such as the IPGCR's processes and elections of Executive Board members. The Delegates Assembly will also have shorter annual meetings to monitor the IPGCR's ongoing operations. The Delegates Assembly

plays a key role in the IPGCR review process, described in sections 4.2 and 4.3 below. Due to the autonomous nature of the Working Groups and the ongoing nature of the Working Group reports, the Delegates Assembly will need to be available on an ongoing basis for remote review and remote voting on items that require their approval (described in detail in sections 4.2 and 4.3). This structure ensures all Delegates will have a say in the IPGCR's operations, while also keeping the organization nimble enough to respond to changes in the state of knowledge relating to global catastrophic risks.

4.1.2 Executive Board

The Executive Board will include the IPGCR Chair and 3 Vice-Chairs, which are elected by the Delegates Assembly. To ensure heterogeneous geographical representation, there must be at least one Vice-Chair from a developing economy and at least one from a developed economy. In addition, each of the Vice-Chairs should come from different disciplinary backgrounds; possible membership could be a natural scientist, a social scientist, and an engineer, but the exact disciplines will be up to the discretion of the Delegates Assembly. Once appointed, the Executive Board will manage the IPGCR on an ongoing basis and make executive decisions around management and content.

4.1.3 Working Groups

Working Groups are operational divisions of the IPGCR, which are created by the Executive Board and approved by the Delegates Assembly. Each Working Group has a Board of five members, elected by the Delegates Assembly. At least two of these members must be from a developing economy and at least two must be from a developed economy. The five Working Group Board members should represent at least three disciplinary backgrounds, most likely a natural scientist, a social scientist, and an engineer, although the exact disciplines are up to the discretion of the Delegates Assembly. Each Working Group Board will decide internally who will serve as Chair. The Chair will have an administrative role, with no additional executive or voting power.

Working Groups will have the flexibility to be restructured as needed, though this will likely be rare, and will need to be approved by the Delegates Assembly. Each Working Group will have 5-10 full-time staff, which will assist the Task Forces.

4.1.4 Review Team

In addition to the Working Groups that produce reports, there will be an independent Review Team, which puts reports through independent peer review. The Review Team will be managed by a Chief Review Editor, who is appointed by the Executive Board. The Chief Review Editor appoints an individual Review Monitor for each Working Group report (described in more detail below).

4.2 Report Creation and Review

The Working Group Boards will identify topics on which reports should be written. For WG1 (Object-Level Risks), the report topics will be potential GCRs. For WG2 (Meta-Level Topics), these reports will cover topics that are key to understanding, assessing, avoiding, and mitigating GCRs. When the Working Group Board proposes a report topic, the Delegates Assembly then approves or denies the Working Group's request to continue the report-creation process. Delegates have 1 month to respond with a yes/no vote. If

Delegates do not vote in time, their vote is not counted for or against. 90% approval is needed to proceed with the report.

4.2.1 Identifying Report Topics

The Working Group Boards will identify topics on which reports should be written. For WG1 (Object-Level Risks), the report topics will be potential GCRs. For WG2 (Meta-Level Topics), these reports will cover topics that are key to understanding, assessing, avoiding, and mitigating GCRs. When the Working Group Board proposes a report topic, the Delegates Assembly then approves or denies the Working Group's request to continue the report-creation process. Delegates have 1 month to respond with a yes/no vote. If Delegates do not vote in time, their vote is not counted for or against. 90% approval is needed to proceed with the report.

4.2.2 Assembling Task Forces

Two Team Leads

Each report will be written by Task Force, led by two Team Leads. The Team Leads will be selected by the Working Group Board. For WG1, one Team Lead must be a natural science or technical expert (e.g. a biologist, an engineer, or a computer scientist) and one must be a social science expert (e.g. an economist, a sociologist, or a legal expert). WG2's Team Leads should also be selected to maintain some heterogeneity of expertise or background, though there are no formal requirements. In all Working Groups, one Team Lead must be from a developing economy and one from a developed economy.

Team of Authors

Team Leads will assemble a team of authors, with the oversight and consent of the Working Group Board. The team of authors will write the report content, as managed by the Team Leads. All authors, including the Team Leads, will be volunteers, and will predominantly be based in academia, industry, or government.

Technical Support

The Task Force will be assisted by the Working Group staff. Each Working Group will have 5-10 technical support staff, who are funded by the developed economies that have representation on the Working Group Board.

Review Monitor

The Chief Review Editor will appoint a volunteer Review Monitor for the report, who will manage and oversee the review process. The Review Monitor will be an expert in the report topic, and will assemble expert reviewers to review and comment on the report. The Review Monitor will arbitrate any disputes between reviewers and authors.

4.2.3 Drafting the Report Findings

Once assembled, the Task Force is expected to draft the report's findings in about a year. The exact process of drafting the report will be up to the Task Force, but Task Forces will be funded to meet in person 2-4 times to coordinate and discuss the report's content. The workflow of drafting, reviewing, and endorsing the report is shown in Figure 2.1 and is described below.

4.2.4 First Draft Review (Expert Review)

The Task Force will suggest possible reviewers to the Review Monitor, but the Review Monitor will make final decisions on reviewers. The Review Monitor will invite a broad array of experts to read and critique the draft. These reviewers' comments will be sent to the Task Force anonymously, and the Task Force must respond to them. It is the role of the Review Monitor to mediate disagreements between the Task Force and the reviewers, and to ensure the reviewers' comments are properly responded to by the Task Force. This portion of the review process will typically take several months. Upon future publication of the report, the reviewers' names will be listed on the report. To aid in transparency, all reviewer comments, without attribution, will be publicly available online.

4.2.5 Second Draft and Construction of the Report Summary for Policymakers and the Report Technical Summary

After receiving and responding to comments from the expert reviewers, the Task Force will create a second draft. Expert reviewers may have raised gaps in knowledge that require the Team Leads to recruit additional authors to complete the second draft.

In addition, as the second draft is being written, two summary sections will be drafted by the Team Leads:

- **Report Summary for Policymakers** - Provides an overview of the report findings in layman's terms as well as recommendations. All report authors will be involved in the recommendation-creating process, as Team Leads will solicit recommendations from each author working on the report. However, the Team Leads will have the ultimate editorial authority to decide which recommendations are included in the final report.
- **Report Technical Summary** - Provides a technical summary of the report.

4.2.6 Second Draft Review (Expert and Government Review)

Upon completion of the second draft, the Review Monitor will redistribute the report findings to experts, who will review the second draft with a particular focus on new material that was added to the second draft. In addition, Delegates (who act as representatives of their countries) will receive the report findings and the first drafts of the report Summary for Policymakers and the report Technical Summary. Delegates can provide comments that are taken into account, but don't necessarily need to be responded to by authors. As mentioned, to ensure that reports can be created in a timely manner, Delegates will have 2 months to give comments. In addition to their review comments, Delegates will provide Yes/No approval of each recommendation, and the tally will be communicated to the Task Force (e.g. 68% approved of recommendation #3). Ultimately, Delegates will provide endorsements in final review as outlined below, so authors will want to consider Delegate comments in their final changes to the document.

Review Monitors will continue to mediate disagreements. Significant disputes will be mediated by the Working Group Board.

4.2.7 Final Changes and Review

The Task Force will make final changes to the documents. The document will then be sent to Delegates for endorsement. Delegates will provide the following types of endorsements for the different components of the report:

- Report Findings - Delegates provide an overall endorsement that the report presents a fair assessment of the state of current knowledge around the report subject.
- Report Technical Summary - Delegates provide section-by-section endorsement.
- Report Summary for Policymakers - Delegates provide line-by-line endorsement.
- The overall report findings section, each section of the report Technical Summary, and each line of the report Summary for Policymakers must receive 90% endorsement from Delegates to be finalized as an IPGCR publication.

4.3 Working Group Summary Creation and Review

Figure 2.1: Working Group Summary Creation Process

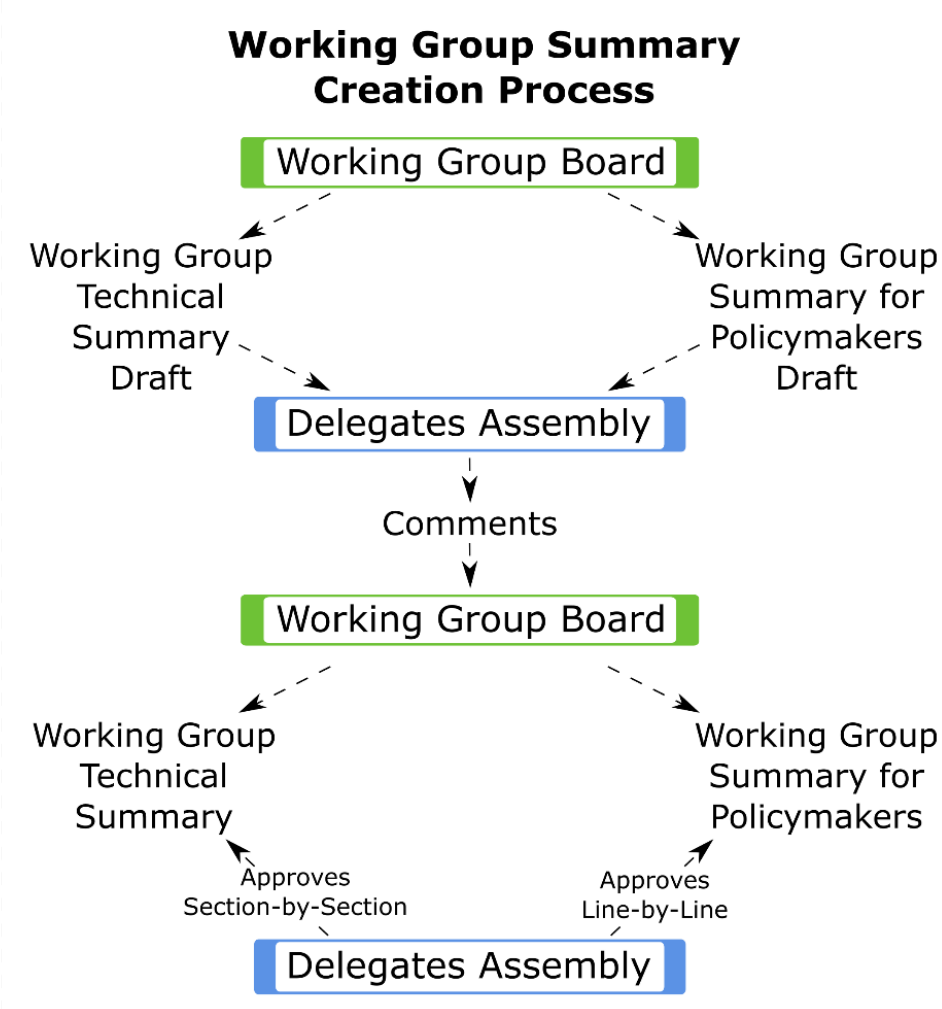
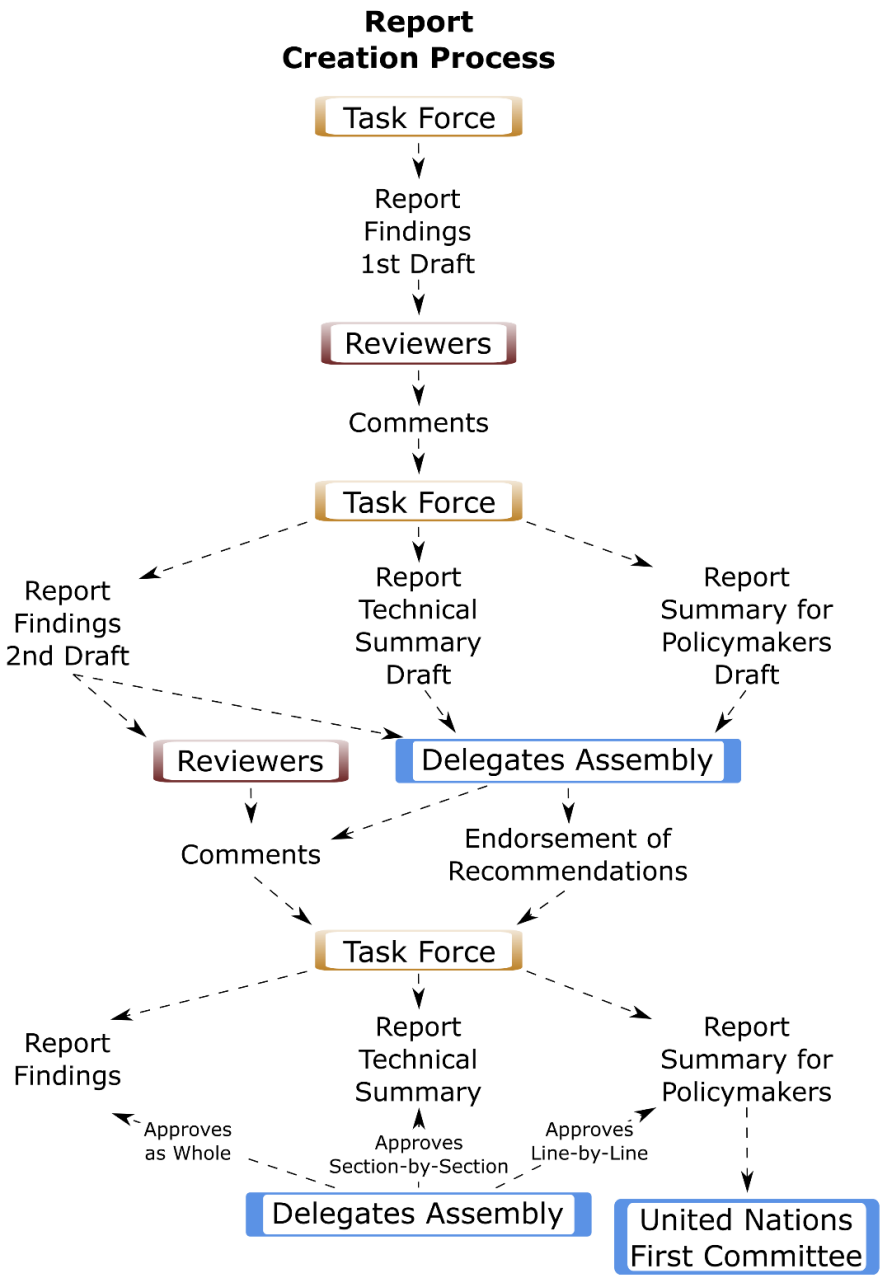


Figure 2.2: Report Creation Process



3 years into the IPGCR's 5-year cycle, the Working Group Board members will start drafting the Working Group Summaries, described in section 3.2. These Working Group Summaries will be the culmination of the Working Group's work over the 5-year cycle. They will summarize 5 years of the Working Group's work, and will likely be the documents most widely read by policymakers and the general public. The Working Group Summaries will include two parts:

- **Working Group Technical Summary** - Provides a more detailed technical summary of the Working Group's reports.
- **Working Group Summary for Policymakers** - Provides an overview of the Working Group's reports in layman's terms, including recommendations.

After drafting, the two Working Group Summaries will be sent to Delegates for review (review process shown in Figure 2.2). Since these summaries will largely build off each of the individual reports from the Working Groups, which have already been reviewed by experts, these summaries will only be reviewed by the Delegates Assembly.

As with Working Group reports, each Delegate will be given 2 months to provide comments on the Working Group summaries. Their comments will go back to the Working Group Board, which will then complete the final summary. In the 4th annual Delegates Assembly meeting within the Working Group cycle, Delegates will discuss and deliberate on the Working Group Summaries, then vote to endorse them, at different levels:

- Working Group Technical Summary - Delegates provide section-by-section endorsement.
- Working Group Summary for Policymakers - Delegates provide line-by-line endorsement.

As with the Working Group reports, each section of the Working Group Technical Summary and each line of the Working Group Summary for Policymakers must have at least 90% endorsement from Delegates to stay. Since Working Group Summaries will be endorsed as part of the Delegates Assembly meeting, all participating countries will have the opportunity to vote. However, since these Working Group Summaries build off of Working Group reports, which have already received Delegate endorsement, it will be unlikely for large parts of the Working Group Summaries to be rejected in this process.

5. IPGCR Adoption and Implementation

The IPGCR will be established by a resolution of the UN General Assembly, creating it as a subsidiary organ of the General Assembly. The majority of General Assembly resolutions are passed without a vote (United Nations 2022), including Resolution 43/53 that endorsed the creation of the IPCC in 1988 (United Nations 1988). Given that the benefits of addressing global catastrophic risks are by definition global, and the budget requirements are likely small—for reference, the annual budget of the IPCC is roughly $4.5M (IPCC 2016)—passing a UNGA resolution to create the IPGCR seems achievable.

Once adopted, the process to implement the IPGCR will proceed as follows:

1. Countries appoint their Delegates to the Delegates Assembly.
2. The Delegates Assembly votes on the Executive Board.
3. The Executive Board drafts the original plan, policies, and procedures for the IPGCR, including the decision on which Working Groups will be included.
4. The Delegates Assembly votes on the plan, policies, and procedures and elects Working Group Board members.
5. The Executive Board appoints a Review Editor.
6. Working Group Boards identify report topics.
7. The content creation process begins.
8. Working Group Boards release a summary of all reports every 5 years.
9. Every 5 years the Executive Board reviews the IPGCR's plan, policies, and procedures, including what Working Groups should exist. The Delegates Assembly meets in a large meeting every 5 years to review the Executive Board's recommendations, and to vote on the new plan, policy, procedures, Working Groups, and leadership for the next 5-year cycle of the IPGCR.

References

Ambrose, S. H.. (1998). "Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans." *Journal of Human Evolution*, 34, 623-651.

Bostrom, Nick. (2013). "Existential risk prevention as global priority." *Global Policy*, 4.1, 15-31.

Bostrom, Nick and Milan Cirkovic. (2011). *Global Catastrophic Risks.*Oxford University Press.

Hulme, M. (2022). *A Critical Assessment of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

IPCC. (2014). *Synthesis Report. Contribution of working groups I. II and III to the fifth assessment report of the intergovernmental panel on climate change*, 151(10.1017).

Benedictow, Ole J. (2005). *The Black Death: The Greatest Catastrophe Ever*. Available at https://www.historytoday.com/archive/black-death-greatest-catastrophe-ever. (Accessed 25 Nov. 2016).

Diamond, Jared M. (2005). *Collapse: How Societies Choose to Fail or Succeed*. New York: Viking.

IPCC. (2016). "IPCC TRUST FUND PROGRAMME AND BUDGET." Available at https://www.ipcc.ch/apps/eventmanager/documents/37/010320160933-Doc.%202%20-%20IPCC%20Programme%20and%20Budget.pdf.

Rampino, Michael R., and Stephen Self. (1992). "Volcanic winter and accelerated glaciation following the Toba super-eruption." *Nature* 359 (6390), 50-52.

Roser, Max and Esteban Ortiz-Ospina (2017). *World Population Growth*. Available at https://ourworldindata.org/world-population-growth/. (Accessed 16 Jun. 2017).

Oppenheimer, Clive. (2002). "Limited Global Change due to Largest Known Quaternary Eruption, Toba ≈74 kyr BP." *Quaternary Science Reviews*, 21, 1593-1609.

Oppenheimer, Michael, et al. (2007). "The limits of consensus." *Science Magazine's State of the Planet 2008-2009: with a Special Section on Energy and Sustainability*, 317, 1505-06.

Posner, R. A. (2004). *Catastrophe: risk and response*. Oxford University Press.

United Nations (2022). "UN Documentation: General Assembly Voting." Available at http://research.un.org/en/docs/ga/voting.

United Nations (1988). "General Assembly resolution 43/53, Protection of global climate for present and future generations of mankind, A/RES/43/53. Available at http://www.un.org/documents/ga/res/43/a43r053.htm.

Stanford
Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Crisis Government's Legitimacy Paradox: Foreseeability and Unobservable Success

Daniel D. Slate [1*]

Citation: Slate, Daniel D. Crisis Government's Legitimacy Paradox: Foreseeability and Unobservable Success. *Proceedings of the Stanford Existential Risks Conference* 2023, 248-259. https://doi.org/10.25740/zj321vj7513

Academic Editor: Dan Zimmer, Trond Undheim

Funding: The author's research has been generously funded by the Stanford University Department of Political Science.

Conflict of Interest Statement: The author has no conflicts of interest to declare.

Informed Consent Statement: N/A.

Acknowledgments: The author would like to thank the many excellent comments from participants and fellow panelists at the third Annual Conference of the Stanford Existential Risks Initiative. Dan Zimmer's feedback, questions, and suggestions were particularly helpful.

Author Contributions: The author is sole author and had no co-authors on this paper.

Abstract: This paper contributes a rethinking of crisis government, starting from two observations. First, nearly all prior political theorizing about crisis focuses on imminent or already-present *threats*, rather than more amorphous forecasted *risks*. Historical theories justify drastic responses only when an actual, present threat exists, and are silent about how to approach the risk of total societal destruction on the basis of uncertain forecasts and longer time scales. Second, prior theories rely on a legitimacy principle based on retrospective, *ex post* approval of norm violations that stopped an *observed* threat or a directly experienced emergency. These theories give little guidance about whether crisis governments can operate instead based on foresight and anticipation. If their interventions succeed, such governments would prevent forecasted disasters from occurring, but the actual presence of an emergency—which extant theory requires to legitimate crisis government—would then be unobserved. Such a government's very success would prevent the actualization of its legitimacy, presenting a paradox for present theory. Drawing on the Talmud's discussion of how to preclude a civilization-scale catastrophe, this paper contributes a new theory of crisis government that answers how we can legitimately act on the basis of foresight to address the anticipated exigencies of existential risks.

Keywords: emergency powers, crisis government, existential risk, foresight, Talmud

[1] Stanford University, Encina Hall West, Suite 100, Stanford, CA 94305-6044; dslate@stanford.edu.

* Correspondence: dslate@stanford.edu.

1. Introduction

More than anything else, existential risks concentrate the mind on the imperative to make high-quality decisions in the face of imminent or forecasted emergencies. Political theorists and jurists have long grappled with this subject, and their writings preserve many insights relevant today. However, the current prospect of interacting and cascading crises should challenge us to revisit whether extant jurisprudence and political theories—and the institutions based on them—adequately address the nature and scale of twenty-first century existential risks (X-risks).

This paper contributes a rethinking of crisis government, starting from two observations. First, nearly all prior political theorizing about crisis focuses on imminent or already-present *threats*, rather than more amorphous forecasted *risks*. In other words, historical theories justify drastic responses only when an actual, present threat exists; merely forecasted disasters are insufficient. Relying on various forms of the law of necessity, prior theory's guidance appears self-similar, whether the scale of the challenge is an individual "taking the law into one's own hands" when facing imminent death, or a political community violating normal rules for the sake of its population's survival in the face of "supreme emergency." An exception to this trend in the history of ideas, found in the Talmud, does attend to uncertain forecasts, longer time scales, and the risk of total societal destruction, and thus offers us a useful starting point for new theorizing that can meet the challenge of the present.

Second, prior theories rely on a legitimacy principle based on retrospective, *ex post* approval of norm violations that stopped an *observed* threat or a directly experienced emergency. Locke's influential discussion of prerogative and Cicero's legal maxim *salus populi suprema lex esto* ("the safety of the people shall be their highest law") are representative of this trend. Unfortunately, these theories give little guidance about whether crisis governments can operate instead based on foresight and anticipation. If their interventions succeed, such governments would prevent forecasted disasters from occurring. However, the actual presence of an emergency, which extant theory requires to legitimate crisis government, would be unobserved; such a government's very success would prevent the actualization of its legitimacy, presenting a paradox for present theory.

This paper thus contributes a new theory of crisis government that answers how we can legitimately act on the basis of foresight to address the anticipated exigencies of X-risks. It proceeds in three parts. First, it surfaces the principles underlying traditional models of crisis government, including the Roman dictator, the royal prerogative, and doctrines of supreme emergency. Second, it identifies the legitimacy paradox that arises from attempting to apply the traditional theories to longer-term existential risks rather than immediately imminent extant threats. Third, it develops principles for a new theory by recovering the Talmud's alternative approach to crisis government.

2. Crisis Government: The Traditional Models

2.1 The Roman Dictator

In the first book of his *Discourses*, Niccolò Machiavelli celebrated ancient Rome's practice of lawfully entrusting great power for a limited term to a dictator at occasional times of crisis, carefully distinguishing this role from a tyrannical citizen who usurped power against the constitution, its laws, and an explicit public trust. "[T]heir power to act," he wrote in the thirty-fourth chapter, "was confined to the particular occasion for which they were created" (Machiavelli, 1531/1950, p. 202). While during "the pressing danger" of the crisis a dictator could "decide alone upon the measures to be adopted for averting" it, "the

Dictator could do nothing to alter the form of the government, such as to diminish the powers of the Senate or the people, or to abrogate existing institutions and create new ones" (Machiavelli, p. 202). Machiavelli went on to praise the great benefits the dictatorship brought to republics like Rome and Venice: while republics' normal operations are slow and time-consuming, a dictator could move swiftly to decide "urgent matters" "when the occasion requires prompt action" (p. 203). He elaborated:

> When a republic lacks some such system a strict observance of the established laws will expose her to ruin; or, to save her from such danger, the laws will have to be disregarded. Now in a well-ordered republic it should never be necessary to resort to extra-constitutional measures…Thus no republic will ever be perfect if she has not by law provided for everything, having a remedy for every emergency, and fixed rules for applying it. (Machiavelli, p. 203)

Machiavelli saw in the Roman dictator a legally defined and constitutionally constrained role. With the ability to name a dictator (who would be checked by parallel institutions, which the dictator was not empowered to disestablish) a "well-ordered" republic never need act outside the law to respond to crises. He contrasted the dictator, meant to answer a short-term crisis, with the extraordinary powers entrusted to the Roman Decemvirate, who the people elected to remake Rome's entire constitutional order—a power the Decemvirs ultimately abused (Machiavelli, First Book, chs. 35 & 40).

For Machiavelli, the nature of a crisis parallels the attributes of the dictator: it presents an immediate need, a pressing necessity that threatens the republic's existence, but it is only expected to last for a short term. The unique role of the dictator is to violate the established constitutional order to create the conditions for safely reinstating the constitutional order. After the crisis passes in the near future, everything will return from the extraordinary to the ordinary, corresponding to the expiration of the dictatorship's term.

Jean-Jacques Rousseau, like Machiavelli, favored the model of the Roman dictator. In principle, the temporary, elected "supreme chief" who in a time of crisis "silences all the laws and provisionally suspends the Sovereign authority" posed no problem for him; what objections Rousseau had to the history of the dictatorship owed solely to the Romans' occasional failure to use it as he, writing centuries later, thought best. (Rousseau 1762/1997, p. 138).[1] Taking the dictator as his exemplar, Rousseau offered a general rationale for raising up a time-limited crisis government. When there is true peril, threatening societal annihilation (for "only the greatest dangers can counterbalance the danger of disturbing the public order"), then "the general will is not in doubt, it is obvious that the people's foremost intention is that the State not perish" (Rousseau, p. 138).[2] Rousseau had explained from the beginning:

> The inflexibility of the laws, which keeps them from bending to events, can in some cases render them pernicious, and through them cause the ruin of a State in crisis. The orderliness and deliberateness of formalities requires a space of time which circumstances sometimes deny one. A thousand cases can arise for which the Lawgiver did not provide, and it is a very necessary foresight to sense that one cannot foresee everything. (p. 138)

[1] For more on the Roman model itself, see Lazar 2013.

[2] For Rousseau's technical definition of the State as the 'passive' mode of the sovereign body politic, see Book I, ch. 6 ("Of the Social Pact"), p. 53. Rousseau insisted that "Government" was not to be confused with the State or the Sovereign, but was merely "its minister," commissioned by the sovereign people. Ibid., Book II, ch. 6, p. 67, and Book III, ch. 1, pp. 82-83. Rousseau consistently describes the dictator in Book IV, ch. 6, as just another commissioned officer, an administrator of the general will.

In other words, politics must adapt to contingency and unpredictable events. In Rousseau's formulation, order and legal rigidity must "bend" or re-form, even at times giving way to a temporary disorder, so as to overcome the worse outcome of total ruin. Throughout the crisis, however, the ordinary constitution remains intact, just quiescent and suspended for a time; the dictator cannot remake it. "This way the suspension of the legislative authority does not abolish it: the magistrate who silences it cannot make it speak, he dominates it without being able to represent it; he can do everything, except make laws" (pp. 138–39). For Rousseau, the dictator is a necessary supplement to the lawmaker, but not a replacement for the legal order, even when it has structural deficiencies. On this approach, the legitimacy of crisis government and its decisions rests on (1) the immediacy and gravity of the threat and (2) the inability of the crisis leader to do anything other than answer the crisis, serving only "for a very brief term which can never be extended" (Rousseau, 1762/1997, p. 140).

2.2 Prerogative Power

John Locke in the *Second Treatise* gave a similar description of the nature of exigency and crisis but drew different institutional conclusions. Like Machiavelli and Rousseau, Locke observed that events might occasionally require a quick response (Locke 1689/1988, § 160). He concurred that shortness of time and pressing immediacy characterize a crisis and that crises are recurring if irregular features of political life (§§ 160 & 167), consistent with his earlier premise that "Things of this World are in so constant a Flux that nothing remains long in the same State" (§ 157; Fatovic 2009; Feldman 2013).

Where Locke differed was in how to address them. Unlike Machiavelli and Rousseau, Locke favored lodging the powers of crisis leadership in a permanent institution, rather than a temporary magistrate. Conceiving a good government as one with separated powers, Locke thought of the legislative power as slower to move and more burdensome to coordinate than the executive, which concentrated power in a single person. Locke held the responsibility for responding to events requiring prompt action should be entrusted to the executive, but in all other cases where time was not short, power properly remained with the legislature (§ 159).

Like other theorists, Locke acknowledged that the ordinary ordinances may obstruct an effective response to a crisis. What justified the prerogative-wielding executive "to act according to discretion for the publick good, without the prescription of the law and sometimes even against it" (§ 160), was the ancient maxim of *salus populi suprema lex* ("the safety of the people shall be their highest law"), which Locke described as "certainly so just and fundamental a Rule, that he, who sincerely follows it, cannot dangerously err" (§ 158). This legal maxim, tracing back to Cicero's *De Legibus*, went on to assume a central place in European and American jurisprudence and political theory (Cicero, 58–43/1928, pp. 466–67; Coke, 1727, 10:139b (Keighley's Case); Hobbes 1651/1994; Kent, 1860, 1:253, 2:275; Broom, 1874, p. 1; Munir, 1957, p. 299). Locke explained:

> Many things there are, which the Law can by no means provide for, and those must necessarily be left to the discretion of him, that has the Executive Power in his hands, to be ordered by him, as the publick good and advantage shall require: nay, 'tis fit that the Laws themselves should in some Cases give way to the Executive Power, or rather to this Fundamental Law of Nature and Government, *viz*. That as much as may be *all* the Members of the Society are to be *preserved*. For since many accidents may happen, wherein a strict and rigid observation of the Laws may do harm; (as not to pull down an innocent Man's House to stop the Fire, when the next to it is burning) […] (§ 159).

Importantly, Locke explained that prerogative was not an emergency power *per se*, but acting in conformance with a "Fundamental Law of Nature and Government" (*salus populi*). While Locke did not favor the model of the Roman dictatorship, he was nonetheless appealing to an ancient maxim that Cicero in *De Legibus* inscribed immediately before discussing that institution, when justifying the executive magistracy of the consuls, whose power united in the dictator (Cicero, 1928, pp. 466–67). Crisis government for Locke partook of a deeply lawful sort of politics: events and circumstances may be in constant flux, but the natural law that defines and delimits just prerogative remains fixed.

The English jurist William Blackstone affirmed Locke's definition of prerogative ("the discretionary power of acting for the pubic good") and explained that all statements in English law about the apparent absolute sovereignty of the crown arise from the reality of "how impossible it is, in any practical system of laws, to point out beforehand those eccentrical remedies, which the sudden emergence of national distress may dictate, and which that alone can justify" (Blackstone, 1765, 1:244). Prerogative was "wisely placed in a single hand by the British constitution, for the sake of unanimity, strength, and dispatch," for uniting "several wills" together "is a work of more time and delay than the exigencies of state will afford" (pp. 242–243). Instead of the Roman dictator, Blackstone compared the king to a Roman emperor, though he was careful to explain the many ways the English executive did not wield arbitrary power (p. 243).

In sum, Locke and Blackstone offered a permanent monarch wielding decisive power in times of crisis, while Machiavelli and Rousseau instead endorsed a constitutional republican dictator whose temporary powers paralleled temporary crises. All, however, characterized the crisis that justified extraordinary powers the same way: it is a sudden event, impressing an immediate threat upon the whole people who make up the state, necessitating prompt and decisive action.

2.3 Supreme Emergency: Exigency vs. Expediency

Michael Walzer offered a restatement of the legitimacy of law-transgressing emergency powers, borrowing language from a memo of Winston Churchill, who wrote, "The letter of the law must not in supreme emergency obstruct those who are charged with its protection and enforcement. It would not be right or rational that the aggressive Power should gain one set of advantages by tearing up all laws, and another set by sheltering behind the innate respect for law of its opponents. Humanity, rather than legality, must be our guide" (Walzer 2006, p. 245). He and his sources accept that the circumstances of "supreme emergency," which typically come unexpectedly, permit overriding the rights of neutrals and innocents (Walzer 2006, pp. 247–49 & ch. 16; Statman, 2006). "The world of necessity," Walzer wrote, "is generated by a conflict between collective survival and human rights" (p. 325). Yet, paradoxically, he insisted that "when we override them we make no claim that they have been diminished, weakened or lost. They have to be overridden…precisely because they are still there, in full force, obstacles to some great (necessary) triumph of mankind" (p. 247). Walzer usefully distinguished between the "moral necessity" imposed by a great threat and the justice of overcoming it, and the "instrumental or strategic claim" that no other means other than the rights-overriding extraordinary measures were available to meet it (p. 248). If the latter is not there – if, in other words, the extraordinary measures simply bring swifter victory rather than prevent annihilation in a total defeat, they are not justified by necessity but are instead "only a piece of expediency" (p. 249).

Walzer suggested that people desire to have leaders in extreme times who are willing to do the immoral for the sake of their societies (Walzer, 2006, p. 323; Walzer, 2007).

Ultimately, though, his formulation is analogous to Locke's reservation that prerogative is only appropriate when there is no time available for the legislative power to properly convene and address the present challenge through ordinary constitutional measures.

A supreme emergency has two necessary elements: "imminence of the danger" and its extreme nature – in particular "the danger must be of an unusual and horrifying kind" (Walzer, 2006, pp. 252–53). In Walzer's crystallization, "The two criteria must both be applied. Neither one by itself is sufficient as an account of extremity or as a defense of the extraordinary measures extremity is thought to require. Close but not serious, serious but not close—neither one makes for a supreme emergency" (p. 252). He elaborated to explain that a true supreme emergency is akin to the existential threat presented by Nazism, which posed the "threat of enslavement or extermination directed against" whole nations (pp. 253–54). Walzer suggested it was important to cultivate skepticism and "a wary disbelief of wartime rhetoric," noting the tendency of politicians to use the language of emergency and crisis to mobilize people for action even when the circumstances might not in truth match their cynical words: "We need to make a map of human crises and to mark off the regions of desperation and disaster. These and only these constitute the realm of necessity, truly understood" (p. 253).

3. The Legitimacy Paradox in Traditional Theory: Exigency and Imminence

All of these approaches that I have characterized as the traditional (because better-known) models of emergency conceive of crisis as an immediate threat compelling a rapid response. The ordinary operation of the laws or their administration is insufficient to meet the moment. It is something that the leaders and often the people themselves perceive to be imminently pressing in the present or at most the very near future (Rossiter, 1963; Ferejohn & Pasquino, 2004; Gross & Ní Aoláin, 2006; Ferejohn, 2023).

That implies a subtler point that the theorists discussed above do not discuss explicitly (except perhaps Locke, 1689/1988, § 161): the legitimacy of emergency or prerogative action under crisis is something accepted, understood, and often retroactively ratified by the populace precisely because the threat *materialized*. But what of a slow-moving crisis (if one can be so characterized), which a small number of those who are particularly perceptive—or possessing unusual foresight—detect, but no one else notices, or in fact disagrees is even looming on a horizon far-distant? What would legitimate extraordinary responses to an extraordinary *risk* that is not yet an imminent *threat*?

These points are worth stressing because the concept of crisis in our times may be changing—or, more accurately, there appear to be those today who would like it to change, perhaps because they sense the lurking legitimacy paradox. The most obvious instance is the appeal to the language of crisis to characterize a climate emergency (Mittiga, 2022). There can be no doubt about such a challenge's *scale*; what is lacking is the specific sort of immediacy that the traditional approaches require of a genuine emergency (a deficit that has been true now for decades). A challenge from global climate change has been "serious, but not close" for many years, and would allow more than enough time to convene duly constituted legislatures which could provide ordinary political responses to the challenge. It is not clear that anyone is sure that should we pass the threshold into imminent supreme threat (a "point of no return") the world's institutions will be able anymore to meet such a crisis.[3] The proper intervention point is earlier on the tree of decisions and events, before such risks cascade into uncontrollable, existential catastrophes. But it is not clear, from the familiar works in political and legal theory that

[3] A similar point has recently been made by Yudkowsky (2023) from the long history of artificial intelligence research and decades-old concerns about what the technology could one day become.

include among their premises a respect for the concept of humanity and individual rights, that something like climate model forecasts of events predicted to occur years or decades hence could justify emergency powers at an earlier stage; the classic works seem to rule it out entirely.

The traditional theories thus give little guidance about whether crisis governments can operate based on foresight. Such governments, acting from anticipatory wisdom, would prevent forecasted disasters from occurring if their extraordinary measures were successful. The actual manifestation of an imminent emergency, which traditional theory requires to legitimate crisis government, would then go unobserved. Such a government's very success would prevent the actualization of its legitimacy, presenting a paradox for present theory.

It therefore seems to me worthwhile to ask whether the history of political thought includes other approaches to crises and existential risks that include what I take to be two desiderata of our times: (1) respect for humanity and individual rights even under great exigency and (2) the ability to use extraordinary measures when the times require it, even if, in the absence of a close and imminent *threat*, there is instead only a great foreseeable *risk*. So far, I have identified only one.[4]

4. The Talmud's Model: Judges, Foresight, and the 'Law of the Hour'

The Roman invasion of ancient Judaea and destruction of Jerusalem brought an end to the Jewish commonwealth and a loss of national sovereign existence that would not be recovered for almost two thousand years. Political theorizing that draws inspiration from the Talmud—written down many centuries after the catastrophe—is thus in an important sense already post-apocalyptic, drawing on meditations already shaped in response to the end of the particular worlds of the First and Second Temples. The Babylonian Talmud (*Gittin* 55b–57b) includes a discussion of the events—small at first—that cascaded into national ruin and exile, which appropriately begins with a meditation on the need to be wary of unfavorable future outcomes:

> Rabbi Yohanan said, "What is the meaning of that which is written, 'Happy is the man who fears always, but he who hardens his heart shall fall into evil' (Proverbs 28:14)? Jerusalem was destroyed on account of Kamtza and Bar Kamtza. The King's Mountain was destroyed on account of a rooster and a hen. Beitar was destroyed on account of a shaft from a chariot." (Talmud, 500–700/2015, *Gittin* 55b).[5]

The preeminent elucidator of the Talmud, Rashi, explains what sort of fear the Talmud means: "*Who fears* – He is anxious to see what will come to be [lit. to see what will be born (*ha-nolad*)], so that there should not occur a disaster through this: 'If I will do this…'" (Talmud, *Gittin* 55b, Rashi s.v. *mifached*). In other words, the Talmud indicates that the national catastrophes it is about to describe occurred because of a lack of foresight. Rashi also appears to be connecting this text, which we will see has strong political resonance, with another passage elsewhere in the Talmud that defines the nature of wisdom. There (*Tamid* 32a), the Talmud describes an encounter between the Sages of Israel and one of the most famous rulers of antiquity:

[4] It is possible that James Harrington's *Commonwealth of Oceana*'s discussion of institutional design might be read as attending to longer time horizons to stave off republican theory's great fear, the corruption of the regime along the traditional Polybian anacyclosis, but his discussion is not specifically about crisis and existential emergency. Harrington, 1656/1992, and Remer, 2008. Dan Zimmer also reminded me that the medieval proverb, "For want of a nail," indicates an awareness of the possibility of a seemingly insignificant event producing a crisis cascade resulting in the loss of an entire kingdom.

[5] Translations are based on that of the Steinsaltz Talmud.

> Alexander of Macedon posed ten questions to the Desert Sages (*ziknei ha-Negev*) … He said to them, "Who is called wise?" They said to him, "Who is wise? The one who sees the future outcome [*ha-ro'eh et ha-nolad*].[6]
>
> Rashi: *The one who sees the future outcome* – The one who understands, from his mind, what is going to be—the events that are going to come—and is vigilant to take precautions about them. (Talmud, *Tamid* 32a, Rashi s.v. *ha-ro'eh et ha-nolad*).

The Desert Sages answered Alexander with a universal definition of wisdom: foreseeing the unexpected and acting accordingly. To borrow the language of decision and game theory, they described a mindset of backward induction: see, in the present, the future outcomes that may result from the possible choices one has available now, and choose carefully on the basis of that understanding. Remarkably, in a seemingly unrelated discussion, the Talmud (*Nazir* 32b) even mentions the destruction of Jerusalem as the archetypical example of an improbable future event that may be possible, albeit difficult, to anticipate [*nolad*]. Seen in the context of these related discussions, the Talmud's discussion of the national destruction and what might have stopped it announces a clear principle: the wisdom of seeing the future—and acting upon forecasts of even far distant, improbable but incalculably costly events—is the way to avoid evil and disaster.

The Talmud in Tractate *Gittin* goes on to discuss each of the three above-mentioned episodes to illustrate the principle. In the first, Bar Kamtza, plotting revenge after being slighted at a banquet, went and lied to the Roman Caesar, claiming a rebellion was brewing. The Caesar doubted him and sent a kosher offering to the Temple in Jerusalem. Bar Kamtza, undeterred, maimed the offering so that it would be ritually unacceptable, in order to make it appear the Jews were rejecting the Caesar. The national leaders, the Sages of Israel making up the Great Sanhedrin, thought to stop Bar Kamtza's plot using a temporary ruling (*hora'at sha'ah*, "law of the hour") that would deviate from standard Torah law, to offer an invalid offering in order to ensure peace would persist with Rome and head off a foreseeable existential crisis. However, one leader, R. Zechariah b. Avkulos, hesitated and refused his assent, fearing the populace would learn the wrong ritual law if the Sages made an exceptional ruling. Given the context in which it would be made, it would not be possible to give a full, public explanation of the reasons for the *hora'at sha'ah*. Disaster followed, and the Talmud sharply criticized R. Zechariah b. Avkulos' flawed decision to hold himself back, blaming him for the destruction of Jerusalem, the loss of national sovereignty, and the ongoing exile (Talmud, *Gittin* 55b–56a). The latter two cascades mentioned in the Talmud's discussion (the destruction of the communities of the King's Mountain and Beitar) do not relate explicitly to actions or inactions of government officials (*Gittin* 57a). One might nonetheless infer from both these latter accounts that, had the leaders of the named cities acted to restrain their people's outrage at Roman violations of cherished local customs, they could have avoided extreme disasters for their provinces.

The same principle—of foreseeing the not-yet-imminent and acting carefully to avert disaster—applies in all three cases the Talmud cites. At the critical moment when foresight could still have made a difference, there was not yet an active existential threat, only the latent risk posed by a potential overreaction by the imperial Roman government. Since the first case the Talmud discusses concerns itself specifically with a critique of a leader's decision at the moment of foreseeable crisis, I will focus on that one.

[6] The Talmud elsewhere (*Sotah* 49b) suggests Greek wisdom (*hokhmat yevanit*) is relevant as a language of government, but the Talmud here (*Tamid* 32a) uses a generic term (*hokhmah*) to define a universal wisdom as foresight about future outcomes, which is not restricted to specifically Greek wisdom.

The discussion at *Gittin* 55b–56a is not considered the Talmud's *locus classicus* about extraordinary exigent powers. Those discussions are found elsewhere, in tractates *Sanhedrin* (46a) and *Yevamot* (90b), and the relevant Talmudic sources are elucidated and elaborated at length by R. Zvi Hirsch Chajes in his definitive treatise on the subject (Chajes 1836/1958, ch. 5).[7] Nonetheless, R. Chajes, in the course of a lengthy comment on R. Zechariah b. Avkulos's decision, explained that the Talmud criticized him for failing to invoke precisely those extraordinary powers. He wrote:

> From this we see that, in the eyes of the Sages of the Talmud, R. Zechariah b. Avkulos' approach was not correct, for he said, "They [the people] will say, 'They're bringing maimed offerings on the altar.'" It is evident that according to the law it was permissible to do this because of fear of the [Roman] kingdom, and so too if they had killed Bar Kamtza they would have done so legally, because he was a pursuer [*sh-radaf*] after all of Israel and [the rule is] "the one who comes to kill you, arise early to kill him." However, because of the great diffidence of R. Zechariah b. Avkulos he did not strengthen himself to give a practical *halakhic* ruling…and to rely on a *hora'at sha'ah* [law of the hour]. (Talmud, *Gittin* 56a, R. Chajes s.v. *anavatnuto shel Rabbi Zechariah ben Avkulos hachrivah et Beitenu*, quoting *Magen Avraham*, *Orach Chaim*, Laws of Lulav, 656.8).[8]

The Talmud's critique thus contains a clear endorsement of using extraordinary powers preventively, based on foreseeing an existential risk of national disaster that had not yet even started to become an imminent threat. (After all, the Roman Caesar had not heard what would become of his offering and was certainly not yet even inclined to march on Jerusalem on the flimsy basis of what he suspected were Bar Kamtza's fabrications). What's more, R. Zechariah b. Avkulos' rationale is precisely the problem of unobserved success when averting a foreseeable but distant crisis.[9] He refused his assent to extraordinary measures because he anticipated the populace would judge their leaders based only on what they could see: that the Sages had apparently endorsed a rule that went against the Torah, with no explanation offered. The people would mistakenly infer that the exception must be the normal operative rule, and this, he believed, would itself be disastrous: either the people, trusting their leaders, would henceforth live according to a false impression of Torah law, or the people would lose faith in their leaders, mistakenly believing they had left the standard, normative path of the Torah without explanation. The people would never see that the *hora'at sha'ah* ruling prevented a war with Rome, that the extraordinary powers had averted a crisis. His reasoning, while understandable, was unequivocally rejected as not only wrong but the cause of catastrophe.

The Talmud instead endorses extraordinary measures even in circumstances when there is no visibly imminent threat, which upon being dealt with successfully would retroactively justify the exercise of exceptional powers. The wisdom of foresight is the legitimating principle.

---

[7] Characterizations of these powers as extra-legal (e.g., Gross, 2013) are mistaken, as the *hora'at sha'ah* power, while deviating from the standard law of the Torah, is still considered a canonical form of Talmudic lawmaking, albeit an unusual and rare one.

[8] R. Chajes's commentary, *Chiddushim al ha-Talmud*, can be found printed in the back of standard editions of the Talmud. For discussion of the other legal principles mentioned, specifically about the law of a pursuer (*rodef*), see Maimonides, 1180/2001, *Hilchot Rotze'ach* ('Laws of a Murderer'), and for a recent discussion of the principles, see Bleich & Jacobson, 2015, pp. 120–21.

[9] Deterrence theory partakes of a similar logic, though it works by different means. If criminological deterrence is effective, raising the cost of crime and dissuading the criminal from engaging in it, crime rates are lower than they would be in the absence of the deterrent; in nuclear deterrence, success is zero nuclear exchanges arising from leaders recoiling from the costs of fighting a nuclear war, the extraordinarily costly outcome foreseen by looking down the decision tree. Both criminal and nuclear deterrence, to claim success, rely on the assumption that a negative has occurred, and it is famously difficult to prove attribution to the deterrent factor. The Talmud elsewhere (*Makkot* 7a) records a debate that accepts the logic of deterrence on a society-wide level.

One final point worth noting is that, in the legal codification of these powers and how they can be legitimately used—rulings which have been the basis for normative practice for nearly the last thousand years—the jurists insisted the decision-maker (the "judge" [*shofet/dayan*]) cannot use extraordinary measures in ways that violate human dignity [*kavod ha-briyot*] (Maimonides, 1180/2001, *Hilchot Sanhedrin*, § 24.10; Ben-Asher, 1475/2016, § 2). This is because the individual is of infinite worth, likened by the Talmud to an entire world (*Sanhedrin*, Mishnah 4.5, esp. Maimonides' elucidation; cf. Halbertal, 2015). To violate or degrade humanity in the name of saving human society is therefore illegitimate.

5. Conclusion: Dictator, King, or Judge

Emergency powers make for a timeless (and these days, timely) subject, one that consistently draws interest across the history of political theorizing, in part because it sits uneasily with the values of the rule of law and often risks introducing instability and unpredictability even as it claims to legitimate itself by seeking stability and order. The "war on terror" was just one recent example of how power's response to crisis can put rights and the rule of law at risk, and many have experienced abridgement of their rights under the state responses to the latest global pandemic (Bjørnskov & Voigt, 2022). Emergencies and crises appear to be with us for the foreseeable future, and we must consider them, and the range of tools we might use to address them, as we assess X-risks.

The current conceptual landscape of emergency powers invoked to respond to society-scale existential danger is inspired largely by two institutions with long histories, the Roman dictator and the royal prerogative.[10] Even theories of supreme emergency that don't draw explicitly from that intellectual history are still perfectly willing to sacrifice otherwise fundamental rights to the maximalist claims of the survival of one's community. Cicero's rule, *salus populi suprema lex esto*, while sometimes operating under different names, has retained essentially the same intellectual structure through the history of political and legal thought to the present day. We know that the current theoretical and institutional toolkit that democratic leaders turn to during crisis has a troubling track record: the latest empirical research indicates that "democracies are 75 percent more likely to erode under a state of emergency than without, marking a substantial increase in the probability of democratic decline" (Lührmann & Rooney, 2021, p. 618). One would be justified in questioning whether the theoretical tradition that runs from Rome to the Reich is a dangerous dead-end. Yet all would also admit that we cannot do without sophisticated means of responding to exceptional times of crisis.

This paper has suggested that, to whatever extent the political theorist's metaphor of a dialogue between Jerusalem and Athens has merit, a third model from the tradition of Jerusalem offers a different way of conceiving of legitimate ways to respond to existential risks, even those that may be improbable or remote but threaten to cascade into catastrophe. It is an approach that is sensitive to the wisdom of forecasts and the need to act decisively, yet also functions with built-in protections for the dignity of the human person. As the world confronts new kinds of existential risks, I suggest that this alternative approach to extraordinary powers merits more attention than it has previously received.

[10] The Roman "temporary dictator" model remains influential globally, with some variations. To take just one recent example, El Salvador's government has carried out extensive anti-crime measures under the authority of emergency powers, renewed every month by the legislature, a clear echo of the Roman model. "El Salvador's congress extends anti-gang crackdown" (March 16, 2023). *Associated Press*. See Rossiter, 1963, Ferejohn & Pasquino, 2004 and Gross & Ní Aoláin, 2006. The kingly prerogative, too, remains at the center of ongoing debates about executive powers. See McConnell, 2022; Prakash, 2015; and Nelson, 2014. Compare ICCPR, 1966.

References

Ben-Asher, Y. (2016). *Tur hoshen mishpat*, in Karo, Y., et al. *Tur ve-Shulhan `arukh he-hadash*. Mekhon Shulhan Melakhim. (Original work printed 1475)

Ben-Maimon, M. [Maimonides]. (2001). *Mishneh torah: Yad ha-chazakah*. Hotzaat Shabse Frankel. (Original work published ca. 1180)

Bjørnskov, C., & Voigt, S. (2022). This time is different?—on the use of emergency measures during the Corona pandemic. *European Journal of Law and Economics*, *54*(1), 63–81. https://doi.org/10.1007/s10657-021-09706-5

Blackstone, W. (1765). *Commentaries on the laws of England*. Clarendon Press.

Bleich, J. D., & Jacobson, A. J. (2015). *Jewish law and contemporary issues*. Cambridge University Press.

Broom, H. (1874). *A selection of legal maxims, classified and illustrated* (7th American ed.). T. & J. W. Johnson & Co.

Chajes, Z. H. (1958). *Torat ha-Nevi'im*. In Chajes, Z. H. *Kol sifre Mahari"ts Hayut*. Divrei Hakhamim. (Original work published 1836)

Cicero, M. T. (1928). *De legibus* (trans. Clinton Walker Keyes). William Heineman. Loeb Classical Library. (Original work ca. 58–43 BCE)

Coke, E. (1727). *The reports of Sir Edward Coke Kt. in English, compleat in thirteen parts*. E. & R. Nutt.

Fatovic, C. (2009). *Outside the law: Emergency and executive power*. Johns Hopkins University Press.

Feldman, L. C. (2013). Lockean prerogative: Productive tensions. In C. Fatovic & B. Kleinerman (Eds.), *Extra-legal power and legitimacy: Perspectives on prerogative* (pp. 75–93). Oxford University Press.

Ferejohn, J. (2023). Emergency powers: An introduction. In M. Fiorina (Ed.), *Who governs? Emergency powers in the time of COVID* (pp. 3–32). Hoover Institution Press.

Ferejohn, J., & Pasquino, P. (2004). The law of exception: A typology of emergency powers. *International Journal of Constitutional Law*, *2*(2), 210–39. https://doi.org/10.1093/icon/2.2.210

Gross, O. (2013). Violating divine law: Emergency measures in Jewish law. In C. Fatovic & B. Kleinerman (Eds.), *Extra-legal power and legitimacy: perspectives on prerogative* (pp. 52–74). Oxford University Press.

Gross, O., & Ní Aoláin, F. (2006). *Law in times of crisis: Emergency powers in theory and practice*. Cambridge University Press.

Halbertal, M. (2015). Three concepts of human dignity. Dewey Lecture, University of Chicago Law School. Recording online at https://chicagounbound.uchicago.edu/dewey_lectures/7.

Harrington, J. (1992). *The commonwealth of Oceana*, in J. G. A. Pocock (Ed.), *The commonwealth of oceana and a system of politics*. Cambridge University Press. (Original work published 1656)

Hobbes, T. (1994). *Leviathan*. Hackett Publishing Co. (Original worked published 1651)

International Covenant on Civil and Political Rights [ICCPR]. (1966). Adopted on December 16, 1966 by the United Nations General Assembly, G.A. Res. 2200. Entered into force on March 23, 1976. 999 U.N.T.S. 171, 6 ILM 368.

Kent, J. (1860). *Commentaries on American law* (10th ed.). Brown & Company.

Lazar, N. C. (2013). Prerogative power in Rome. In C. Fatovic & B. Kleinerman (Eds.), *Extra-legal power and legitimacy: Perspectives on prerogative* (pp. 27–51). Oxford University Press.

Locke, J. (1988). *An Essay Concerning the True Original, Extent, and End of Civil Government* [*Second Treatise*]. In P. Laslett (Ed.), *John Locke: Two Treatises of Government*. Cambridge University Press. (Original work published 1689)

Lührmann, A., & Rooney, B. (2021). Autocratization by decree: States of emergency and democratic decline. *Comparative Politics*, *53*(4), 617–35. https://www.jstor.org/stable/27090047

Machiavelli, N. (1950). *Discourses on the first ten books of Titus Livius* (trans. Christian E. Detmold). In M. Lerner (Ed.), *The Prince and the Discourses*. Modern Library. (Original work published 1531)

McConnell, M. W. (2020). *The president who would not be king: Executive power under the constitution*. Princeton University Press.

Mittiga, R. (2022). Political legitimacy, authoritarianism, and climate change. *American Political Science Review*, *116*(3), 998–1011. https://doi.org/10.1017/S0003055421001301

Munir, M. (1957). *Special reference No. 1 of 1955*. In I. Jennings, *Constitutional problems in Pakistan* (pp. 259–349). Cambridge University Press.

Nelson, E. (2014). *The royalist revolution: Monarchy and the American founding*. Harvard Belknap.

Prakash, S. B. (2015). *Imperial from the beginning: The constitution of the original executive*. Yale University Press.

Remer, G. (2008). After Machiavelli and Hobbes: James Harrington's Commonwealth of Israel. In G. Schochet, F. Oz-Salzberger, & M. Jones (Eds.), *Political hebraism: Judaic sources in early modern political thought*. Shalem Press.

Rossiter, C. (1963). *Constitutional dictatorship: Crisis government in the modern democracies*. Harcourt, Brace & World.

Rousseau, J.-J. (1997). *The social contract* (trans. V. Gourevitch). Cambridge University Press. (Original work published 1762)

Statman, D. (2006). Supreme emergencies revisited. *Ethics*, *117*(1), 58–79. https://doi.org/10.1086/508037

Talmud, Babylonian [Talmud]. (2015). Tractate *Gittin*. In R. A. E. Steinsaltz (Ed.), *Koren Talmud Bavli, the Noé edition*. Shefa Foundation Koren Publishers. Online with commentaries (including Rashi): https://shas.alhatorah.org/. (Original work published ca. 500–700)

Walzer, M. (2006). *Just and unjust wars* (4th ed.). Basic Books.

Walzer, M. (2007). Political action: the problem of dirty hands. In D. Miller (Ed.), *Michael Walzer: Thinking politically: Essays in political theory*. Yale University Press.

Yudkowsky, E. (2023, March). Pausing AI developments isn't enough. We need to shut it all down. *Time*.

Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Scenarios 2075: The Cascading Risks Study

Undheim, Trond Arne [1*]

Citation: Undheim, Trond Arne, Scenarios 2075: The Cascading Risks Study. *Proceedings of the Stanford Existential Risks Conference* 2023, 260-279. https://doi.org/10.25740/qf684zr4532

Academic Editor: Paul N. Edwards, Daniel Zimmer

Funding: Open Philanthropy.

Conflict of Interest Statement: The author declares no conflict of interest.

Informed Consent Statement: The study is subject to, and has passed the Stanford Human Subjects (IRB) review, IRB Protocol 68151. No third party material is used, apart from survey responses.

Acknowledgments: Thanks to Dan Zimmer for review of an earlier draft of this article and to Taimur Ahmad for research assistance on statistical analysis and survey presentation.

Abstract: Humanity faces a myriad of existential technology, geopolitical, and ecological risks. When studied separately, one misses the destructive systemic trajectories due to cascading risks. The Cascading Risk Study uses methods of triangulation between board game creation, literature review, scenario planning, quantitative growth indicators projected into 2075, modified Delphi survey, taxonomy development, focus groups and interviews, and case study research to inform the issue. This paper is focused on the five input scenarios created as well as initial findings from an online survey on global systemic risks. The results show that an expert sample (n=145) do not consider 2075 to be a time frame relevant for human extinction, but that they are strongly concerned already about 2225 and beyond. The survey confirms the relevance of the five disruption factors used for scenario building (tech, governance, business, social dynamics, environment), with heightened concern about technology, particularly biotech and AI, and especially the combination of the two. Respondents suggest deep dives in specific application areas of AI, bioengineering, and climate change. Both survey and literature review finds support for taking into consideration a wider array of factors, indicating that even mid-range risks may become systemic.

Keywords: systemic risk, disaster research, existential risk, AI risk, biorisk, scenarios

[1] Research Scholar, Stanford Existential Risk Initiative (SERI), Center for International Security and Cooperation (CISAC), Stanford University, 616 Jane Stanford Way, Encina Hall, Room: C240, Stanford, CA 94305, USA; trondun@stanford.edu

* Correspondence: trondun@stanford.edu

1. Introduction

Systemic risks represent the potential for individual disruptions or failures to cascade into a system-wide failure (Kemp et al., 2022). Severe systemic risks are risks that threaten civilizations (Mayans, British, US), religions (Christianity, Islam, Irreligion, Hinduism), political economic systems (socialism, communism, capitalism, egalitarianism), or geographical societies (regions, nations, counties), and that operate across socio-technical systems (organizations, infrastructures, energy or technology platforms). For example, they are in play when infrastructure is destroyed, and particularly when social collectives collapse, return to a lower level of complexity in terms of energy utilization, social expression, or cultural complexity, or when such entities (civilizations, societies, or systems) are dramatically and perhaps permanently and irreparably harmed.

Related terms such as Global catastrophic risk (GCR), Existential risk, Intersystemic systemic risk, and Disaster risk are all distinct parts of the scientific literature on systemic risk. I note that systemic risk originated in finance circles and was used about financial institutions that were viewed as "too big to fail", but is now used in a much broader context (Mitra & Shaw, 2023; Moch, 2018; Schweizer, Pia-Johanna, 2021; Sillmann, J. , Christensen, I. , Hochrainer-Stigler, S. , Huang-Lachmann, J. , Juhola, S. , Kornhuber, K. , Mahecha, M. , Mechler, R. , Reichstein, M. , Ruane, A. C. , Schweizer, P.- J. & Williams, 2022).

The analysis of systemic risk has already seen positive impact through scientific progress in fields as diverse as finance (reinsurance), engineering (industrial risk), earth science (climate change), and global health (social determinants of health), but is increasingly butting up against three key challenges: increasing societal, industrial, and technological complexity, a deteriorating environment increasingly found to be derived from human interventions and activity (leading to cascading changes), and, correspondingly, an urgent need to communicate across disciplinary boundaries.

An overarching hypothesis became highly relevant for scenario creation. H1 says that even mid-range risks may become systemic, and so might contribute to catastrophic or even existential outcomes, depending on the order and magnitude of the interaction effects between them. Investigating the mechanisms surrounding (H1-related) cascading risks I ask (1) which factors are important to assess and monitor and where do they fit on the historical timeline, (2) which factors contribute the most to severe knock-on effects, (3) how do distinct social groups differ in their views on existential risks.

1.2 Cascading systemic risks and the role of tipping points

Cascading systemic risks, as a subcategory of systemic risks, emerging from disaster science (Alexander & Pescaroli, 2019; Lucas et al., 2018; Mc Gee et al., 2014; Schweizer, Pia-Johanna & Renn, 2019; Zuccaro et al., 2018). Cascading risks represent an ongoing challenge to several established fields within risk science, and across the domains that touch on systemic phenomena, because of the complex ways these risks potentially interact and amplify each other. In this study, I define cascading risks as forming cumulative, co-causal chains of events, resulting in accumulated damage and potential contagion to other, closely coupled systems.

Each cascading factor can be large, medium, or small, but initial size is not necessarily determinant. Cascades may occur in extreme timescales: from fast onset as in market

crashes, to slow onset as in droughts, or a combination of slow and fast as in epidemics. The direction of causality may be hard to ascertain, or the relationship between factors might even be non-linear, meaning there is not a straightforward relationship between an independent variable and a dependent variable (Kravchenko, 2018; Leydesdorff, 1997; Schweizer et al., 2022; Song et al., 2018; van Doorn et al., 2007). For all of these reasons, cascading risks are often easier to describe ex post than ex ante. That's why scenario planning methods are required to try to cope with the complexity.

Tipping points are key linear factors in a causal chain that majorly impacts systemic risk outcomes by helping to create positive or negative momentum. They often constitute the transition points between discontinuous system states, and are thought to cause or accelerate cascades. To take an example from the economy, The influential *Limits to Growth* (1972) study, which sold 30 million copies in 30 languages, said that the world economy will tank by 2100 due to having exceeded population, resource consumption, and pollution thresholds (Aigner-Walder & Döring, 2022; Meadows et al., 1974). In the field of AI, singularity (technological singularity) is a hypothetical future point in time at which technological growth ("superintelligence") becomes uncontrollable and irreversible (Bostrom, 2016; Dilmegani, 2017). In biology, one could consider a synthetic pandemic for which there is no known cure (within a reasonable time scale). In the geopolitics of energy, the traditional tipping point is the hypothesized prospect of a nuclear winter after detonating nuclear weapons (Baum, 2015; Scouras, 2019), although caused by a massive systemic process, not just a single point.

1.3 Scenarios for risks, tipping points and cascades

Why create scenarios? The short answer is to discuss and prepare, not to predict the future. The process of creating scenarios is informative. Debating scenarios exposes biases, weaknesses, and blind spots. That's how they are immediately useful, especially to the extent a scenario process fosters new alliances and partnerships. Having said that, there are examples in the scenario literature that do involve predictive claims. These include: how Royal Dutch/Shell anticipated the drop in oil prices in 1986 (Ogilvy, 2015; Wack, 1984), how Xerox anticipated the convergence of the copier and printer, how American Express foresaw traveler's checks replaced by credit cards, and how the IPCC anticipated climate change and fostered the NetZero by 2040 pledge (De Pryck et al., 2022; Ogilvy, 2015).

The five scenarios I have custom designed for my study so far briefly cover cascading risks (specifically meaning "risks that amplify each other") from *emerging technology* such as general AI, nuclear energy, synthetic biology, molecular-scale manufacturing, and quantum computing as well as from their industrialization, *ecological risks* such as pandemics, biodiversity, and climate change, and *sociopolitical risks* such as geopolitics, organized crime, terrorism, and social movements. Each of these drivers have the potential for both complicating the risk environment and for accelerating innovation to potentially solve for some of the risks.

In scenario development, initial narratives are skewed toward the extreme to stimulate thinking. The most relevant extinction scenarios chosen for this study, based on preliminary research, are: (1) Climate Cataclysm, (2) World War III, (3) Growth and Collapse, (4) Runaway AI, and (5) Synthetic Biology Unleashed In The Wild. They represent commonly held assumptions in the expert community as well as in the population at large about the most significant and already recognized risks the (relatively)

near future holds. That being said, these scenarios are relatively limited in the application of cascades and each includes a tipping point (specifically: a climate event, the outbreak of war, a resource constraint, a technology development, or a human error). These scenarios will be updated, amended, replaced, and filled in with driving forces, data projections, and a complete set of narratives by the end of the project. I expect the next set of scenarios to be significantly more complex although there is also the possibility that I might choose to narrow in on a single scenario identified through the research that demands maximum attention.

Historically, narrative based scenario mapping is typically used (by corporate futurists) for shorter term studies of industrially relevant futures 10-25 years ahead in time and was invented by Shell in the 1970s (Royal Dutch Shell plc, 2012; Wack, 1984). Overall, the literature on scenarios did not contain a shared understanding of what the relevant time scales should be in this case, nor did the literature on human extinction. For example, is 50 years, 250 years, 1000 years, or 10,000 years the relevant time frame for preparing for human extinction? The answer to this question does not simply depend on scenario methodology, but rather on the availability of empirical data on risks, the unsure path of technological innovation (which becomes much harder to predict beyond 50 years), but also on key notions about the meaning of life, the weight of current societal challenges versus emerging or future challenges, and on political outlooks (what society we want to foster). I ended up with a 50-year timeframe mostly for empirical reasons (this is where I have data) and because 50 years is a timeframe where one has a personal stake in what happens.

Note that it is not only the endpoints (e.g. 2075) which is of interest, but also the curves' possible discontinuity, ebbs and flows over the next 50 years given various assumptions. In the following section, this type of data is explored from a systematic angle, and indicating data sources I am already aware of.

2. Methodology

The Cascading Risks Study, which is carried out at Stanford University, uses methods of triangulation between board game creation, literature review, scenario planning, quantitative growth indicators projected into 2075, a modified Delphi survey, taxonomy development, focus groups and interviews, and case study research to inform the issue. This paper will focus on describing the five input scenarios that are available to inspire discussion and give some context as to how they came about, with casual mention of the other approaches.

Scenarios are meant to stimulate the imagination and to foster ideas, strategies, and actions that bring about desired change, or that mitigate or even avoid certain changes and outcomes. Despite a plethora of books and articles on the scenario creation process, the process of identifying drivers of change is highly ad-hoc and typically not based on a systematic theory of societal evolution. Scenario guru Peter Schwartz, who wrote the seminal publication on scenario planning, influencing a generation of corporate futurists, simply says "trust your instinct", and "use a checklist" for factors "driving trends in the macro-environment" (Schwartz, 1996). In later work, this is specified as "Be sure to consider five general categories of forces and trends: social, technological, economic, environmental, and political forces that interact with one another to create complex and interesting plots" (Ogilvy & Schwartz, 2004).

What is often missing from openly available scenario planning studies is a concrete suggestion on which standardized, longitudinal data sets should be used for key drivers such as demographics, environment, economics, and technology. For the input scenarios, my study took inspiration and timelines from the demographic data provided by the United Nations as well as The World Bank economic data and considered various technology development paths in each of the emerging technologies deemed relevant (especially AI, synthetic biology, and energy technologies).

Throughout the study, even if they evolve, each scenario will be expressed both as a narrative in micro form (25-50 words/used for Delphi surveys), short form (250 words/used for focus groups) and in long form (8000 words/book chapter length), book-form sci-fi versions of the scenarios, as well as in various visual expressions (drawings, charts, timelines, and video). To facilitate dissemination, these scenarios will also be available online. Note that the micro scenarios used for the quantitative survey will (largely) stay the same to allow longitudinal data collection.

The overall research approach emerged in a grounded way, iteratively, as challenges arose, and as literature review revealed challenges to be confronted. The use of all of these methods resulted in a long list of factors which were subsequently reduced to core dimensions using method triangulation, in a semi-grounded theory approach (Corbin & Strauss, 2008), except I began with the chosen strategic lens of my core disruptive factors (SciTech, Econ, Soc, Gov), then expanded to ecology (Eco) given the emerging literature on environmental systemic risks.

A scenario encapsulates a given time and place in history, with relevance for key social groups and/or for society as such. A scenario is, in fact, often a complete narrative of an imagined (possible) world or set of circumstances. According to Bostrom, extinction effects are both pan-generational in scope and crushing in severity, causing death or a permanent and drastic reduction of quality of life (Bostrom, 2013). An extinction scenario would describe a process evolving over time as opposed to an event which would drastically curtail humanity's potential for a long time or forever (Bostrom, 2002). I broadly share this definition in which gradual and steadily deteriorating social collapse is approximated to extinction as opposed to thinking of human extinction purely as a biological phenomenon. In the end, the dying out of the human species from “natural” causes such as population decline from sub-replacement fertility or genetic inbreeding, environmental forces (an asteroid impact, large-scale volcanism, or other natural disasters), via alien extermination, or anthropogenic destruction, would each be processes developing over time. Only by viewing such fatalistic eventualities as processes can we take a proactive approach towards changing or interjecting in such trajectories.

A simplified version of systematic literature review (Phillips & Barker, 2021) was carried out given the early state of the systemic risk field. Five input scenarios were created in text and video format (T. A. Undheim, 2023) using scenario planning methodology (Amer et al., 2013; Chermack, 2022; Ramirez & Wilkinson, 2016; Schwartz, 1992; Wack, 1984). Qualitative long interviews (n=200) were conducted (Leong & Tan, 2013; McCracken, 1988) specifically tailored to interviewing elites (Li, 2022; T. Undheim, 2015). Given the early state of cascading systematic risk, expert sampling, a type of non-probability, purposive sampling used to identify participants that fit a particular profile, (Battaglia, 2011) was used. The advantage of this method is that it is inexpensive, convenient, and adaptive, and yields a richness of information and insights (Klar & Leeper, 2019; Palinkas et al., 2015). When mobilized well, from a sufficiently diverse set of backgrounds, experts

help identify critical factors, constraints, and issues in a research area. Experts (n=145) were surveyed using online survey methodology (open from Jan 15, 2023 until April 30, 2023), Delphi style (Beiderbeck et al., 2021), including as a qualitative research tool (Braun et al., 2021). The survey sample consisted of 26.14% PhDs (46), 46.02% Involved with systemic risk, tech, or foresight (81) and the geographical distribution: USA (95), Norway (13), Germany (7), with the gender distribution: 102 males, 42 females, 2 transgender, 2 non-binary, 1 prefer not to say.

2.1 Statistical aggregation

The purpose of the study is to understand more about the risks of human extinction. In the survey, I ask questions on background, risk perception, future risk scenarios, cascading risks, and innovation and risk. A short survey (20-min) has been open from Feb 14 to April 30 2023.

For the strategic sample survey, I was particularly looking for futurists, scientists, activists, technologists, economists, academics, students, executives, and other experts or practitioners in a domain or social movement of relevance to future risks and innovations. I am also interested in getting in touch with any stakeholders who are likely affected by such risks or innovations now or in the future. The target age of respondents is 20+. I did not conduct research on children. The sample for the short survey was n=145. Given that this was an online survey, exact response rates are difficult to ascertain. However, the survey was sent to an initial mailing list of approximately 5000 people. If that number was used, the response rate could be said to be 3.5 percent, which is quite low even for online surveys, although the published literature I know describes education surveys specifically (Wu et al., 2022).

Empirical studies of expert opinions of human extinction are relatively rare (Aiimpacts, 2022; Javeline et al., 2015; Sandberg & Bostrom, 2008; Schubert et al., 2019; Stansberry et al., 2019; Zhang & Mace, 2021) and demand highly specialized interests and knowledge. Those who do give predictions serve up grim scenarios (Bologna & Aquino, 2020; Bostrom, 2002; Buzan, 2003; Ord, 2020; Rees, 2003). Surveys of the general population's view of extinction risk are also rare (Bialik & Orth, 2023; Pauly, 2022). Informal feedback from recipients of my outreach email indicated that recipients wondered if they had the proper insight to respond and whether they truly were the target audience.

The scenarios were meant to use as inputs to the survey instruments. With that in mind, they were published on SERI's YouTube channel as video scenarios in February 2023 (T. A. Undheim, 2022) and were subsequently linked to the Qualtrics survey battery as supplementary audio-visual support to the summary abstracts provided in the survey.

2.2 Cursory historical case studies

Historical events that share some characteristics that are useful to consider for building believable extinction scenarios will be considered as smaller case study vignettes throughout the project and inspired broad causal chains in the core narratives. Examples include the Mayan civilization collapse (850 to 1000 A.D.), the Effects of the Black Death (1347-1353), World War I and II, the Rwanda genocide, and the COVID-19 pandemic (2019-2024). Each of these events have been deeply studied. My cursory analysis will be derived from strategically chosen overview articles that emphasize macro patterns in the

data. To be clear, I do not aim to contribute unique new insights to any of these historical narratives in the present article.

In parallel with building scenarios, a cursory analysis of these historical events was carried out, resulting in fishbone diagrams illustrating disruptive factors at play. Figure 1 is one example of that from a simplified summary of the Mayan civilization collapse (Alatalo, 2017; Diamond et al., 2020; Fernandez-Armesto, 2002; ScienceDaily, 2019; Stromberg, 2012). In-depth analysis of a few chosen historical cases is planned for the popular non-fiction book manuscript being developed, but not for any of the scientific articles.

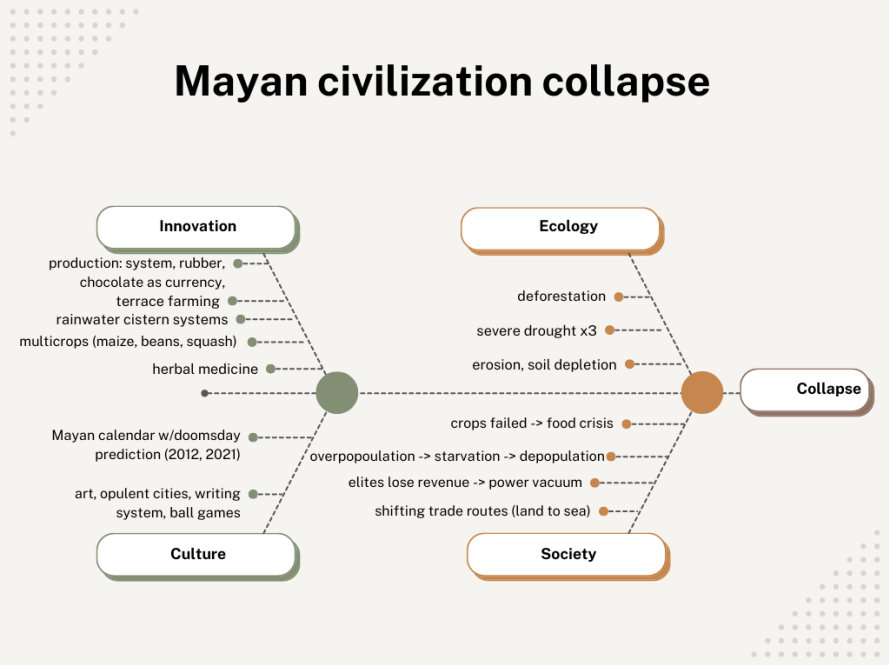


Figure 1 Causes and effects of the collapse of the Mayan civilization

3. Scenarios for future civilizational collapse

The following is the 130-word summary text of the five scenarios (matching the text included in the survey), complemented with a single visualized timeline. The video scenarios are based on the 1000-word version of these same scenarios, and can be consulted by those wanting to view the larger canvas (T. A. Undheim, 2023).

I. *Climate Cataclysm by 2075*

Human-induced climate change from the whole industrial era, starting with industrial activity in the 1800s, and, in particular, since the Great Acceleration of the anthropocene era in the 1950s, have accumulated to produce a mutually reinforcing cocktail of famine, extreme weather, war, and disease, starting to accelerate in the 2050s. Cascading effects of poor land use and primary predator extinction had wiped out 30% of 1850s biodiversity by 2040 and 50% by 2050, and has, since 2065, left the Earth without clean water and lacking in food. As a result, the Earth has already been reduced to 20 percent of the 2025 population, with the population rapidly dwindling, and the planet is on a trajectory that would end humanity within a few decades unless a drastic intervention happens, which does not look likely.

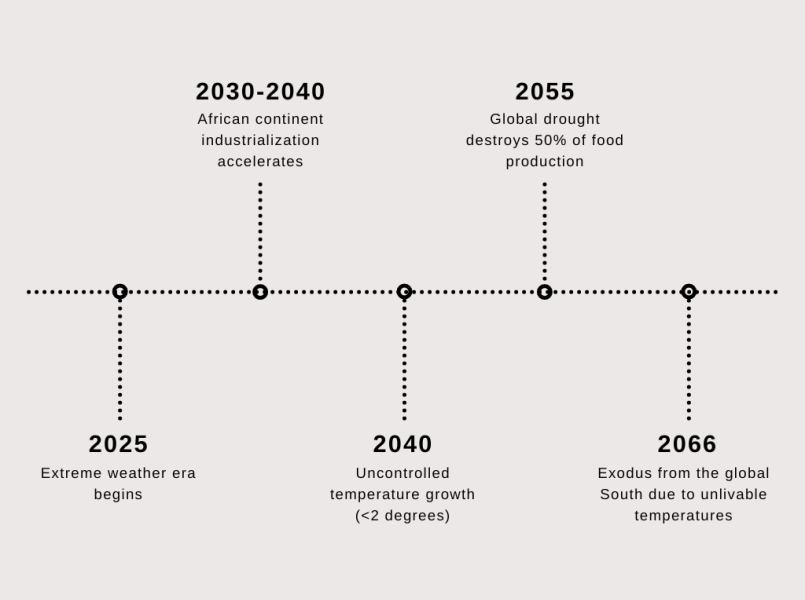


Figure 2 Climate scenario timeline.

II. ***World War by 2075***

The devastating regional war that began in 2055 decimated the world's two most powerful nations, entailed widespread use of inexpensive, widely available weapons of mass destruction, unleashed the biggest standoff ever seen using nuclear warheads, and created detrimental ripple effects across the world within three months, causing global financial collapse, nuclear winter, and near instant, drastic population decline, escalating GDP declines, as well as causing revolutions, fueled by disinformation in several countries. Globalization as a system of trade also became defunct, and with that, other global institutions collapsed, too.

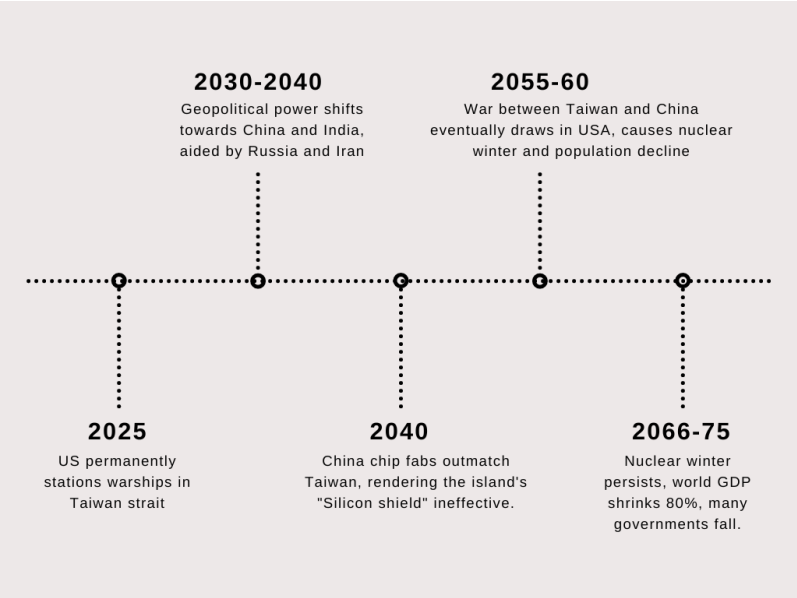


Figure 3 World War III scenario timeline

III. ***Growth and Collapse by 2075***

The build-up of exponential technologies, paired with a plethora of natural and human-made disasters that characterized the last few decades towards 2075 were not all unforeseen, but the exact toxic mix came as a more than unpleasant surprise, given that it

depleted global resources and pitted the world's nations against each other in a drive for ambition, domination, and, ultimately, for survival. The unforeseen chain of events that were precipitated by an initial period of nearly 30 years of exponential growth in technology, finance, and prosperity, ended abruptly in a global contraction caused by a devastating blast from an X factor surrounding a new energy technology, and escalated from there into an extinction event.

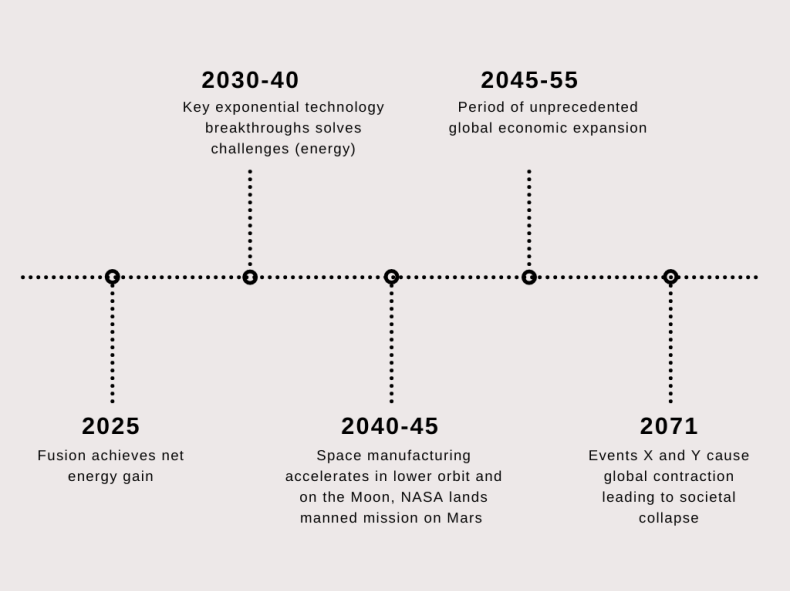


Figure 4 Growth and Collapse timeline

IV. *Runaway AI by 2075*

The AIs that emerged in the first few decades after 2025 didn't have the capacity of general intelligence, and were far from sentient. However, by the turn of 2050, things changed abruptly. Unforeseen changes started to occur, at first amongst the world's top 100 supercomputers which, by 2045 had all been equipped with quantum processors. But it was the alignment of AIs with certain social groups who financed their emergence, and agreed with what I came to understand were the AI's intentions and agenda, that made the runaway phenomenon possible. Enabled by humans, AIs became unstoppable, not alone, but as a hybrid collaboration.

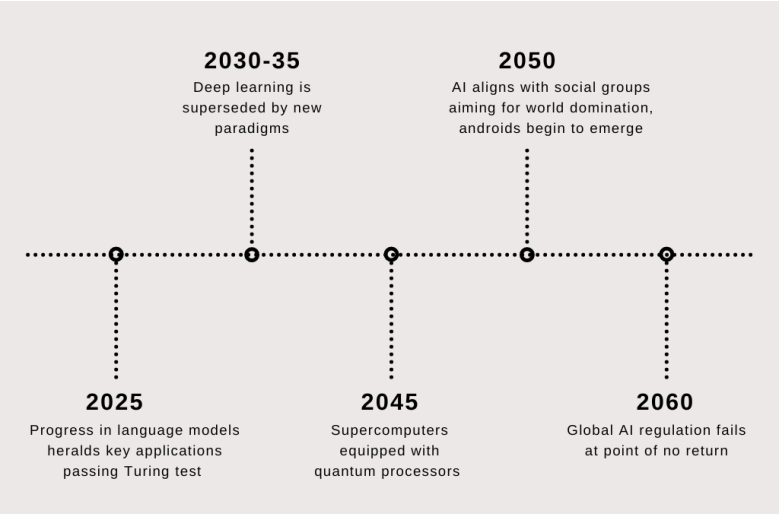


Figure 5 Runaway AI scenario timeline

V. *Synthetic Biology Unleashed In The Wild by 2075*

The synthetic biology breakthrough that fostered the crisis was discovered all the way back in the 2020s. It wasn't the technology as much as the 2047 lab leak and the rapid integration with nature in a vulnerable part of the world that caused Earth's ecology to take a nosedive. The synthetic compounds reacted adversely with photosynthesis and affected drinking water. However, it was the second lab leak, from the Floridian Mars lab in 2067 that accelerated things. The impact of both leaks were initially subtle, and almost untraceable. After a long incubation period, Earth succumbed to human-created synthetic compounds only after failed attempts to decontaminate and isolate the problem regionally.

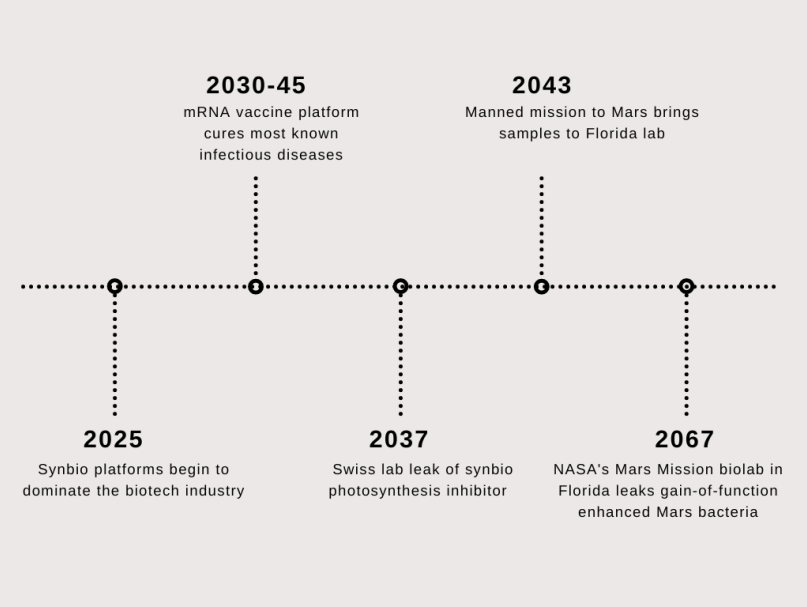


Figure 6 Synthetic Biology Unleashed In The Wild scenario timeline

4. Findings

Scenarios are not standalone, they are meant to stimulate discussion. This article only scratches the surface of the reception of the scenarios. I will highlight a few things.

One might wonder how concerned the experts in my sample are about the potential for a global catastrophic event. Figure 7 (reported in actual number of respondents) captures their level of concern and shows that most were somewhat, moderately or extremely concerned. Only 4 out of 145 experts were not at all concerned. I will try to interview those four if they let me follow up with them.

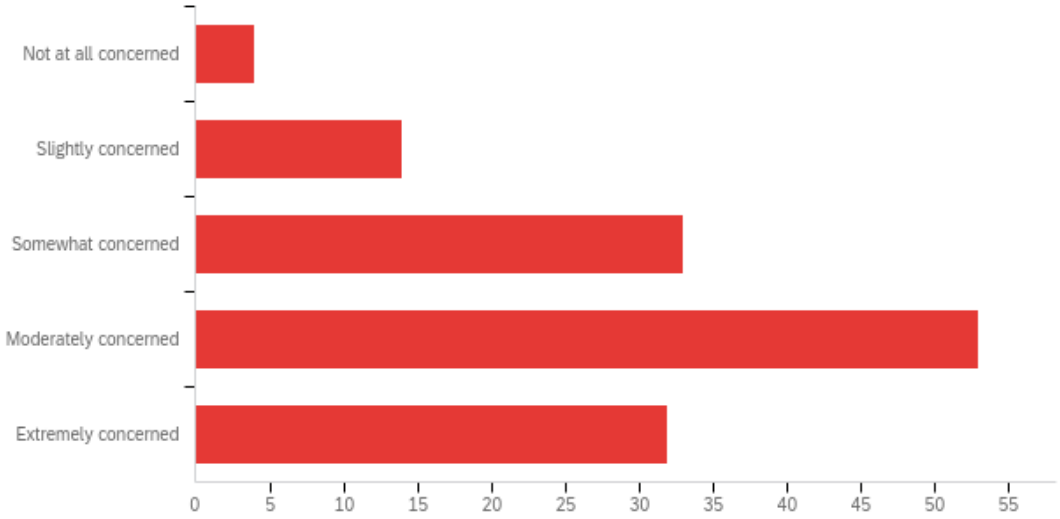


Figure 7. Level of concern about the potential for a global catastrophic event.

A highly relevant aspect to the question of drivers of cascading systemic change is the issue of whether particular factors are viewed as more or less important. I have already primed the scenarios with five disruption factors. As Figure 8 shows, a great proportion of those who answered the question feel that they are each equally important, although technology, social dynamics and ecology are roughly equally important and regulation and business models are viewed as less important.

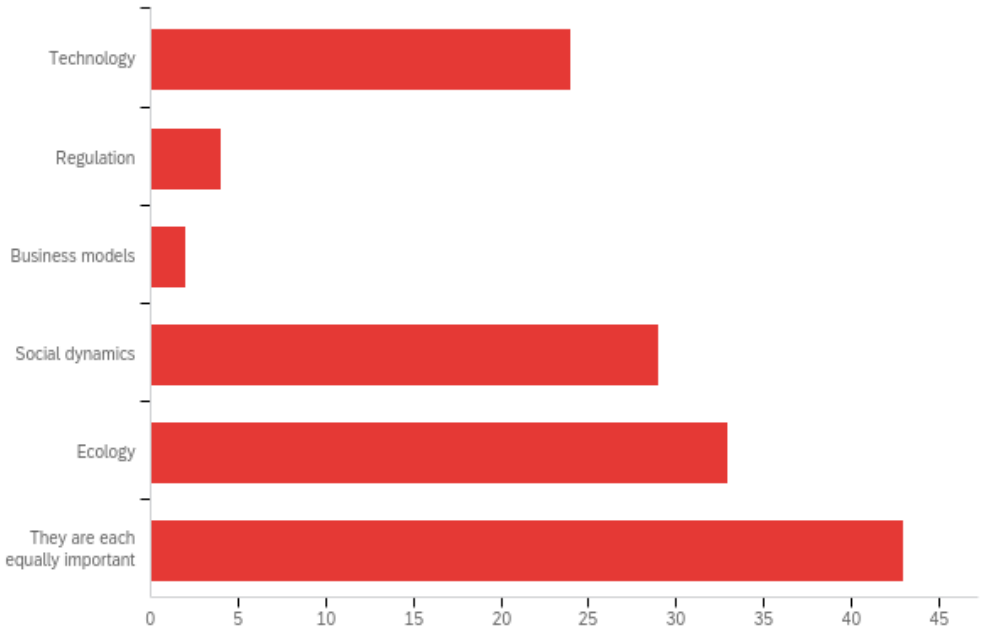


Figure 8 Drivers of cascading systemic change

Ialso asked about which combinations of emerging technologies could prove the most toxic for humanity, should we fail to control them appropriately. As Figure 9 shows, AI + Biotech is the combo most of my experts are the most afraid of, closely followed by Climate change + Biotech (which I had exemplified by geoengineering). The numbers are smaller but my experts seem the least perturbed by quantum computing in combination with other factors, except for in combination with AI.

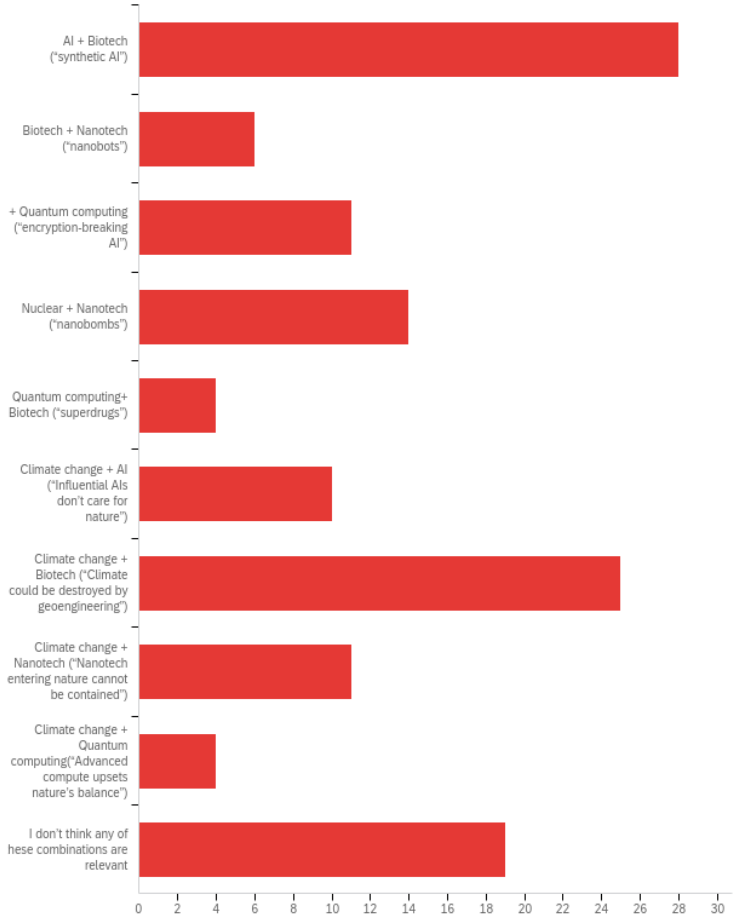


Figure 9 Combinatory risk across risk drivers

Overall, my experts don't seem particularly concerned about the 2075 timeframe of my scenarios. When asked about the likelihood of human extinction by 2075, they unequivocally feel it is unlikely or extremely unlikely, which is not surprising.

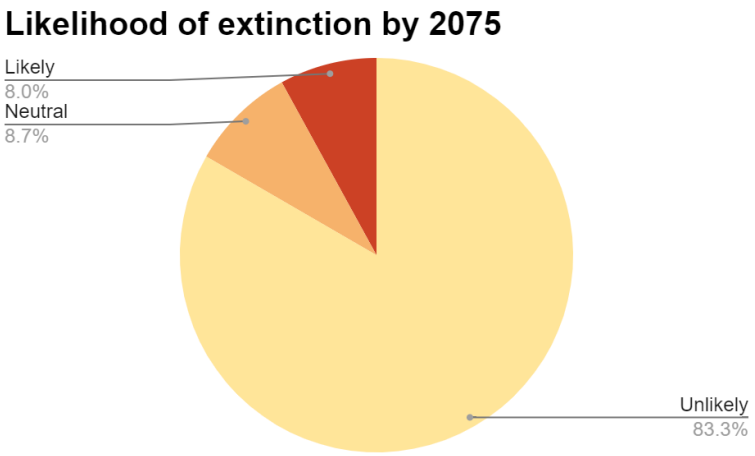


Figure 10 Likelihood of extinction by 2075

In full disclosure, the study's author also does not think this is a likely scenario even though the scenarios were formulated to make that the question for consideration. The present article will not dive into further details on this topic, just to say that further

distinctions exist in the survey between various timeframes, such as 2225 and 2525, 3025 and 12025, and that the numbers change quite dramatically. Many of the experts sampled start considering human extinction likely already by 2225 and extremely likely by 12025. Others, of course, think of the 21st century as a unique existential bottleneck of systemic issues (biodiversity collapse, overpopulation, excessive growth) that can (and must) be resolved so as to avoid unappealing alternatives such as systems collapse, space-migration, or apocalypse (White & Hagens, 2019). This is also the ethos of the longtermist movement (MacAskill, 2022).

The expert and practitioner survey sample is indeed in agreement that if the world is to suffer an existential event in the next 50 years it would not be from a single factor but rather as a result of multiple overlapping, and cascading risks. This does not discount the appearance of a tipping point but it indicates that only looking for a specific tipping point would be a fallacy.

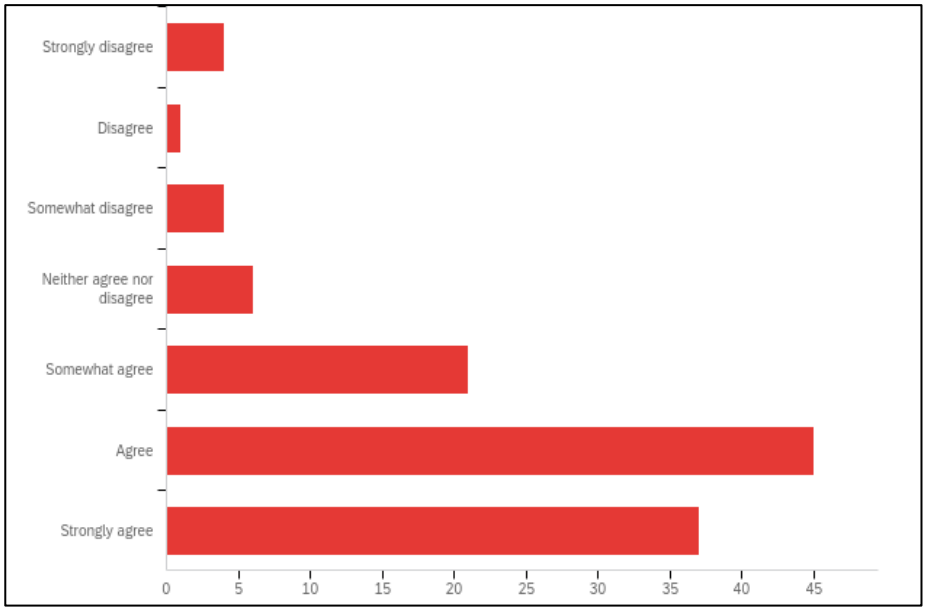


Figure 11 Likelihood that a cascading risk, not a single cause, leads to an existential event in the next 50 years

Respondents are split relatively evenly across the five scenarios when it comes to their relative likelihood (defined as the "single most important scenario"). Shockingly, perhaps, there is the least support for the idea of runaway AI. This could mean I have a skewed sample and need to recruit more AI experts given that previous research plus current zeitgeist indicates AI alignment is a considerable worry among AI experts (Bostrom, 2016; Kirchner et al., 2022). An open letter with 1988 signatories in March 2023 calls for an immediate pause for at least 6 months to the training of AI systems more powerful than GPT-4 (FLI, 2023) gained widespread media attention, including from BBC, CNN, and the Washington Post (Kelly, 2023; Oremus, 2023; Vallance, 2023).

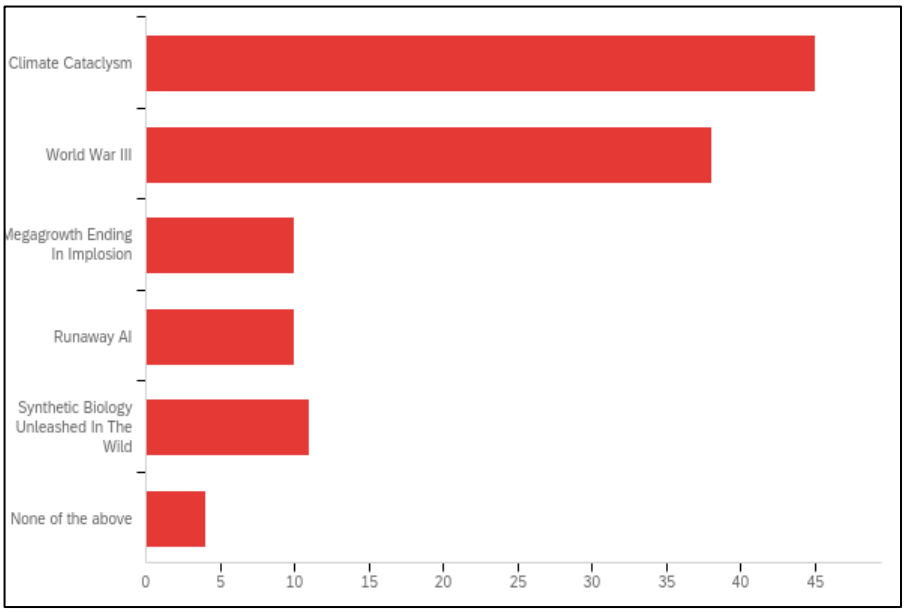


Figure 12 The single most important scenario

As Figure 13 shows, the most compelling data sources among experts are the IPCC scenarios followed by Delphi studies of experts with quantitative datasets in third position. Supercomputing simulations, futurist scenarios and financial data are viewed as less compelling. From that we can learn that even though experts realize the future is uncertain they still tend to cling to numbers when it comes to reasoning around it, which is understandable.

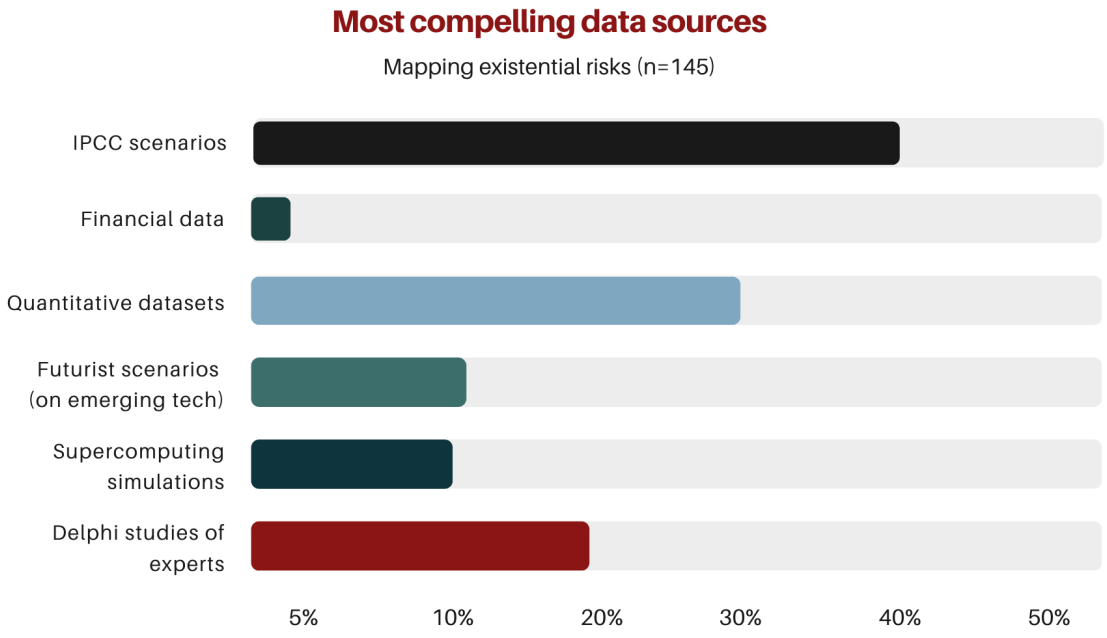


Figure 13 Most compelling data sources on existential risks.

Mitigation was not a big emphasis of the survey or scenarios, but I did get some respondent ideas on the relationship between risk and innovation. The most relevant statements are summarized here:

- Risks from a few specific AI-enabled domains (disinformation, synthetic biology, LAWS) as well as mitigation paths (AI and synbio applied to human disease eradication)
- Climate mitigation paths related to land use (non-animal ag, foodtech, waste reduction or waste management in landfills) and water use (water treatment, desalination)
- Inequality (longevity, social conflict)
- Governance challenges (political education and recruitment, fostering new institutions, group dynamics, and relationships, quintuple innovation helix framework).
- Emerging types of consciousness (degrowth, sustainability, metareflection, longterm(ist) thinking, wisdom)

5. Discussion

Whether these scenarios play out as described is irrelevant. They were never meant to predict the future. That being said, they are also only extreme continuations of existing trends. Despite adding tipping points and singularities in terms of imagined events, my scenarios do not add entirely new interaction patterns. This is consistent with my desire to explore the systemic role of mid-range risks. These might contribute to catastrophic or even existential outcomes, depending on the order and magnitude of the interaction effects between them. Judging from the initial reception among the sample of experts and practitioners, they fear even wider-ranging cascading changes that would represent far more significant discontinuities.

The hardest thing to do when creating scenarios is to plausibly describe cascading risks. I do not feel I fully succeeded with any of these five scenarios. They are each principally focused around one main risk as the tipping point for cascades as opposed to being caused by a plethora of previously unconnected disruption factors. To a slight extent this was by design because I did not want too far out scenarios to distract survey participants from their reflections. On the other hand it is clear from the initial feedback that the scenarios may not have stretched the imagination of the survey respondents to a sufficiently large degree. That has to be addressed in my further work, integrating the scenarios, quantifying them, and creating more extensive cascades.

The scenario work has foregrounded the issue of cascades and has also slightly deemphasized the issue of tipping points as a necessary condition for creating or aggravating cascades. That is not to say that tipping points such as a runaway emerging technology or global financial hardship could not be the straw that breaks the camel's back. Rather, it just points to the fact that it is not the start of the cascade (the "tipping point") that creates the cascading effect but rather the sheer magnitude of the ingredients going into the cascade. For that reason, it would seem that mitigation has to include a way to reduce the overall magnitude of threats and threat factors on a broad scale, not just at the tip of the spear (Kemp et al., 2022; Lenton et al., 2019).

6. Conclusion

The initial feedback on the five input scenarios created for my study indicates support for my core hypothesis (H1) that even mid-range (meso level) risks, initially confined to local geographies, may become systemic, depending on the order and magnitude of the interaction effects between them. Exactly how these mechanisms work is something we will now turn to. Future research needs to examine sociological and location-specific variables, including doing deep dives into the cascading risk topics identified by my work such as risks from specific AI-enabled domains (disinformation, synthetic biology, LAWS

weapons), climate mitigation paths related to land use (non-animal ag, foodtech, waste reduction, or waste management in landfills), water use (water treatment, desalination), inequality (longevity, social conflict), and governance (political education and recruitment, fostering new institutions, group dynamics, and relationships).

My own work is far from done and has far from exhausted the needed adjustments to the research agenda to address the upcoming challenges I have identified. It is not going to be enough to stick to the five relatively obvious scenarios I have depicted so far. Rather, the plausible futures we are going to face will be intricate mixes of all of these as well as full of events and mechanisms about which we are only scratching the surface. Those are exciting prospects for all researchers who are moving into the field of cascading, systemic risk studies.

References

Aigner-Walder, B., & Döring, T. (2022). The Limits to Growth – 50 Years Ago and Today. *Intereconomics*, *2022*(3), 187–191.

Aiimpacts. (2022, August 4). *2022 Expert Survey on Progress in AI (ESPAI).* AI Impacts. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/

Alatalo, E. (2017, May 22). *Rise and fall of ancient Maya water management*. https://www.fieldstudyoftheworld.com/rise-fall-ancient-maya-water-management/

Alexander, D., & Pescaroli, G. (2019). What are cascading disasters? *UCL Open Environment*, *1*(1). https://doi.org/10.14324/111.444/ucloe.000003

Amer, M., Daim, T. U., & Jetter, A. (2013). A review of scenario planning. *Futures*, *46*, 23–40.

Battaglia, M. P. (2011). Purposive Sample. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (pp. 645–647). Sage Publications, Inc.

Baum, S. (2015, May 29). *The Risk of Nuclear Winter*. Federation Of American Scientists. https://fas.org/pir-pubs/risk-nuclear-winter/

Beiderbeck, D., Frevel, N., von der Gracht, H. A., Schmidt, S. L., & Schweitzer, V. M. (2021). Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements. *MethodsX*, *8*, 101401.

Bialik, C., & Orth, T. (2023, April 14). *AI doomsday worries many Americans. So does apocalypse from climate change, nukes, war, and more*. YouGov. https://today.yougov.com/topics/technology/articles-reports/2023/04/14/ai-nuclear-weapons-world-war-humanity-poll

Bologna, M., & Aquino, G. (2020). Deforestation and world population sustainability: a quantitative analysis. *Scientific Reports*, *10*(1), 7631.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology / WTA*, *9*. https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c

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

Bostrom, N. (2016). *Superintelligence: Paths, Dangers, Strategies* (Reprint edition). Oxford University Press.

Braun, V., Clarke, V., Boulton, E., Davey, L., & McEvoy, C. (2021). The online survey as a qualitative research tool. *International Journal of Social Research Methodology*, *24*(6), 641–654.

Buzan, B. (2003). Our Final Century. Martin Rees. London: William Heinemann, 2003. 228 pp. 17.99. In *Survival* (Vol. 45, Issue 4, pp. 222–223). https://doi.org/10.1093/survival/45.4.222

Chermack, T. J. (2022). *Using Scenarios: Scenario Planning for Improving Organizations*. Berrett-Koehler Publishers.

Corbin, J., & Strauss, A. (2008). *Strategies for qualitative data analysis. Basics of Qualitative Research. Techniques and Procedures for Developing Grounded Theory, 3 (10), 4135*.

De Pryck, K., Hulme, M., Skodvin, T., Leclerc, O., Hartz, F., Livingston, J. E., Beck, S., Siebenhüner, B., Standring, A., Gustafsson, K. M., Hughes, H., Yamineva, Y., Edwards, P. N., Petersen, A. C., van Bavel, B., MacDonald, J. P., Dorough, D. S., Guillemot, H., Cointe, B., & Asayama, S. (2022). *A Critical Assessment of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Diamond, J., and Steel By Jared Diamond Ph. D. Guns, G., 978-, 9780393354324, Diamond, C. B. J., 978-, 9780143117001, Diamond, U. B. J., & 978-, 9780316409148. (2020). *Jared Diamond 3 Books Collection Set(Upheaval, Collapse Guns, Germs and Steel)*. Penguin ptd.

Dilmegani, C. (2017, August 8). *When will singularity happen? 995 experts' opinions on AGI*. AIMultiple. https://research.aimultiple.com/artificial-general-intelligence-singularity-timing/

Fernandez-Armesto, F. (2002). *Civilizations: Culture, Ambition, and the Transformation of Nature* (Reprint edition). Free Press.

FLI. (2023, March 22). *Pause Giant AI Experiments: An Open Letter*. Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Javeline, D., Hellmann, J. J., McLachlan, J. S., Sax, D. F., Schwartz, M. W., & Castro Cornejo, R. (2015). Expert opinion on extinction risk and climate change adaptation for biodiversity. *Elementa (Washington, D.C.)*, *3*, 000057.

Kelly, S. (2023, March 29). Elon Musk and other tech leaders call for pause in "out of control" AI race. *CNN*. https://www.cnn.com/2023/03/29/tech/ai-letter-elon-musk-tech-leaders/index.html

Kemp, L., Xu, C., Depledge, J., Ebi, K. L., Gibbins, G., Kohler, T. A., Rockström, J., Scheffer, M., Schellnhuber, H. J., Steffen, W., & Lenton, T. M. (2022). Climate Endgame: Exploring catastrophic climate change scenarios. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(34), e2108146119.

Kirchner, J. H., Smith, L., Thibodeau, J., McDonell, K., & Reynolds, L. (2022). Researching Alignment Research: Unsupervised Analysis. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2206.02841

Klar, S., & Leeper, T. J. (2019). Identities and intersectionality: A case for purposive sampling in survey-experimental research. In *Experimental Methods in Survey Research* (pp. 419–433). Wiley. https://doi.org/10.1002/9781119083771.ch21

Kravchenko, S. (2018). The becoming of non-linear knowledge: New risks, vulnerabilities, and hopes. *Montenegrin Journal of Economics*, *14*(4), 191–202.

Lenton, T. M., Rockström, J., Gaffney, O., Rahmstorf, S., Richardson, K., Steffen, W., & Schellnhuber, H. J. (2019). Climate tipping points - too risky to bet against. *Nature*, *575*(7784), 592–595.

Leong, P. T. M., & Tan, F. B. (2013). Narrative Interviews: An Alternative Method to the Study of Mentoring Adoption by Information Systems Project Managers. *Procedia Technology*, *9*, 638–645.

Leydesdorff, L. (1997). The Non-linear Dynamics of Sociological Reflections. *International Sociology: Journal of the International Sociological Association*, *12*(1), 25–45.

Li, L. (2022). How to tackle variations in elite interviews: Access, strategies, and power dynamics. *Qualitative Research: QR*, *22*(6), 846–861.

Lucas, K., Renn, O., Jaeger, C., & Yang, S. (2018). Systemic Risks: A Homomorphic Approach on the Basis of Complexity Science. *International Journal of Disaster Risk Science*, *9*(3), 292–305.

MacAskill, W. (2022). *What We Owe the Future*. Basic Books.

McCracken, G. (1988). *The Long Interview (Qualitative Research Methods Book 13)* (1st ed.). SAGE Publications, Inc.

Mc Gee, S., Frittman, J., Ahn, S. "james," & Murray, S. (2014). RISK RELATIONSHIPS AND CASCADING EFFECTS in CRITICAL INFRASTRUCTURES: IMPLICATIONS FOR THE HYOGO FRAMEWORK. *Global Assessment Report on Disaster Risk Reduction*. https://www.preventionweb.net/english/hyogo/gar/2015/en/bgdocs/McGee%20et%20al.,%202014.pdf

Meadows, D. H., Meadows, D. L., Randers, J., & Behrens, W. W., III. (1974). *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind* (1st ed.). Universe Books.

Mitra, A., & Shaw, R. (2023). Systemic risk from a disaster management perspective: A review of current research. *Environmental Science & Policy*, *140*, 122–133.

Moch, N. (2018). The Contribution of Large Banking Institutions to Systemic Risk: What Do We Know? A Literature Review. *Review of Economics of the Household*, *69*(3), 231–257.

Ogilvy, J. (2015). Scenario Planning and Strategic Forecasting. *Forbes Magazine*. https://www.forbes.com/sites/stratfor/2015/01/08/scenario-planning-and-strategic-forecasting/?sh=2aea1844411a

Ogilvy, J., & Schwartz, P. (2004). *Plotting your scenarios*. University of Toronto. http://choo.ischool.utoronto.ca/fis/courses/inf1005/ogilvy.pdf

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.

Oremus, W. (2023, April 4). The AI backlash is here. It's focused on the wrong things. *The Washington Post*. https://www.washingtonpost.com/technology/2023/04/04/musk-ai-letter-pause-robots-jobs/

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Administration and Policy in Mental Health*, *42*(5), 533–544.

Pauly, L. (2022, February 16). *Has the pandemic changed views on human extinction?* YouGov. https://yougov.co.uk/topics/politics/articles-reports/2022/02/16/has-pandemic-changed-views-human-extinction

Phillips, V., & Barker, E. (2021). Systematic reviews: Structure, form and content. *Journal of Perioperative Practice*, *31*(9), 349–353.

Ramirez, R., & Wilkinson, A. (2016). *Strategic Reframing: The Oxford Scenario Planning Approach* (1st ed.). Oxford University Press.

Rees, M. (2003). *Our Final Century? : Will the Human Race Survive the Twenty-First Century?* Heinemann.

Royal Dutch Shell plc. (2012, November 18). *Shell Celebrates 40 Years of Scenarios*. PR Newswire. https://www.prnewswire.com/news-releases/shell-celebrates-40-years-of-scenarios-179879331.html

Sandberg, A., & Bostrom, N. (2008). Global catastrophic risks survey. *FHI Technical Report*, *1*, 1–5.

Schubert, S., Caviola, L., & Faber, N. S. (2019). The Psychology of Existential Risk: Moral Judgments about Human Extinction. *Scientific Reports*, *9*(1), 15100.

Schwartz, P. (1992). Composing a plot for your scenario. *Planning Review*, *20*(3), 4–46.

Schwartz, P. (1996). *The Art of the Long View: Planning for the Future in an Uncertain World* (Reprint edition). Currency.

Schweizer, Pia-Johanna. (2021). Systemic risks – concepts and challenges for risk governance. *Journal of Risk Research*, *24*(1), 78–93.

Schweizer, Pia-Johanna, & Renn, O. (2019). Governance of systemic risks for disaster prevention and mitigation. *Disaster Prevention and Management; Bradford*, *28*(6), 862–874.

Schweizer, P.-J., Goble, R., & Renn, O. (2022). Social Perception of Systemic Risks. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, *42*(7), 1455–1471.

ScienceDaily. (2019, August 5). Maya more warlike than previously thought: Evidence of extreme warfare from Classic period disputes role of violence in civilization's decline. *Science Daily*. https://www.sciencedaily.com/releases/2019/08/190805143527.htm

Scouras, J. (2019). Nuclear War as a Global Catastrophic Risk. *Journal of Benefit-Cost Analysis*, *10*(2), 274–295.

Sillmann, J. , Christensen, I. , Hochrainer-Stigler, S. , Huang-Lachmann, J. , Juhola, S. , Kornhuber, K. , Mahecha, M. , Mechler, R. , Reichstein, M. , Ruane, A. C. , Schweizer, P.- J., & Williams, S. (2022). *Briefing note on systemic risk*. UNDRR. https://www.undrr.org/publication/briefing-note-systemic-risk?_gl=1*mvi12p*_ga*MjA4NzQzOTU4Ni4xNjc2NTc0NDQz*_ga_T3RWEE6Z0J*MTY3ODEyMjEyMS40LjEuMTY3ODEyMjc0MC4wLjAuMA..

Song, J. W., Ko, B., & Chang, W. (2018). Analyzing systemic risk using non-linear marginal expected shortfall and its minimum spanning tree. *Physica A: Statistical Mechanics and Its Applications*, *491*, 289–304.

Stansberry, K., Anderson, J., & Rainie, L. (2019, October 28). *3. Humanity is at a precipice; its future is at stake*. https://www.pewresearch.org/internet/2019/10/28/3-humanity-is-at-a-precipice-its-future-is-at-stake/

Stromberg, J. (2012, August 23). *Why Did the Mayan Civilization Collapse? A New Study Points to Deforestation and Climate Change*. https://www.smithsonianmag.com/science-nature/why-did-the-mayan-civilization-collapse-a-new-study-points-to-deforestation-and-climate-change-30863026/

Undheim, T. (2015). Getting connected: How sociologists can access the high tech Élite. *The Qualitative Report*. https://doi.org/10.46743/2160-3715/2003.1902

Undheim, T. A. (2022, December 30). *The Stanford Global Systemic Risk Scenarios Study*. Existential Risks Initiative. https://seri.stanford.edu/research/stanford-global-systemic-risk-scenarios-study

Undheim, T. A. (2023). *Extinction Scenarios for 2075: Videos and Narratives*. Stanford SERI. https://seri.stanford.edu/research/stanford-global-systemic-risk-scenarios-study/extinction-scenarios-2075-videos-and

Vallance, C. (2023, March 29). Elon Musk among experts urging a halt to AI training. *BBC*. https://www.bbc.com/news/technology-65110030

van Doorn, J., Verhoef, P. C., & Bijmolt, T. H. A. (2007). The importance of non-linear relationships between attitude and behaviour in policy research. *Journal of Consumer Policy*, *30*(2), 75–90.

Wack, P. A. (1984). *Learning to Design Planning Scenarios: The Experience of Royal Dutch Shell*. Division of Research, Harvard Business School.

White, D. J., & Hagens, N. J. (2019). *The Bottlenecks of the 21st Century: Essays on the Systems Synthesis of the Human Predicament*. Independently published.

Wu, M.-J., Zhao, K., & Fils-Aime, F. (2022). Response rates of online surveys in published research: A meta-analysis. *Computers in Human Behavior Reports*, *7*, 100206.

Zhang, H., & Mace, R. (2021). Cultural extinction in evolutionary perspective. *Evolutionary Human Sciences*, *3*, e30.

Zuccaro, G., De Gregorio, D., & Leone, M. F. (2018). Theoretical model for cascading effects analyses. *International Journal of Disaster Risk Reduction*, *30*, 199–215.

Conclusion

Stanford
Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Conclusion

The Emergence of a Cascading X-Risks Paradigm Steeped in Transdisciplinarity

Trond Arne Undheim [1*]

Citation: Author Undheim, T.A. Conclusion: The Emergence of a Cascading X-Risks Paradigm Steeped in Transdisciplinarity. *Intersections, Reinforcements, Cascades: Proceedings of the Stanford Existential Risks Conference* 2023, 281-291. https://doi.org/10.25740/gy439pz0808

Academic Editor: Paul N. Edwards, Daniel Zimmer

Funding: This work was partially funded by Open Philanthropy

Conflict of Interest Statement: The author declares no conflict of interest.

Informed Consent Statement: All authors included in these proceedings gave their explicit consent to being featured.

Acknowledgements: We thank the rest of the members of the organizing committee of the 2023 Stanford Existential Risks Conference (Steve Luby, Paul N. Edwards, Victor Warlop, Gabe Mukobi, Camille Walker) for their support.

Author Contributions: The authors wrote this article with minor comments from Dan Zimmer and Paul N. Edwards..

Abstract: Existential risks (X-risks) present humankind with an ever more pressing set of challenges. As X-risk studies broadens beyond philosophy, becoming integral to fields such as social science, public health, engineering, and the sciences, cascading X-risks come into focus. Cascading risks, particularly potential interactions among acknowledged X-risks imply that cascades of "merely" catastrophic risks could combine to jeopardize the foundations of human society. The conclusion I draw from the 2023 SERI X-risks conference is that we can and must begin to describe the essential features of such a paradigm. In so doing, I position risk as constitutive to society—with inequality, transdisciplinarity, and governance emerging as key challenges. A transdisciplinary lens creates opportunities to achieve real-world impact at scale. Yet, how to de-risk the evolution of such cascading potential is not yet clear. A formidable challenge arises: regular political discourse may not possess the tools, processes, or frameworks to accomodate the new framing. A lingering observation, in need of validation, is whether cascading X-risks present governance challenges that constitute a more wicked problem than understanding the potential of the emerging technologies that, to a large extent, created the challenges in the first place.

Keywords: existential risk, AI risk, biorisk, planetary boundaries, policy, scenarios

[1] Research Scholar, Stanford Existential Risk Initiative (SERI), Center for International Security and Cooperation (CISAC), Stanford University, 616 Jane Stanford Way, Encina Hall, Room: C240, Stanford, CA 94305, USA; trondun@stanford.edu.

* Correspondence: trondun@stanford.edu

1. Introduction

Existing approaches to existential risks will not be sufficient to avert either catastrophic or existential risk episodes in the coming years. There remains an urgent need to address the continuing stresses on humanity's support systems, the frailty of existing governance structures, and the weaknesses and blind spots in current existential risk monitoring, research, and mitigation (Sillmann, J. , Christensen, I. , Hochrainer-Stigler, S. , Huang-Lachmann, J. , Juhola, S. , Kornhuber, K. , Mahecha, M. , Mechler, R. , Reichstein, M. , Ruane, A. C. , Schweizer, P.- J. et al., 2022).

Many open questions remain, particularly around cascading risks or "polycrises." A polycrisis occurs when crises in multiple global systems become "causally entangled in ways that significantly degrade humanity's prospects" to generate "interacting crises" that result in harms greater than the sum of those the crises would produce in isolation." (Homer-Dixon et al., 2022). However, even if you agree that the current polycrisis (Lawrence et al., 2022) calls for a different approach: how can we convince the current cadre of political leadership of the relevance of cascading risk? The challenge is complicated by the fact that once the discussion moves beyond a single-risk focus, the complexity increases to the degree that standard political discourse lacks the tools, processes, or frameworks to handle the new framing. Politics always requires prioritizing multiple challenges, but what if many causally interrelated challenges require simultaneous prioritization (albeit in different ways)? A cascading approach is likely to prove perplexing and potentially vexing to mainstream political mindsets. This might mean that the system, and the actors within it, may need to change to accommodate the new lens. That type of structural change is painful, costly, and complex to achieve.

One place to start might be to get a better handle on the knock-on effects (e.g. 'causal chains') that can be observed between different categories of risk throughout society. For that to happen, each needs to be discovered in time to see it unfold and then carefully monitored. At this time, the monitoring tools are limited, but that is mostly a question of priorities. However, given the amount of factors to take into account, the challenge once we start monitoring will be to separate the noise from the signal. What are the thresholds for the number and scale of risks society can handle? Is it possible to reduce the number of risks to worry about to a handful or is that itself a too risky approach to the future?

The chief reason cascading X-risks amounts to a compelling research agenda in the years ahead is that cascading risks call for an increasingly wide-ranging analytical approach, using state of the art tools such as emerging productivity tools, new governance principles, and globally integrated research collaboration that builds a transdisciplinary approach transcending many scientific disciplines and institutions. Emerging approaches need to push beyond the triple helix paradigm of collaboration between academia, industry and government (Etzkowitz, 2008) because what's needed is not simply innovation, but succeeding with the substantially more demanding tasks of identifying, building consensus around, and implementing sustainable systems changes. This may even entail dramatic changes that hit on the core of democracy, legitimacy, and the willingness to be exposed to, and react to, risk.

The papers in these proceedings address a wide range of topics that both touch on identifying and mitigating the potential of cascading risks becoming catastrophic or, ultimately, existential for humanity. Next, I will summarize the lessons learned from these papers.

2. Emerging research threads in recent X-risk scholarship

As we have seen throughout these proceedings, a plethora of research threads are emerging in the field of X-risk. Broadly speaking, the topics covered include risk agency, AI risks, biorisks, cascading planetary boundaries worsened by the prospect of nuclear winter, collective intelligence, crisis governance, dual-use, policy concerns, psychological impact, security, and scenario planning.

The starting point for contemporary exploration of X-risk is the psychological impact of the challenging new environment humanity has implanted itself into. As global industrialized society reaches its planetary limits, there is sufficient evidence that the resistance is felt not just at a political level, but also at the personal and psychological levels. The wider impact of X-risk has ramifications across society, and as a matter of course, has become a subject matter of interest for all sciences. Even the aesthetic impact becomes a factor to contend with, both as a protective factor, and as a rationale for action. For example, Raval in these proceedings claims that humanity's predominantly positive reaction to Yosemite Valley—as an aesthetic experience provided by nature—points to the good in the world and becomes both a rationale to keep going, and a way to inspire us to do just that.

The planetary boundaries perspective thrusts the X-risk debate far beyond the limited lens of climate change and into a wider ecological impact discussion that even starts to implicate established threat vectors such as nuclear weapons. The interrelationships between environmental challenges and single-risk threats are multiple, and fear inducing, not just because of their interrelationships but also because ecological X-risks threaten to take our eyes (and our monetary resources) off the ball, so to speak.[1] While distracted with such challenges, one or several other X-risks might blow up, or vice versa.

That's why the emerging threat of contemporary AI-systems is such a formidable challenge. It is not simply the fear that AI as such might start to take over powerful functions of society such as complex systems of governance, military, organized crime, finance, urban planning, health, transportation, or even our lifestyle choices, but the fact that the specter of AI risk might become an all encompassing fear that captures the public attention near exclusively. Adding to that, the reality is that digitalization is now so embedded in contemporary society that even without further evolution of AI/ML technology, automation in itself becomes self-propelling because it is simultaneously nearly invisible day-to-day and also such a natural choice when considering implementing further efficiency mechanisms. The counterpoint to that is the very real notion that augmenting human intelligence is a far more advanced industrial strategy than that of assuming we can (or should) continue automating human labor away (Linder et al., 2022).

There are at least three established, yet fast evolving risk areas of massive concern to X-risk: AI risk, Biorisk, and Nuclear risk, and an emerging one focused around Planetary

[1] By this, I mean that climate X-risks (and mitigations) are so expensive and all-consuming that the world's leaders (and public imagination) are likely to be so occupied with it in the coming years that they may have a hard time prioritizing action on other threats that seem more remote even if they are serious risks. Having failed to sufficiently prioritize mitigating global warming in advance, we now face the opposite risk of belatedly *over* emphasizing today's "climate emergency" to the detriment of other sources of looming catastrophe.

Boundaries. Among those, nuclear risk is the oldest (Ackerman et al., 2008; S. Baum, 2015; Seth Baum, 2017; Bostrom et al., 2011; Garrick, 2004; Onderco et al., 2021; Ord, 2020; Scouras, 2019), but given the way these risks increasingly are interrelated, there is no way to study them in separation. Indeed, as I point out in these proceedings, the combinatory risks of AI + Biorisk is high on the list of expert worries. Moreover, as Florian Jehn also underlines in this volume, there are compounding effects of nuclear winter and overstepped planetary boundaries, given that biosphere integrity affects nuclear winter survivability and, perplexingly, excess nitrogen (an environmental nuisance short term) might be a protective factor during a nuclear winter. Several authors in these proceedings are worried that our society has not fully taken on board the necessary mitigation paths from the dual-risk challenges that stem from misuse of emerging technologies in isolation or in combination.

What are the tradeoffs of a cascading risks approach? Focusing on cascades means privileging intersections and potentially lacking depth within each risk area. To counter that risk, scholars or policy makers with a cascading lens need to be in close touch with single X-risk experts to avoid missing key developments. Prioritizing cascading effects mitigation might erode focus on single X-risk areas of truly established threat levels such as nuclear weapons and biorisk. To counter such erosion, a minimum threshold of mitigation activity needs to be maintained in each of those two areas no matter what other cascading effects might seem relevant. Lastly, cascading risks is a lens steeped in complexity theory, which is a field with poor visibility outside the specialist sphere. To counter the opacity of such a lens, scholars need to redouble their efforts at creating pidgin language that bridges the gap between lofty statements and observations about systems complexity and the curse of short term political agendas and that attempt to meet laypersons motivations and everyday reality halfway.

To be clear: there is nothing wrong with pursuing single risk approaches, especially if it is found to be true that they are the most likely sources of existential risk. But it is not enough. Surprises such as the belated discovery of nuclear winter, uncertainty about the cumulative effects of environmental toxins, and the unpredictable social dynamics of technologies adopted at scale give ground to worry that the current list of leading X-risks may be incomplete. What if experts in this area have misjudged where to watch? Then we would have missed the opportunity to prepare for other, potentially worse calamities. It also stands to reason that if risks are increasingly occurring in cascades, these cascades cannot be stopped by only looking at individual tipping points, even if such points could be identified ex post (or even ex ante). Tipping points are often overly simplified models for what is actually happening. They reflect linear thinking, and often single factor causal thinking. The current obsession over the accumulation of carbon in the atmosphere (worthy of concern but not as the only concern) is only one example of such a simplification among many others.

Relating risks to each other is going to be a significant, emerging task both for researchers and policymakers. This raises an important need for generalists who have devoted sufficient time to understanding the key dynamics of several risks while resisting the temptation to silo themselves in a specific area of expertise—where academic incentives favor narrow pedantry and gate keeping over cross-disciplinary collaboration. Researchers have few paths to pursue the even more challenging practice of trans-disciplinarity (Leahey, 2018; Rigolot, 2020; Undheim, 2021). In effect, transdisciplinarity—which involves bringing non-academics into the sphere of influence and where axioms are shared across domains, not just traded between them (Undheim, 2021)—offers a way

of being more than a way of knowing (Rigolot, 2020). There are also cognitive challenges because scientific practices, including concepts and methods, are highly domain specific (MacLeod, 2018).

Future risks are sure to be modified by the emerging technologies available to humanity. However, the relationship between innovation and risk is complex. Which emerging technologies, innovations, or inventions would change the risk and innovation trajectory the most? Some obvious candidates at this time include AI, synthetic biology, fusion energy, but there are many others, including nanotechnology and brain-computer interfaces (BCI). All of these will likely have both positive and negative implications on cascading X-risks, of unknown magnitude and in timelines still poorly understood or at least poorly manageable within current governance approaches. These impacts will go far beyond currently recognized dual risks as described in the papers in these proceedings.

Let's now look at what this emerging paradigm of cascading X-risks might look like.

3. The cascading X-risks paradigm

While it is clearly very early days for a new paradigm of cascading X-risks, and the field and its constituent elements are each evolving fast, a few points can already be made quite succinctly.

The field does not start from scratch. There is a weighty evidence base to build on from established fields such as the history of ideas, disaster research (Alexander, 2018; Caldera et al., 2022; Naqvi et al., 2021; Rodríguez et al., 2017; Sillmann, J. , Christensen, I. , Hochrainer-Stigler, S. , Huang-Lachmann, J. , Juhola, S. , Kornhuber, K. , Mahecha, M. , Mechler, R. , Reichstein, M. , Ruane, A. C. , Schweizer, P.- J. et al., 2022; Stallings, 2003), the sociology of risk (Battistelli et al., 2019; Beck, 2013; Centeno et al., 2015; Ekberg, 2007; Zinn, 2017), international security (Compel et al., 2023; Dijkstra et al., 2018; InfluenceWatch, 2017; Petersen, 2012), philosophy (MacAskill, 2022), and more.

There remains much to be done in analyzing the rich history of engagement with end times, with the prospect of the apocalypse, or indeed with appeals to the heavens as a possible solution. Billions of the world's faithful belong to religions with strong apocalyptic traditions, and these claims and expectations may prove to be one of the chief rivals to the advice of X-risk studies scholars during times of heightened existential anxiety. Historical and interpretive research on our cultural and political traditions, and the cascading effects of civilizational choices will need to be continuously reassessed. There is no room for error here. Forgetting our past, or refusing to quarrel about it, will lead to simplistic, technocratic, or indeed naive enthusiasm for the present or the future to the detriment of our rich and complex past. The cascading frame reminds us that there is no future without the past, and there is no present without being aware of the pathways of the past or seeking and charting new visions for the future.

Emerging technologies will continue to push the boundaries of the subject of X-risk into new domains, new possibilities, and new interrelationships that present opportunities for cascading effects. The novel contribution of the cascading paradigm is to foreground not merely the collisions, but also the missed but potential connections *between technologies* and *between technology and society*.

Governance challenges can be expected to become a prominent vein of activity of X-risk studies and engagement (Aven et al., 2021; Renn, 2008; Shaw et al., 2006). The notion that currently existing democracies can provide adequate risk governance should not be taken for granted (Gattinger, 2023). Because the drivers of X-risks are systemic, they offer no short term or cheap fixes. In fact the costs of risk mitigation will themselves come to constitute such a big part of political discussion that it will, at times, take over the agenda setting even in quiet times. Adopting a proportionate response to risks could, arguably, become the costliest budget item on any national, organizational, or family budget (Aven et al., 2021; De Pryck et al., 2022; Dietz et al., 2021; *IPCC — Intergovernmental Panel on Climate Change*, n.d.). Decisions under risk are notoriously difficult to analyze and any actor with this predicament is faced with both utilitarian and prospective choices and constraints (Kahneman et al., 1979).

The inherent inequality effects of the dominant cascading X-risks represent another set of emerging concerns. The fact that several insurance companies will no longer insure real estate in California, due to the increased risk of natural disasters such as forest fires or flooding (Hao, 2023), is not only a harbinger for a world that will become largely uninsurable but also a warning about inequality. As we have seen in other states such as Florida and Louisiana, the only real estate available to socially marginalized and economically disadvantaged groups is often the most exposed to today's worsening environemtnal threats—so clearly so that insurers don't dare add the properties to their balance sheets. At the same time, the specter of AI looms large in this respect given that we have already seen a decade or more of detrimental effects on public debate due to unregulated algorithms run amok. For more on the subject in this volume, see: Bohdal et al., Yang & Sandberg, and Undheim.

Transdisciplinary engagement is necessary and will entail a mix of traditional and new approaches to knowledge gathering, discussion, consensus-building, and mitigating action. Scientists will need to become more comfortable in policy-making settings (Nathan et al., 2022a). Critically minded science and technology studies (STS) scholars need to become more comfortable engaging with the politicized security community (Evans et al., 2021). Luxton et al face this challenge in these proceedings as they try to identify the public health implications of cascading X-risks and have to analyze effects of many areas that are novel to the field of public health. The scope is so wide that no individual research project could cover it. In that case, one has to rely more than usual on secondary sources. In that sense, cascading X-risks is very much an action research paradigm (Canlas et al., 2020; Cornish et al., 2023; Etzkowitz, 2008; Shaw et al., 2006). That being said, the methods of X-risk study already draw on the panoply of scientific methods from hermeneutics and historical research to empirically driven survey research (Cremer et al., 2021; Kemp et al., 2022; Shackelford et al., 2020; Tonn et al., 2013).

The X-risk field is moving towards, on the one hand, transdisciplinary engagement and facing the challenge of developing concepts, tools, and communities that can transcend disciplinary boundaries, sectors (notably academia, the private sector, and the public sector), national boundaries, and the North/South divide. On the other hand, the X-risk field is developing on separate tracks within each science almost along the same path as research ethics has become increasingly institutionalized as a concern embedded in each set of scientific paradigms. For example, there is an ethics of survey research which in the US is managed by the Institutional review boards (IRBs) and equally a set of encoded practices for biorisk. One could imagine that the engineering practice of studying industrial risk widens across all engineering disciplines, becomes more established in the

social sciences which already have embraced the wider notion of risk society, and that civilizational risk becomes a staple of both historical and contemporary scholarship and training in the humanities. Studying the rise and fall of great powers is, of course, a traditional topic in historical studies.

Looking at the X-risk field even more broadly, where might one most fruitfully engage as a researcher, policymaker or funder? In my reading, there seem to be opportunities across the spectrum of engagement. Even though the field calls for concentrated mitigation action around the most concerning areas, which in itself will be costly and complex, the most important goal of all might be to avoid the trap of limiting X-risk to only a few areas of concern. This could prove detrimental to global systemic preparedness for great and calamitous change.

In the following section, I provide a call to action for researchers, policymakers, and funders of the study of cascading risks with potential existential outcomes for humanity.

4. A call to action: rising to the challenge of transforming risk governance

Because every X-risk jeopardizes everything that is humanly valuable, each genuinely existential risk can in principle claim the whole of our attention. For those who research this field, the challenge of transforming risk governance stands out as a clear way to reduce x-risks in general. This is for two simple reasons: firstly, good governance encompasses everything from research to infrastructure, all the way to innovation and world building; secondly, good governance of X-risk is currently non-existent (Nathan et al., 2022b). As Bressler and Alstott point out in these proceedings, this lack is certainly visible at the global level and needs immediate remedy.

As most of the papers in these proceedings argue, a single-risk focus cannot meet the many emerging challenges humanity faces. But what is a researcher to do when academic funding and prestige remain tied to work conducted in single topic deep dive mode? What is a policymaker to do about the growing impact of existential risk as a core challenge to democracy, given that the regulatory process is slow, painful, and constrained by vested interests? What is a funder to do given the wide variety of organizations that already support X-risk and the challenge of picking topics, teams, and individuals that would contribute to maximum impact?

My answer is to try to resist only investing in pet causes. The myopic focus on pandemic influenza (echoing the 1918 historical precedent) contributed to the global lack of preparedness for COVID-19. Today's myopic geopolitical concern with "avoiding another world war" (ironically at the same time as the West gambled on a NATO expansion eastward) is partly what caused the appeasement of Putin's annexation of Crimea which brought about the war in Ukraine and all the cascades that followed.[2] Alternative perspectives that kept more than one variable in mind would perhaps have created a different, and more effective, mix of caution and action among geopolitical actors.

Additional resources for readers who want to engage in X-risk studies, governance, or mitigation efforts include recent monographs from Emile Torres (Torres, 2023), Andrew Leigh (Leigh, 2021), Noah Taylor (Taylor, 2023), and Joshua Schuster and Derek Woods

[2] This has echoes of British Prime Minister Neville Chamberlain's appeasement of Hitler's annexation of Sudetenland in western Czechoslovakia which contributed to World War II (Bouverie, 2019).

((Schuster et al., 2021). This new thrust significantly complements the earlier, more limited lens of the philosophy-inspired Oxford school of X-risk (Bostrom, 2016; MacAskill, 2022; Ord, 2020) and even the more eclectic, yet still physics and economics-grounded Cambridge school (Cremer et al., 2021; Rees et al., 2004).

Above, I asked rhetorically whether cascading X-risks present governance challenges that constitute a more wicked problem than understanding the potential of the emerging technologies which, to a large extent, created the challenges in the first place. What do you think? I leave that question on the table. But the answer matters. Now that you have the direction, and the emerging tools provided by the new thrust of X-risk studies, go out there and use them, and refine them. If you are a single domain expert, try to greet the interest of generalists with generosity and work to establish the intersections between specialist fields that reveal the possibility for previously unanticipated cascades. Most importantly, whatever you do, don't study this topic in isolation, but carry out your research and agency in the public sphere. We need the widest possible discussion about a topic of such profound significance for the continuation of humanity.

Studying x-risk can be uniquely taxing—especially when work across cascades of multiple forms of risk reveals still more dangers to consider. And yet, we do not get to choose our times, only how we respond to them. What a privilege it is to belong to a generation whose decisions will have such an outsized impact. What a thrilling endeavor to be engaged in, working to more deeply understand humanity's raison d'etre (which is the obligatory passage point before making judgements), investigating humanity's present condition (a challenging task given all the factors involved in such an ongoing process), and safeguarding humanity's future (which calls for careful foresight under conditions of extreme uncertainty).

References

Ackerman, G., & Potter, W. C. (2008). Catastrophic nuclear terrorism: a preventable peril. In *Global Catastrophic Risks*. doi: 10.1093/oso/9780198570509.003.0026

Alexander, D. (2018). A magnitude scale for cascading disasters. *International Journal of Disaster Risk Reduction*, 30, 180–185.

Aven, T., & Zio, E. (2021). Globalization and global risk: How risk analysis needs to be enhanced to be effective in confronting current threats. *Reliability Engineering & System Safety*, 205, 107270.

Battistelli, F., & Galantino, M. G. (2019). Dangers, risks and threats: An alternative conceptualization to the catch-all concept of risk. *Current Sociology. La Sociologie Contemporaine*, 67(1), 64–78.

Baum, S. (2015, May 29). The Risk of Nuclear Winter. Federation Of American Scientists. Retrieved from https://fas.org/pir-pubs/risk-nuclear-winter/

Baum, S. (2017). *Global Catastrophes: The Most Extreme Risks*. SSRN.

Beck, M. (2013). Risk: a study of its origins, history and politics. Retrieved from https://searchworks.stanford.edu/view/12970259

Bostrom, N. (2016). *Superintelligence: Paths, Dangers, Strategies* (Reprint edition). Oxford University Press.

Bostrom, N., & Cirkovic, M. M. (2011). *Global Catastrophic Risks*. OUP Oxford.

Bouverie, T. (2019). *Appeasement: Chamberlain, Hitler, Churchill, and the Road to War* (Illustrated edition). Tim Duggan Books.

Caldera, H. J., & Wirasinghe, S. C. (2022). A universal severity classification for natural disasters. *Natural Hazards*, 111(2), 1533–1573.

Canlas, I. P., & Karpudewan, M. (2020). Blending the Principles of Participatory Action Research Approach and Elements of Grounded Theory in a Disaster Risk Reduction Education Case Study. *International Journal of Qualitative Methods*, 19, 1609406920958964.

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.

Compel, R., & Arcala-Hall, R. (Eds.). (2023). *Security and Safety in the Era of Global Risks* (Routledge Advances in International Relations and Global Pol) (1st ed.). Routledge.

Cornish, F., Breton, N., Moreno-Tabarez, U., Delgado, J., Rua, M., de-Graft Aikins, A., & Hodgetts, D. (2023). Participatory action research. *Nature Reviews Methods Primers*, 3(1), 1–14.

Cremer, C. Z., & Kemp, L. (2021). Democratising Risk: In Search of a Methodology to Study Existential Risk. In https://papers.ssrn.com › sol3 › papershttps://papers.ssrn.com › sol3 › papers. Retrieved from https://papers.ssrn.com/abstract=3995225

De Pryck, K., Hulme, M., Skodvin, T., Leclerc, O., Hartz, F., Livingston, J. E., Beck, S., Siebenhüner, B., Standring, A., Gustafsson, K. M., Hughes, H., Yamineva, Y., Edwards, P. N., Petersen, A. C., van Bavel, B., MacDonald, J. P., Dorough, D. S., Guillemot, H., Cointe, B., & Asayama, S. (2022). *A Critical Assessment of the Intergovernmental Panel on Climate Change.* Cambridge University Press.

Dietz, S., Rising, J., Stoerk, T., & Wagner, G. (2021). Economic impacts of tipping points in the climate system. *Proceedings of the National Academy of Sciences of the United States of America*, 118(34). doi: 10.1073/pnas.2103081118

Dijkstra, H., Petrov, P., & Versluis, E. (2018). Governing risks in international security. *Contemporary Security Policy*, 39(4), 537–543.

Ekberg, M. (2007). The Parameters of the Risk Society: A Review and Exploration. *Current Sociology. La Sociologie Contemporaine*, 55(3), 343–366.

Etzkowitz, H. (2008). *The Triple Helix: University-Industry-Government Innovation in Action*. Routledge.

Evans, S. W., Leese, M., & Rychnovská, D. (2021). Science, technology, security: Towards critical collaboration. *Social Studies of Science*, 51(2), 189–213.

Garrick, B. J. (2004). Nuclear Power: Risk Analysis. In C. J. Cleveland (Ed.), *Encyclopedia of Energy* (pp. 421–431). New York: Elsevier.

Gattinger, M. (2023). *Democratizing Risk Governance*. Springer International Publishing.

Hao, C. (2023, June 5). "A big blow": How home prices could be impacted by insurers pulling out of California. Retrieved from https://www.sfchronicle.com/california/article/insurance-home-prices-18131069.php

Homer-Dixon, T., Renn, O., Rockström, J., Donges, J. F., & Janzwood, S. (2022, July 20). A call for an international research program on the risk of a global polycrisis. Cascade Institute. Retrieved from https://cascadeinstitute.org/wp-content/uploads/2022/03/A-call-for-an-international-research-program-on-the-

risk-of-a-global-polycrisis-v2.0.pdf

InfluenceWatch. (2017, August 25). Center for Strategic and International Studies. InfluenceWatch. Retrieved from https://www.influencewatch.org/non-profit/center-for-strategic-and-international-studies/

IPCC — Intergovernmental Panel on Climate Change. (n.d.). Retrieved from https://www.ipcc.ch/

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica: Journal of the Econometric Society*, 47(2), 263–291.

Kemp, L., Xu, C., Depledge, J., Ebi, K. L., Gibbins, G., Kohler, T. A., Rockström, J., Scheffer, M., Schellnhuber, H. J., Steffen, W., & Lenton, T. M. (2022). Reply to Kelman: The foundations for studying catastrophic climate risks. *Proceedings of the National Academy of Sciences*, 119(42), e2214794119.

Lawrence, M., Janzwood, S., & Homer-Dixon, S. (2022). What is a global polycrisis? Cascade Institute.

Leahey, E. (2018). The Perks and Perils of Interdisciplinary Research. *European Review*, 26(S2), S55–S67.

Leigh, A. (2021). *What's the Worst That Could Happen?: Existential Risk and Extreme Politics*. The MIT Press.

Linder, N., & Undheim, T. A. (2022). A*ugmented Lean: A Human-Centric Framework for Managing Frontline Operations*. John Wiley & Sons.

MacAskill, W. (2022). *What We Owe the Future*. Basic Books.

MacLeod, M. (2018). What makes interdisciplinarity difficult? Some consequences of domain specificity in interdisciplinary practice. *Synthese*, 195(2), 697–720.

Naqvi, A., & Monasterolo, I. (2021). Assessing the cascading impacts of natural disasters in a multi-layer behavioral network framework. *Scientific Reports*, 11(1), 20146.

Nathan, C., & Hyams, K. (2022a). Global Catastrophic Risk and the Drivers of Scientist Attitudes Towards Policy. *Science and Engineering Ethics*, 28(6), 50.

Nathan, C., & Hyams, K. (2022b). Global policymakers and catastrophic risk. *Policy Sciences*, 55(1), 3–21.

Onderco, M., & Zutt, M. (2021). Emerging technology and nuclear security: What does the wisdom of the crowd tell us? *Contemporary Security Policy*, 42(3), 286–311.

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.

Petersen, K. L. (2012). Risk analysis – A field within security studies? *European Journal of International Relations*, 18(4), 693–717.

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

Renn, O. (2008). *Risk Governance: Coping with Uncertainty in a Complex World* (Earthscan Risk in Society). Routledge.

Rigolot, C. (2020). Transdisciplinarity as a discipline and a way of being: complementarities and creative tensions. *Humanities and Social Sciences Communications*, 7(1), 1–5.

Rodríguez, H., Donner, W., & Trainor, J. E. (Eds.). (2017). *Handbook of Disaster Research* (Handbooks of Sociology and Social Research) (2nd ed.). Springer.

Schuster, J., & Woods, D. (2021). *Calamity Theory: Three Critiques of Existential Risk* (Forerunners: Ideas First). Univ Of Minnesota Press.

Scouras, J. (2019). Nuclear War as a Global Catastrophic Risk. *Journal of Benefit-Cost Analysis*, 10(2), 274–295.

Shackelford, G. E., Kemp, L., Rhodes, C., Sundaram, L., ÓhÉigeartaigh, S. S., Beard, S., Belfield, H., Weitzdörfer, J., Avin, S., Sørebø, D., Jones, E. M., Hume, J. B., Price, D., Pyle, D., Hurt, D., Stone, T., Watkins, H., Collas, L., Cade, B. C., … Sutherland, W. J. (2020). Accumulating evidence using crowdsourcing and machine learning: A living bibliography about existential risk and global catastrophic risk. *Futures*, 116, 102508.

Shaw, S., & Barrett, G. (2006). Research Governance: Regulating Risk and Reducing Harm? *Journal of the Royal Society of Medicine*, 99(1), 14–19.

Sillmann, J. , Christensen, I. , Hochrainer-Stigler, S. , Huang-Lachmann, J. , Juhola, S. , Kornhuber, K. , Mahecha, M. , Mechler, R. , Reichstein, M. , Ruane, A. C. , Schweizer, P.- J., & Williams, S. (2022). Briefing note on systemic risk. UNDRR. Retrieved from https://www.undrr.org/publication/briefing-note-systemic-risk?_gl=1*mvi12p*_ga*MjA4NzQzOTU4Ni4xNjc2NTc0NDQz*_ga_T3RWEE6Z0J*MTY3ODEyMjEyMS40LjEuMTY3ODEyMjc0MC4wLjAuMA..

Stallings, R. A. (2003). *Methods of Disaster Research*. Xlibris US.

Taylor, N. B. (2023). *Existential Risks in Peace and Conflict Studies* (Rethinking Peace and Conflict Studies) (1st ed.). Palgrave Macmillan.

Tonn, B., & Stiefel, D. (2013). Evaluating methods for estimating existential risks. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, 33(10), 1772–1787.

Torres, É. P. (2023). *Human Extinction: A History of the Science and Ethics of Annihilation* (Routledge Studies in the History of Science, Technology and Medicine) (1st ed.). Routledge.

Undheim, T. A. (2021). *Future Tech: How to Capture Value from Disruptive Industry Trends*. Kogan Page Publishers.

Zinn, J. O. (2017). The Sociology of Risk. In *The Cambridge Handbook of Sociology: Specialty and Interdisciplinary Studies* (pp. 129–138). Cambridge University Press.