The B. John Garrick Institute for the Risk Sciences
UCLA ENGINEERING

GIRS-2017-04

FIRST INTERNATIONAL COLLOQUIUM ON CATASTROPHIC AND EXISTENTIAL RISK

PROCEEDINGS

Editor: Dr. B. John Garrick

Sponsored by:
The B. John Garrick Institute for the Risk Sciences
Henry Samueli School of Engineering and Applied Science
University of California
Los Angeles, California

Held at:
Luskin Convention Center, UCLA
March 27-29, 2017

FIRST INTERNATIONAL COLLOQUIUM ON CATASTROPHIC AND EXISTENTIAL RISK

PROCEEDINGS

Editor: Dr. B. John Garrick

Copyright © The B. John Garrick Institute for the Risk Sciences. All rights reserved.

ACKNOWLEDGMENTS AND DISCLAIMERS

The colloquium met all our expectations and it was due to the contribution of many, the presenters, participants, UCLA staff, volunteers, and the UCLA Luskin Convention Center. Ali Mosleh, Director of the Institute, was the key force behind the scene in getting the university administration to support the colloquium, developing the necessary funds for sponsorship, and adding his wisdom to the overall planning process. Special recognition is given to Christopher Jackson, Barbara Hamrick, Mihai Diaconeasa, and Liz Ward, not only for their contribution to the colloquium, but their contribution to the editing and compiling of the proceedings. Senior Fellows Roger McCarthy and Henry Petroski of the Institute performed with distinction as the moderators for Breakout Sessions 1 and 2 as David Kosson did for Breakout Session 3. David Kosson "was standing in" for Senior Fellow George Apostolakis who was ill, but joined us to make contributions during the wrap up session. Many thanks to scribes Alex Mennen, Barbara Hamrick, and Allison Duettmann for Discussion Groups 1, 2, and 3, respectively. Student volunteers and others performing key assignments on the colloquium were Alexis Umoye, Nicholas Yin, Samaneh Balali, Keo-Yuan Wu, and Yuang-Shang Chang.

TABLE OF CONTENTS

I INTRODUCTION

At the University of California, Los Angeles (UCLA), some 40 distinguished philosophers, physicists, engineers, lawyers, social scientists, administrators, students and writers spent two and a half days at an event sponsored by UCLA's B. John Garrick Institute for the Risk Sciences, examining the issue of catastrophic risks facing humans, including risks that threaten the entire future of humanity – that is "existential risks." The attendance was by invitation to professionals active in the field of catastrophic and existential risk as well as related subject matter experts. The attendees were mostly from the United States, but there was representation from Asia and Europe.

The question has been asked in many ways and by many scholars how long will humanity exist, a hundred years, a thousand years, millions of years, or more? One thing is clear, unless we confront the threats to humanity, that period could be uncomfortably short - some experts believe as short as a century. The risks are both anthropogenic (people caused) and natural. There is growing concern that such risks are now more likely to occur as a result of anthropogenic events as opposed to natural events. Human actions of concern are runaway new technologies (such as nanotechnology weapons and super intelligence machines) and weapons of mass destruction (such as bioterrorism and nuclear terrorism). The thrust of the colloquium was the discussion of such threats and the consideration of methods for predicting, preventing, mitigating, or delaying their occurrence. Emphasized in the colloquium was how these risks can be better "quantified" to enable meaningful and cost-effective defensive actions. Social, legal, and ethical issues were part of the discussion.

The colloquium was held in the new on campus full-service UCLA Luskin Convention Center. The two-and-a-half-day colloquium consisted of lectures the first day, breakout discussion sessions the second day, and presentation and discussion of the results the third half day. The breakout sessions enabled all the participants to become fully engaged in the discussions.

The attendees were welcomed by Executive Vice Chancellor and Provost Scott L. Waugh; Garrick Institute Director, Professor Ali Mosleh; and

Founder of the Institute Dr. B. John Garrick. The welcoming remarks stressed the need for much more characterization and exposure of the evidence of extreme risks. It was stressed that "knowledge" and the "evidence" supporting that knowledge should be the tenants of the quest for the truth about catastrophic and existential risks. "Knowledge, rather than judgment" is necessary for advising our leaders for making the right decisions, not only on how to save lives, but to save humanity.

II SUMMARY AND EVALUATION

II.1 WHY THE COLLOQUIUM?

Throughout human history, and across geological and cultural divides, humankind has engaged in apocalyptic myth-making. Whether the myths derive from purely religious revelations, predictions of anthropogenic (human-caused) induced doom, the imagination, or the threat of hostile artificial or alien intelligences, the common theme of ultimate extinction is something with which all cultures have mythologized.

Perhaps humankind's fascination with the end of the world stems from a dark desire to envisaging the world ending when one's own life ends, or conversely, from a deep-seated fear that all traces of humanity may someday be erased from the physical universe (so our existence was all for naught). It could also be simply a subconscious nudge to find the ways and means for humanity to persevere.

Modern humans have inhabited this planet for about 200,000 years: only a tiny fraction of this planet's 4,500,000,000 years. It is inevitable that one-day human life on this planet will end. But how it will end and when it will end is an open question: much like the endings of our individual lives. And, like our individual lives, the choices we make now with our collective lives may have bearing on how and when life on this planet ends, and where it might continue.

The risk sciences bring these issues into focus, cataloguing and prioritizing the dangers through quantification of the risks, their uncertainties, and the contributing factors.

Just how long will humanity exist on planet earth—hundreds of years, thousands of years, or millions of years? The corollary question is "what can we do to maximize our time here?" Our destiny is at risk from both natural and anthropogenic threats; but with an adequate understanding of the risks, implementation of reasonable prevention strategies, and stockpiling of mitigation resources, we increase the boundaries of the survival of our species.

In 2008, a small but distinguished group of experts on global catastrophic risks at a Global Catastrophic Risk Conference at the University of Oxford suggested there is a 19% probability of human extinction over the next century.[1] While the caveats are many on this prediction, it is nevertheless an indication from those who think most about such risks of just how fragile our existence is. In fact, it is an amazing characteristic of society that such a prediction from renowned scholars on existential risk has little to no impact on our world leaders for international action.

Some may argue that we can't do anything about existential events - at least for those arising from natural phenomena such as a super volcano, the impact of a very large asteroid, the long-range effects of a neighboring super nova, or a super geomagnetic disturbance. Certainly, some events (such as the transition of our sun to a red giant) will likely overwhelm humanity unless we conquer the challenges of intra-galactic travel. Yet others may be preventable or at least survivable with adequate preparation. Science and engineering may develop the necessary tracking and diversion tools to reduce or remove the risk of an asteroid impact. Electronic systems may be hardened to minimize the impact of a major geomagnetic storm event. Further study in this area may provide the basis for new extreme-survival technologies. Understanding the risks and preparing for them is key to maximizing our survival, but we need to be asking the right questions to begin down this path—questions such as these:

Why is knowledge about catastrophic and existential risk important to society?

Are our leaders providing adequate resources to address these issues?

Are all the necessary scientific and engineering disciplines involved in addressing the issues related to catastrophic and existential risks?

[1] "Global Catastrophic Risk."

Are contemporary methods of quantitative risk assessment and decision analysis being adequately deployed to address questions related to these risks?

What tools do we currently have for the assessment and management of catastrophic and existential risks?

What schemas and methodologies have been (or need to be) developed to prioritize catastrophic and existential events?

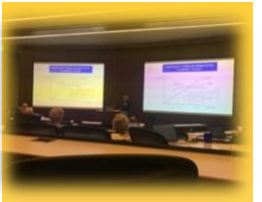
Should prioritization of the risks, and risk prevention and management strategies be driven by potential outcome (localized or regional devastation versus potential extinction)?

How should actions respecting prevention and threat-reduction versus post-event mitigation be prioritized?

Should extreme risk management strategies (such as inter-planetary emigration) be included in the consideration of existential risk management?

These questions only begin to scratch the surface of an area of study that quite literally could save humanity. One of the greatest challenges to researchers and investigators of global human survival is elevating the consciousness of the public and our leaders such that they fully appreciate the importance of taking action now to assure the future of humanity over the long term – be it thousands or even millions of years. These challenges exist in spite of the fact that many elite universities and other institutions have been studying this issue for decades. The most famous institutions making inroads to formalizing the disciplines having to do with the sustainability of the human race are The Future of Humanity Institute at the University of Oxford and The Center for the Study of Existential Risk at the University of Cambridge, both of the United Kingdom. There are others beginning to fall in line including several scholarly groups in the United States, but given that this is a world issue of extreme magnitude the voices about it have been silent

from most of the nations of the world. This suggests that experts need to engage and collaborate with other nations to find ways to better manage global and existential threats to humanity.



While the interest is in all risks that can have catastrophic consequences, the risks of primary interest in the colloquium were those that could result in adverse global consequences. More importantly, those risks that could lead to the end of the human species.

Global catastrophic risks have been defined by the Global Catastrophic Risk Institute (a United States based institute) as events large enough to significantly harm or even destroy human civilization at the global scale. The metric of global risks varies but examples are human fatalities, economic collapse, and massive cultural change or their combinations. An existential risk has been defined by Bostrom (2003)[2] as "a risk where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential." Global risks are distinguished from existential risks as being generally endurable.

The University of California Los Angeles' B. John Garrick Institute for the Risk Sciences is one of many in the United States including in its research and development agenda the quest for a better understanding of global and existential risks. Other examples of institutes engaged in addressing these risks are the Foresight Institute, Palo Alto; Machine Intelligence Research Institute, Stanford; the Center for Catastrophic Risk Management, University of California Berkeley; the Future of Life Institute, Boston; and the aforementioned Global Catastrophic Risk Institute.

Among the tasks these and other organizations are engaged in are improved methods and activities for preventing, predicting, mitigating, or delaying global and existential risks. The Garrick Institute has an international reputation in developing and applying the risk sciences to the risk of severe accidents involving nuclear power, transportation, chemical and petroleum and other industries. It is logical to extend this experience to global and existential risks as many of the established risk

[2] Bostrom, "Existential Risks."

assessment methods will apply. See the Sidebar on the General Theory of Quantitative Risk Assessment.

General Theory of Quantitative Risk Assessment (QRA)

It should be noted that there are proponents who believe a general theory of QRA exists and is applicable to any kind of risk, including global catastrophic and existential risk. The hypothesis is that there exists a set of overarching and axiomatic rules that apply to any type of risk. That was the motivation behind conceiving the "triplet definition of risk" (Kaplan and Garrick, 1981)[3]: a general framework for considering any kind of risk. Embedded in the axioms is a specific definition of probability based on the supporting evidence that obeys Bayes theorem - the fundamental framework for developing and processing the probabilities. The primary difference between applications is in the boundary conditions, scope, and representation of the basic phenomena driving the risk. These basic elements of an attempt at a general theory of QRA are highlighted in Chapter 2 of Garrick (2008)[4] with an example in Appendix A. The generality of this theory has been tested in dozens of applications varying from catastrophic industrial accidents, natural events, national defense, to the risk of new diseases being contracted by mixing different animal species.

An important aspect of the general theory is how it embraces all forms of evidence including actuarial experience, modeling and analysis insights, and expert knowledge. The theory includes a robust

[3] Kaplan and Garrick, "On The Quantitative Definition of Risk."
[4] Garrick, *Quantifying and Controlling Catastrophic Risks*.

means to coherently combine information (including uncertainty) from these diverse sources.

There are some caveats to the claims of the referenced general theory. The general theory proponents make a clear separation between risk analysis and decision analysis (and therefore risk analysis and risk management). This was to avoid diluting the focus on specifically answering the question "what is the risk?" Risk analysis is interpreted by the general theory proponents as quantifying the answers to the three overarching questions about systems and events that constitute a risk. The questions are "what can go wrong?" "how likely is it if it does go wrong?" and "what are the consequences?" On the other hand, decision analysis generally must consider costs and benefits in addition to risk. Formal decision analysis must also address the difficult issues of preferences and value judgments, as well as other issues that might enter into the decision such as what the decision analysts call "affect,"[5] or "decision affect theory."

As to the proclaimed general theory of QRA, the "what can go wrong?" question is represented by a set of scenarios; the "how likely is it if it does go wrong?" is based on a formal definition of probability and its supporting evidence; and the "what are the consequences?" question deals with the terminating event of the scenarios. Thus, the core calculations are scenarios and likelihoods. The final step is assembling the scenarios into coherent representations of the risk - the calculus of which is well developed. Of course, decisions have to be made on the risk measures most appropriate to the application. Typical risk measures are injuries,

[5] Mellers et al., "Decision Affect Theory."

fatalities, damage levels, dose levels, damage and evacuation costs, and environmental impacts.

Another major caveat of the general theory proponents is that while the framework is completely general, the analytical processes within the framework are not. They must be application-specific. This is where many of the questions having to do with QRA limitations raised by Breakout Session 1 of the colloquium really apply. Subject matter experts must change with the application to assure the most informed process for answering the triplet of questions. There is no question that scenarios, likelihoods, and consequences are completely general components that can constitute the architecture of the risk of any system or event. What matters is the quality and rigor of the analyses used to answer the questions and the choices made on how to present the results.

What makes generalizing the approach possible is embracing the uncertainty sciences. Quantifying the uncertainties, in principle, enables the consideration of any kind of risk, including global and existential. To be sure, there will need to be additional algorithms and techniques for different types of risk as noted in Breakout Session 1, but the overall framework, has been demonstrated to be completely general. In those cases where there were flaws in the analyses, it was a matter of incompleteness in the scenarios, not the lack of generality in the method. This is the reason for assuring that the analysis team includes the appropriate subject matter experts, a requirement not always fulfilled in practice.

II.2 LECTURE SUMMARIES

Nine lectures were given across a broad range of catastrophic and existential risk issues to be the basis for intense debate and discussion. A summary is provided below. The full lectures are provided in Section III.

The following presentations are the basis for the discussion in this section.

- The State of Research in Existential Risk – Seán Ó hÉigeartaigh
- Towards an Integrated Assessment of Global Catastrophic Risk – Seth Baum, Tony Barrett
- Nuclear Terrorism – Albert Carnesale
- Biological Terrorism: An Existential Threat or Merely a Weapon of Mass Disruption – Peter Katona
- Biological and Nuclear Terrorism Risk Assessment – Detlof von Winterfeldt
- The Tragedy of Uncommons: Psychology, Politics and Policy – Jonathan Wiener
- Societal and Ethical Issues Related to Catastrophic and Existential Risk – Anders Sandberg
- Risks and Risk Management in Systems of International Governance – Catherine Rhodes
- Some Approaches for Reducing Catastrophic and Existential Risks – Christine Peterson, Mark S. Miller[6], and Allison Duettmann

Essential to begin a discussion on existential and catastrophic risk is a common understanding of the terms and current status of risk science research in this area. Lecturer Ó hÉigeartaigh's presentation, "The State

[6] Researcher, Google.

of Research in Existential Risk," opened the lecture session with an overview of the terms, topics, resources, and challenges relevant to the study of both existential and catastrophic risk. Lecturer Baum followed with his presentation, "Integrated Assessment of Global Catastrophic Risk," also addressing these core issues. What was immediately clear from both lectures and the follow-up question and answer sessions is that the application of quantitative analyses in this area is yet nascent, but the field is ripe for cultivation.

While formally defined earlier, in simple terms existential risk may be thought of as the risk of the premature extinction of humankind. Global catastrophic risk generally does not pose the threat of extinction. However, it still may produce large, irreversible impacts on humanity (including large numbers of fatalities, injuries, or disabling illnesses) regionally, socially, or culturally. While colloquium participants all had a sense of the types of events that could trigger extinction or a global catastrophe (asteroid collision, catastrophic climate change, pandemics, nuclear war, or extreme space weather), it became apparent through both the lectures and discussion there is a need for a more formal, risk-informed catalogue of such events, which could be used to systematically prioritize preventive and mitigating actions.

Categorization of risks still relies, in large part, on disparate and incomplete efforts to capture the totality of existential and catastrophic risks, and various methods tend to make artificial distinctions (e.g., "natural" versus "man-made"), which may obscure important elements

in the universe of risks. As Ó hÉigeartaigh and other participants brought to light, there are risks of synergies that the current ad hoc system of categorization may not identify. These might include cascading events – such as one catastrophic event (an earthquake) triggering a second (a tsunami); or interactive events – such as a famine, leading to war, leading to loss of infrastructure, exacerbating the famine; or mitigation risks – such as when an action taken to prevent or mitigate a risk gives rise to a different risk of the same or greater magnitude as the original.

Baum brought to light the somewhat subtle, but important distinction between simply talking about risks in the language of initiating events, and risks in the more fluid universe of the outcomes to pursue or avoid. In very simple terms, this is the difference between talking about the risk of a specific event occurring versus the risk associated with various outcomes (survival and recovery to normal, survival and recovery to something new, survival without recovery, or extinction). This distinction between event and outcome underscores yet another element necessary to the effective categorization of existential and catastrophic risks.

Baum also raised the point that prevention of dangerous technologies is unlikely. In particular this reinforces the point from Lecturer Carnesale's presentation, "Nuclear Terrorism," that the most stable state relative to nuclear weaponry is one where each state actor maintains some arsenal, and that total disarmament is probably not feasible since it is in one's own interest to be the last to disarm. Similarly, research in artificial general intelligence is likely to continue despite laws or sanctions against them, making prevention by prohibition unlikely.

With respect to survival, Baum made the point that societal resilience will be crucial. While one may not be able to prevent dangerous technologies from moving forward, one may build robust and diverse infrastructures that can withstand catastrophic and near-existential events. A resilient foundation for infrastructure coupled with pre-staged knowledge resources, on science and engineering for local and regional survival, may ensure adequate residual survival to overcome most events short of planetary annihilation.

Finally, Baum touched on strategies for communicating the need for research and support in the area of existential risks, including 1) direct communication to politicians, media, and public on risks, which is fact-based and non-alarmist; 2) indirect communications to the same audience, focusing on extreme results and potential harm (i.e., to alarm the audience into action); and, 3) very indirect communications to the same audience, essentially coercing cooperation via an offer for collateral benefits of actions being taken for a different purpose.

A litany of acknowledged (if yet potential) risks began with both Ó hÉigeartaigh's and Baum's views on the risks of artificial general intelligence, and continued with Carnesale's presentation on the nuclear dangers; Lecturer Katona's presentation, "Biological Terrorism: An Existential Threat or Merely a Weapon of Mass Disruption;" and Lecturer von Winterfeldt's presentation, "U.S. Department of Homeland Security Activities on Nuclear and Bioterrorism, "touching on both prior topics. As the lecturers examined specific existential and catastrophic risks, there was a rising awareness, apparent in both the lectures and the follow-up question and answer sessions, that existential risks are indeed quite exotic, while catastrophic risks are relatively commonplace in the history of humanity.

Lecturer Peterson in her presentation "Some Approaches for Reducing Catastrophic and Existential Risks," took on two important anthropogenic risks: advanced nanotechnology and artificial general intelligence. Her perspective was specifically oriented toward risk reduction, a critical element of all risk management strategies, but also playing a role in risk assessment. Peterson examined this aspect by example, offering various scenarios of attacks and the different types of advance defense mechanisms that should be considered, in this way demonstrating the need for flexibility and innovation in the approach to accounting for reduction of likelihood (preventing an event) or mitigating consequences.

The demonstration of the usefulness of creative approaches to risk reduction was enhanced by the enumeration of different levels of nanotechnology and artificial general intelligence attacks, requiring, perhaps, different preventive or mitigating measures. Examples were provided of how society has met such challenges in the past, with the most notable example being the avoidance of a nuclear holocaust since development of nuclear weapons over seventy years ago. As pointed out by Peterson and others, the use of nuclear weapons has been kept in check, primarily through the implementation of non-proliferation treaties supported by direct monitoring of the activities of nations with nuclear capability - a complex, multinational effort that has demonstrated significant success thus far.

As was made clear by Lecturer Rhodes in her presentation "Risks and Risk Management in Systems of International Governance," global and existential risk management involves not only solving many technical and social problems, but international governance problems as well. In particular, there must be agreed upon preventive and mitigating plans and actions to address global or regional risks, potentially impacting many diverse nations, or of sufficient magnitude to render the action of a single impacted state inadequate.

Rhodes went on to discuss how international governance systems can be effective in managing risk, but also how they may be a source of risk as well. One major challenge to our resilience is ensuring a robust international system in place to effectively cope with global and existential risks. The political challenge is that an effective system must

be truly universal, with no restrictions on participation based on geographic, economic, or other grounds.

Rhodes illustrated one possible set of elements an effective system might have. Among the components are reliable risk analysis procedures, intergovernmental reporting requirements, surveillance and monitoring systems, expert networks, and prohibitions as necessary. Specific examples of each component were provided in the context of biotechnology.

The dissonance of the probability of an existential event versus the consequences percolated amongst participants throughout the lecture series and the discussion sessions the following day. Some participants questioned the feasibility of constructively applying the risk sciences to existential risks (see earlier Sidebar). Although one of the fundamental purposes of engaging in QRAs is to provide a coherent structure for prioritizing risks and proposing preventive or mitigating actions, it was generally agreed that even a less rigorous (though still highly structured) analysis can provide extremely valuable information to decision-makers by revealing gaps and uncertainties, providing a rational basis for optimizing actions. In practice, the most valuable output of QRA is exposing and quantifying the contributors to risk; such information enables specific corrective actions to either eliminate or reduce the risk.

Lecturer von Winterfeldt identified tools available to supplement and enhance traditional risk assessment methodologies, including game theory, possibility theory, fuzzy set theory, expert elicitation, and risk scoring and ranking methods. These tools may be of distinct value in the assessment of the category of "intentional events" – i.e., the risks faced by society as a result of humans' intent upon harming other humans on a large scale.

Another major topic of lecture and discussion was that regarding support for the risk sciences and public and political awareness of the threats facing humanity. Lecturer Wiener's presentation, "The Tragedy of Uncommons: Psychology, Politics and Policy," highlighted several challenges, such as the challenge of "unavailability," which refers to the fact that, unlike rare events (which often elicit exaggerated public support for prevention), ultra-rare events (such as a near-extinction

event) tend to result in less concern or support. As well, Wiener pointed to the "psychic numbing" effect which, perhaps counterintuitively, finds that the willingness to authorize spending to prevent or mitigate an event goes down as the number of (potential or real) lives lost goes up. Wiener also discussed the need for "smart management" in the prioritization of risks and optimization of prevention and response options; overly simplistic approaches (e.g., the "precautionary principle"[7] may result in unintended and unacceptable risk trade-offs, making the cure, as it were, worse than the disease).

Opportunities for advancing the risk sciences primarily lie in effective communications. Wiener highlighted several approaches to spotlighting the need for support for this area of research: 1) emphasize the benefits – where net benefits are very large, opposition to funding and regulation may be overcome; 2) identify high-profile policy ambassadors (e.g., Bill Gates) who can increase the acceptability of regulation and contribute to increased collective funding for assessment, prevention strategies, and mitigation strategies; and, 3) take advantage of crisis events, which are teachable moments, and will drive an increase of support for regulation and funding.

Lecturer Sandberg's presentation, "Societal and Ethical Issues Related to Catastrophic and Existential Risk," raised questions about how to value that which may be lost in a catastrophic or near-existential event, including the calculation of future lives lost. Communication of the true value of humanity, both now and long into the future, is not trivial and is crucial to conveying the importance of addressing the risks we face today. Sandberg also pointed to the need for engagement of the public on these issues. Specifically, public distrust of experts is mitigated by sincere engagement and may ultimately result in a more stable prioritization of risks and potential actions. Prediction markets are also a useful tool in overcoming public resistance to expert advice. Not only may the prioritization become more stable when there is public buy-in to the process, but from a moral and governance perspective, maximizing inclusion reflects the fundamental values of society.

[7] "Precautionary Principle."

II.3 BREAKOUT SESSION SUMMARIES

The breakout sessions were organized around three themes, with participants self-selecting the session they preferred to attend. Prior to the sessions, the moderators met to select prescriptive questions appropriate to the session topic to focus the discussion. A narrative discussion of each breakout session follows. The narrative discussions are based on notes taken during the session and session presentations delivered at the colloquium the day following the breakout sessions.

II.3.1 BREAKOUT SESSION 1 - IDENTIFICATION, PREDICTION AND QUANTIFICATION OF CATASTROPHIC AND EXISTENTIAL RISKS

Question 1: *Are the contemporary methods of QRA (QRA is interpreted to be the same as probabilistic risk assessment (PRA)) being employed to better calibrate the types of risks of concern (scenarios, likelihoods, consequences)?*

Question 2: *To what extent have the risk sciences been applied to quantifying such risks and have they been successful? What has happened to the results?*

Question 3: *Is the lack of action "knowledge based" or just the view that there are other more important issues facing society?*

Question 4: *How do we get attention and action on being better prepared to cope with a catastrophe whether it is regional, global, or existential?*

It was the view of the Breakout Session 1 participants that current contemporary methods of QRA and PRA are not being applied to global and existential risk scenarios nearly as much as expected, given the opportunities to do so. Organizations such as the Global Catastrophic Risk Institute are doing some excellent work tying together various pieces needed for a global risk assessment, including consequence

profiles. Unfortunately, they have not had the opportunity to advance their model to the level of a quantitative analysis. Quantification involves a full probabilistic analysis of the risks with uncertainty. The lack of quantitative applications may be due to the absence of a state-sponsored international effort to better manage the prevention, prediction, mitigation, and postponement of such risks.

Participants also pointed out that while catastrophic risks (such as asteroid strikes, pandemics, and nuclear terrorism) are being studied, there lacks the use of rigorous quantitative models in a visible and impactful way. There are isolated attempts at applying QRA methods to catastrophic threats (Garrick, 2008), but they need greater support, continuity, and more rigor. Breakout Session 1 participants are correct in that there is no known internationally or nationally organized effort to apply such methods to global or existential risks. Efforts to date have been primarily to illustrate how QRA could be used to assess catastrophic risks, but have not yet reached the level of a rigorous application.

A possible exception on the national level is the effort of the United States Department of Homeland Security (DHS). They have major science and technology activities relating to homeland security. These activities include 12 University Centers of Excellence, each with a specific focus. Among their areas of research is risk analysis. Based on the participants' understanding of DHS' efforts, it appears that programs addressing global and existential risks are still in the early stages of development, and the role of QRA has not yet been resolved. DHS' mission is strongly focused upon the risk of terrorism. Therefore, in addition to QRA (and how it may be applied to human-driven events), DHS is actively pursuing adjunct methodologies, including game theory, possibility theory, fuzzy set theory, and risk scoring methods.

Breakout Session 1 participants brought forward an important point relating to the clarity of the results of QRA efforts. In particular, the participants suggested that the risk measure be transparent and easily interpreted. Metrics, such as fatalities, economic loss, environmental impact or their combinations, should be the defined outcome, rather than something with no direct physical meaning, such as a damage coefficient, safety index, or a utility factor. The use of commonly

understood metrics can make these efforts more accessible to the public and policy makers.

An additional observation was that the promotion of QRA to assess global catastrophic risks would greatly benefit from some pilot-project successes. Examples of projects which could serve this effort were the risks associated with asteroid impacts or climate change, where there has already been considerable study and success in raising public consciousness of the need for action. More in-depth assessments of these risks using QRA might make the case for its utility or at least expose its limitations and make clearer what new methods need to be developed to get the desired results.

Another means of raising the consciousness of the public and the decision makers discussed by Breakout Session 1 participants was to work through the organizations whose mission it is to advise the United States Federal Government on matters important to our nation. One such organization is the National Academies of Sciences, Engineering, and Medicine (NASEM). Forming a NASEM committee specifically addressing the urgency of better managing global and existential risk issues could provide a path directly to the decision makers for developing initiatives for a more consistent and deliberate global risk management program. One possible result from such a committee

would be a protocol for monitoring the progression of precursor events to global and existential events. While the majority of the participants believed current quantitative methods to be a reasonable starting point for taking action to become more informed about such events, it was noted that the unique characteristics of virtually every risk with potential global consequences could require event-specific augmentation of (or modification to) the traditional QRA methods. See earlier Sidebar for more on this point.

II.3.2 BREAKOUT SESSION 2 - EXAMPLES OF CATASTROPHIC AND EXISTENTIAL RISKS

Question 1: *How feasible is physicist Stephen Hawking's suggestion that we should be looking for another planet as a way to avoid an existential event? Kepler - 452b, 1400 light years away, may offer some hope on this option. There would have to be colossal breakthroughs in transportation and communication. At the speed of the New Horizons Spacecraft [37,000 mph] it would take 26 million years to get there.*

Question 2: *How involved is the engineering community in developing defensive measures against extreme events having the potential for catastrophic consequences? What are the engineering challenges to better managing catastrophic risks? Why doesn't engineering have a stronger presence in implementing real solutions to many catastrophic risks?*

Question 3: *What are the options for reducing the risks of nuclear and bioterrorism? What is actually being done and by whom? Reducing this threat is an example where engineering could play a major role.*

The first question taken up by Breakout Session 2 participants was that relating to the feasibility of migration to another planet in the face of planetary calamity. This question kicked off the discussion, which then

moved organically to other topics, and did not necessarily cover all other questions assigned to Breakout Session 2 in the order presented.

The idea of intra-galactic emigration to a new planet was dismissed rather quickly by the participants as infeasible, including intra-solar-system emigration. Participants noted there exist theoretical non-planetary-based options, such as O'Neill Colonies (O'Neill, 1976)[8], which may prove the most realistic possibility for human survival if the earth were to verge on the uninhabitable.[9] In addition, participants noted that however out-of-reach certain ideas may appear in the present, it is worth noting that ideas born of science fiction eventually came to fruition, and that it can still serve as a fertile ground for ideas to explore. It was also suggested that while inter-planetary migration may be infeasible at this time, information can still be gathered and assimilated as an adjunct to space missions that may take place for other purposes.

The discussion turned to whether QRA is even feasible for assessing potential human extinction events, since there is little or no data with which to begin the exercise. On the other hand, it should be noted that data is much too narrow a perspective when performing a risk assessment. The more appropriate term is "evidence" which embraces all forms of relevant information, including "data," modeling and analysis insights, and expert knowledge.

Some participants felt that a focus on existential risk was a distraction - policy makers would lack interest in events that have not yet been part of the human story (a point also made during the lecture sessions)[10]. Others countered that the exercise is useful to scope problems, build models, identify bounding conditions, and identify specific knowledge gaps.

Breakout Session 2 participants also discussed the fact that even near-existential events are relevant to these analyses. Consideration of events that could result in the collapse of civilization, with the attendant loss of knowledge, or events involving a sufficient number of fatalities

[8] O'Neill et al., *The High Frontier*.
[9] Hadhazy, "How We Could Actually Build a Space Colony."
[10] From lecture 6

to place successful repopulation and recovery at risk over the long term should be included in the universe of existential events. It was noted that these themes are also often explored in fiction and provide inspiration for realizing the full universe of existential risks.

The focus then turned to what event(s) could kill a sufficient number of people that it might cause the collapse of a society as a whole, or regionally. One issue raised was the subject of much discussion. This discussion revolved around the risks posed by humanity's growing dependence upon power and interconnectivity, pointing out that major interference with power generation, communication or information security or transmission could lead to national and trans-national instability, potentially resulting in war, social unrest, and devolution of society, which could threaten millions or billions.

Our current information systems (e.g., banking, power distribution, etc.) are worse than simply vulnerable and unsecured. Three major points were made in this regard: 1) the current systems are not securable, 2) a high level of sophistication is not required to take them down, and 3) there is a system, wholly unlike our current operating systems, that can meet our needs, but transition will be difficult.

It was reported that it will take a concerted effort to begin a transition, perhaps starting with using the technology in the "internet of things,"[11] and advances in "bridge-building" technologies (i.e., to bridge the old technology with the new, without compromising the new system). This will then begin the integration of a more secure platform (i.e., to "grow" it) into society, until eventually it underlies our interconnectedness. Breakout Session 2 participants felt this information was important enough to bring to the attention of the full colloquium, since it represented a potentially global risk that had perhaps not received

[11] See, e.g., Xia, F., et al., "Internet of Things," INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS Int. J. Commun. Syst. 2012; 25:1101–1102 Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/dac.2417

adequate attention as the risk relates to potential catastrophic or existential risk.[12]

In a similar vein, another Breakout Session 2 participant raised the point of our grid vulnerability to "space weather," and in particular to coronal mass ejections, which could also result in a breakdown in our communications and financial systems, leading to national or trans-national instability. Since this had not been raised specifically during the lecture session, the participants chose to include this risk as one that requires more consideration.[13]

Other risks discussed by Breakout Session 2 participants included those that had been raised during the lecture session: 1) nuclear war, 2) pandemics, 3) artificial general intelligence, and 4) genetic manipulation. With the possible exception of the genetic manipulation risk, these other risks were not seen as rising to the level of an existential risk in and of themselves, but could clearly pose catastrophic risk, and serve as a precursor to additional calamity (via a cascading effect, or through another unidentified synergy).[14]

Ultimately, Breakout Session 2 participants determined that we need better characterization and elucidation of the risks, with concomitant evidence to support prioritization. Everyone agreed that a panel of experts should be convened to examine the issues discussed, and perhaps scope out a research agenda.

II.3.3 BREAKOUT SESSION 3 - GOVERNANCE, SOCIETAL, AND ETHICAL ISSUES RELATED TO CATASTROPHIC AND EXISTENTIAL RISKS

[12] Miller also provided me with reference materials for his comments relating to our un-securable operating systems, and the development of secure systems, which I've listed on the last page.

[13] There is a relatively recent example of such an event (1859 solar storm, which took our telegraph systems in Europe and North America), (see, e.g., https://science.nasa.gov/science-news/science-at-nasa/2008/06may_carringtonflare).

[14] Group 2 specifically asked Katona to discuss biological (pandemic) risks, and the conclusions were that they were large, but historically not existential.

Question 1: *What should be the driving considerations for prioritizing catastrophic and existential risks?*

Question 2: *What are the benefits to society of a better understanding of catastrophic and existential risks, and why is such knowledge important?*

Question 3: *What is missing in terms of technology and governance to put initiatives in place for the better management of catastrophic and existential risk?*

Question 4: *What means and mechanisms exist or need to be created to raise the consciousness of our political leaders of the importance of taking action on the management of catastrophic and existential risks?*

Session 3 participants discussed a number of metrics that should be included when considering how to effectively prioritize catastrophic and existential risk. Among those discussed were the basic risk triplet. Additionally, prioritization should include consideration of a variety of time-dependent variables, describing (for the universe of events) expected time to occurrence (related to likelihood); expected warning time (once the known event is in motion); time required to adequately prepare for event; and time required to recover or adapt.

Furthermore, risk management decisions must play a role in prioritization to maximize the expected (undiscounted) utility of action, taking into consideration the impact of synergistic (multi-risk or cascading) events, as well as co-benefits and countervailing harms of a given set of actions. Finally, defining acceptable outcomes require stakeholder input with multiple cultural perspectives.

The participants also reflected upon why this endeavor is important, with the most obvious reason that the survival of humanity may be at stake, but (for less than existential events) also the survival of civilization, culture, governments, and knowledge. Awareness of the

risks can provide a basis for collective, multi-national action. Bringing these issues to a larger stage provides perspective on our present institutions and what is important to humanity, and may provide occasions to identify opportunities for improvement as well as knowledge and process gaps. The examination of these risks also provides context that may enlarge people's frame of mind with respect to risks in general, and provides hope and a sense of control over our collective destiny, influencing and imagining posterity in a positive way.

There is also value in shining a light on humanity's past to identify potential future problems and provide insights into the relevant actors and roles in humanity's history. For success, it is imperative that multidisciplinary and multicultural approaches are brought to subject. Different levels of understanding and examination will be important to different groups (e.g., policy makers versus scientists versus general public). Presentation of the information will require contextual balance to spur desirable action but not incite panic.

Exigent needs for this effort include both refinement of simulations and tools, as well as an approach to motivate political will to expend capital and resources on prevention strategies. Success would mean the impacts do not occur or are not severe, which in turn can undermine support. In such cases, where success is something not happening, it can be difficult to provide empirical proof of benefit. In this regard, the scientific community may also need to seek lessons from the environmental community, and consider the changing intergenerational ethics with respect to motivating action. Recent literature assesses cost-benefit in terms of ever-increasing prosperity rather than imminent disaster; and, emphasizes sustainable development goals, adhering to a positively focused vision that enhances resilience.

Specific strategies might include the use of co-benefits to engage policy-makers (e.g., pandemic preparedness activities also provide important information on prevention and mitigation actions); or, as Von Winterfeldt suggested, the use of high-profile public figures (e.g., Elon Musk or Bill Gates) to attract media attention to the issue, provide an opportunity for media education, and reach a larger and more diverse audience.

For the catastrophic and existential risk agenda to achieve a place of prominence, it will be important to encourage national security leadership to take ownership of the issue. A suggested approach is to establish an interagency task force under executive branch leadership. For example, DARPA might be well-suited to perform the horizon-scanning function (beginning the efforts to categorize and prioritize the universe of risks), while DHS could take on preparedness and response assessments, with the United States Department of Energy and the national laboratories supporting research efforts.

As a beginning, it may be appropriate to stand-up a committee of NASEM with multidisciplinary representation to scope the issues, and engage with international partners to produce a focused research agenda that addresses identification, prioritization, prevention, mitigation, recovery, use of technology for governance, and decision options. Ultimately, an international advisory body (e.g., modeled after the Global Catastrophic Risk Institute) is probably a reasonable avenue to ensure appropriate global perspective and assist individual nation-states in developing programs for both research and governmental action.

II.4 TAKEAWAY MESSAGES

1. Strong consensus on convening a national panel to raise the consciousness of the public and decision makers on catastrophic and existential risks.

As expected, there was a variety of opinions in the colloquium on how best to capture the attention of our national leaders on the seriousness of global and existential risks. While the frequency of catastrophic and existential events is believed to be extremely small based on natural events like asteroid impacts and gamma bursts from exploding stars, modern technology has led to a situation where anthropogenic events, that is, people caused events, are now believed to be the greatest risk for human extinction.

This seriousness of the risk or threat has not been grasped by society and it is clear that in order to get the attention of our national leaders it will be necessary for this issue to be taken up by a national level

institution that Congress looks to for advice. An example of such an institution is NASEM. Upon the authority of the charter granted to it by the Congress, NASEM has a mandate that requires it to advise the federal government on scientific and technical matters. For example, through NASEM's National Research Council (NRC) a committee could be formed with internationally recognized experts to answer explicit questions about existential risk. Among the questions to be addressed are 1) "what does the government need to do to deal with this issue?" and 2) "how should the technical and scientific community contribute to better understanding of such risks?" Important to realizing an NRC committee is finding an appropriate sponsor.

2. A primary need is to develop and implement a protocol for prioritizing a variety of issues associated with improving the management of catastrophic and existential risks. Among the issues are the risks of concern, the research required, and the decisions that need to be made.

A systematic process needs to be developed for identifying and quantifying catastrophic and existential risks to aid the decision-making on actions to better manage such risks. The quantification of the risks will expose where the greatest uncertainties are and enable a clearer path for fruitful research activities. QRA is mature for many rare types of risk, such as industrial accidents, transportation systems and natural events such as earthquakes and severe storms. The nuclear power industry is the most advanced in the application of probabilistic risk assessment (PRA), their term for quantitative risk assessment. Some of these methods are just beginning to be applied to global and existential risks, but much more needs to be done to match the maturity of their use in nuclear power safety. Most of the basic ideas and methods exist to quantify catastrophic and existential risks. Clearly, there will have to be new algorithms and extensions to fit the needs of existential risk. One major difference will be in the scope and the boundary conditions employed. The investigations in this area, while beginning to occur, are very limited due in part to the lack of support and funding.

3. Catastrophic and existential risks need to be better characterized in terms of the nature of their threat and the evidence supporting their existence.

In order to perform evidence based assessments of the risk of global and existential catastrophic events, there needs to be more transparent information on just what the risks are and how they are manifested. The supporting evidence needs to be quantified using the uncertainty sciences to enable propagating the uncertainties through the risk assessment models. Such characterization of the supporting data is essential to properly representing the parameters in the risk model. Characterization of anthropogenic risks is more complicated and likely more uncertain than natural events.

Natural events have the benefit of a much longer observation time, while anthropogenic events such as the risks associated with nanotechnology weapons and super computer machines of artificial intelligence have no history and their consequences are not really well defined. Also, while "natural events" can be understood in terms of their physical nature, anthropogenic risks may have elements of deliberation and intentionality. For the most part, the risks have only involved thought experiments (speculated scenarios) as evidence, thus the uncertainties at the parameter level are very large and may be one of the reasons for thinking such events are now our greatest concern from an existential risk perspective. That is, the major contributor to existential risk is uncertainty. This translates into a critical need for research and analysis to have a basis for better characterization of the existential risk threat. Research may, in fact, not support the view that anthropogenic events are our greatest threat.

4. Coupled processes and synergies are important factors in identifying potential catastrophic and existential risks. A greater understanding of the events having such potential may expose far more opportunities for catastrophic events than currently considered. It is even possible that synergies represent the greatest threat of a global and existential event.

Single events such as asteroids and cataclysmic galactic events are obvious candidates for catastrophic and existential risks as are new and uncontrolled technologies, but there are many other subtler threats of a synergistic and interactive nature. Threats may start out being confinable, but due to a catalytic effect may become uncontrolled and cascade into something much more serious that may even have global consequences. For example, the combination of a super storm (hurricane or tornado) and the diseases caused by the storm, of which one may have a deadly and contagious strain, could lead to an uncontrolled pandemic. The result is the need to know which precursor events have the potential to trigger other processes and events. Results from such research could be an entirely different outcome than currently perceived.

5. As with the risk field in general, communication with different sectors of society on matters of catastrophic and existential risk is one of the greatest challenges to generating interest in understanding and supporting actions to better manage such risks. More "communication-friendly" methods need to be developed for raising the consciousness of society about the types of risks we face to get the support needed to do something meaningful about predicting, preventing, mitigating or delaying these rare but real risks to humanity.

The results of rare event analysis suffer from a society that thinks more in terms of events that occur in our life spans. We comprehend the frequency of major storms, election cycles, horrific murders, depressions, wars, and pandemics. We even take seriously historical events over periods of time involving a few thousand years. As a rule, we don't do very well with events that have frequencies of a thousand or millions of lifetimes. We tend to just not take them seriously, even if in reference to the existence of humanity. The important fact is we may not be dealing with thousands or millions of years for an existential event. We simply don't know because serious studies have not been made.

The much lower frequencies generally associated with global and existential events simply do not register with the general public. Of course, the exception is the existential risk research community, which is growing as manifested by the number of institutions and professionals engaged in this new and challenging field. Nevertheless, this is a miniscule fraction of society and the burden is on them to communicate their message in an effective manner. It is clear there is no one way to communicate to the different sectors of society. For example, politicians, the public, government agencies, and the media will all require different types of information, with different interpretations of the risks and what the risk analyses really mean. It is most likely that different metrics or a combination of metrics and narratives will have to be used to effectively communicate risks to different groups.

III LECTURER PAPERS

III.1 THE STATE OF RESEARCH IN EXISTENTIAL RISK (SEÁN Ó HÉIGEARTAIGH)

Seán Ó hÉigeartaigh is the Executive Director of the Centre for the Study of Existential Risk (CSER) at the University of Cambridge. He is also a Senior Research Associate at the Leverhulme Centre for the Future of Intelligence, where he leads CFI's Policy and Responsible Innovation project, and is a co-investigator at the Strategic AI Research Centre.

In the last fifteen years there has been substantial growth in research on existential risk – the category of risks that threaten human extinction, or the permanent and drastic reduction of humanity's future potential. A number of new organisations focused explicitly on existential and global catastrophic risk have been founded in recent years, complementing the long-standing work of existing centres focused on specific risk areas such as nuclear war, biosecurity, climate change and systemic risk. This paper provides a brief overview of the emergence of this new research community, and provides a case study on the community's research on potential risks posed by future developments in artificial intelligence. There exists the opportunity for powerful collaboration between the new approaches and perspectives provided by the existential risk research community, and the expertise and tools developed by the risk sciences for risks of various magnitudes. However, there are a number of key characteristics of existential and global catastrophic risks, such as their magnitude, and their rare or unprecedented nature, that are likely to make them particularly challenging to submit to standard risk analysis, and will require new and specialised approaches.

III.1.1 EXISTENTIAL AND GLOBAL CATASTROPHIC RISK

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, 2002)[15]. For most practical purposes, this refers to developments that might wipe out humanity, or lock us into a situation we (or other intelligent life on earth) cannot recover from, such as a major and permanent global civilizational collapse.

The Centre for the Study of Existential Risk, among other organisations within the existential risk research community, also includes a focus on

[15] Bostrom, "Existential Risks."

global catastrophic risks. Here, a common definition is that used by the Global Catastrophic Risk Institute:

Global catastrophic risk is the risk of events large enough to significantly harm or even destroy human civilization at the global scale" ((http://gcrinstitute.org/concept).[16]

For this definition of global catastrophic risk, human extinction risks would be a subset of a broader set of global catastrophic risks. However, precise definitions for these classes of risk are less important than having a shared sense of the magnitude of scope of events or developments under consideration, and the key characteristics of such events or developments. The study of existential and global catastrophic risk restricts us to events or trends that might lead to a full civilizational collapse. It rules out localised catastrophes, and global-scale events that would represent a tragedy but that would not impact our civilisation in the longer-term – unless these events were likely to play a key role in more severe and permanent cascades and collapses.

Events as historically significant as Chernobyl, Hurricane Katrina, the Ebola outbreak, most of our wars in the 20th century, and even the Spanish influenza would fail to constitute global catastrophes or existential threats. For this research community, their primary relevance is in what they can tell us about more severe possibilities within the relevant risk categories. On the other hand, various near-misses during the Cold War (Lewis et al, 2014)[17] might plausibly have led to a global thermonuclear war with global catastrophic consequences, and thus this topic is firmly in scope.

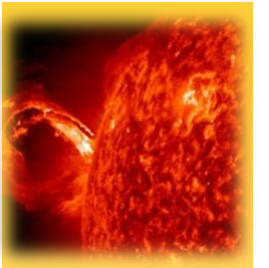
A clear example of an existential risk is the risk of an asteroid impact on the scale of that which wiped out the dinosaurs 66 million years ago (Schulte et al, 2010)[18]. An example of a global catastrophic risk that is not existential might include a large-scale pandemic disease outbreak

Previously, he ran the Oxford Martin Programme on the Impacts of Future Technology, and set up the FHI-Amlin Collaboration on Systemic Risks, both at the Future of Humanity Institute at Oxford.

Dr. Ó hÉigeartaigh's research spans technology policy and strategy, catastrophic risk, and horizon-scanning and foresight.

[16] "GCR Concept Project | Global Catastrophic Risk Institute."
[17] Dr. Patricia Lewis, Sasan Aghlani, and Benoît Pelopidas Heather Williams, "Too Close for Comfort."
[18] Schulte et al., "The Chicxulub Asteroid Impact and Mass Extinction at the Cretaceous-Paleogene Boundary."

(Palmer et al, 2017; Millett & Snyder-Beattie, 2017)[19]. Research and expert opinion indicates that it is unlikely that a natural pandemic outbreak could wipe out all humans across the globe given our distribution, immune variation, and other factors; and in most plausible scenarios global recovery seems likely in the longer term.



It is also important to consider factors that increase stress or resilience on a global scale – events or developments that in of themselves would not be a global catastrophe, but might make it more or less likely for a global or existential catastrophe to occur. Climate change has the potential to result in global catastrophic consequences at the more severe end of the possibility spectrum – e.g. 5 degrees and up (Wagner & Weitzman, 2016)[20]. But we might also consider less severe climate change as a stressor, as it could be expected to lead to major droughts and famines and other resource shortages, mass migration, geopolitical tension that could result in local or global war, and so forth. It could also lead to international conflict, for example over the use of controversial mitigation techniques such as sulphate aerosol geoengineering technologies.

More generally, many of the specific risks we will look at need to be placed in the context of a world with a rising population, rising resource footprint, more extreme weather events, increasing pressures on ecosystem services, a changing physical and electronic infrastructure, and changing geopolitical pressures - and a world with a range of technologies more powerful than any we've had in previous centuries.

III.1.2 TOPICS OF FOCUS WITHIN THE EXISTENTIAL AND GLOBAL CATASTROPHIC RISK RESEARCH COMMUNITY

Risks of particular focus within the existential risk community include those relating to our interaction with the global environment (for example, extreme climate change; globally catastrophic biodiversity loss; and risks from natural pandemic outbreaks). They also include risks from scientific and technological advances, such as risks from

[19] Palmer et al., "On Defining Global Catastrophic Biological Risks"; Millett and Snyder-Beattie, "Human Agency and Global Catastrophic Biorisks."
[20] Gernot Wagner and Martin L. Weitzman, "Climate Shock."

engineered pandemics and future advances in synthetic biology and other biotechnologies; nuclear winter; and risks from future advances in, and applications of, artificial intelligence. Some focus is given to risks from asteroids and supervolcanoes, but these have been less highly prioritised to date, due to the low frequency of occurrence of the relevant events; the last global catastrophe-level asteroid impact event was the impact that created the Chicxulub Crater 66 million years ago, and the last likely global catastrophe-level supervolcano is thought to have been the Toba eruption 70,000 years ago (Ambrose, 2000)[21].

The community also has research groups working on global systemic risks, interactions and cascades between risks, risks from future technologies such as neurotechnologies, advances in nanotechnology and other technologies. Researchers also focus on broader or more cross-cutting themes, such as governance challenges associated with global catastrophic risks, ethical considerations relating to the value of future generations, analysis of the costs and value of reducing existential and global catastrophic risk relative to other global priorities, foresight, horizon-scanning and road-mapping exercises for risks and relevant sciences and technologies, and methodological issues relating to reasoning about extreme events under great uncertainty.

For most individual global catastrophic and existential risks (with the possible exception of risk from advanced artificial intelligence), there are individual research communities working on these topics; albeit sometimes focusing more so on risks at a lower end of the spectrum. Some have been doing so for decades. For example, there are many centres doing valuable work on nuclear non-proliferation and security, pandemic preparedness and surveillance, and bioweapon governance – research leaders from these communities are represented at the Garrick Colloquium. There are numerous centres working on different aspects of climate change, biodiversity loss and other environmental risks and resource-related challenges. There are a number of excellent groups working on global systemic risks. NASA, the Planetary Defense community and others work on asteroid scanning and mitigation strategies. Furthermore, there are a range of centres in academia, think tanks, government and elsewhere working on more cross-cutting issues

[21] Rampino and Ambrose, “Volcanic Winter in the Garden of Eden.”

such as risk governance and international security, foresight and scenario planning, and so forth. At a level below global catastrophic risk, there is excellent work on risk modelling being done in, and associated with, the reinsurance industry.

The existential and global catastrophic risk research community collaborates with, and draws on, research from these communities extensively. It aims to complement such work by looking at existential and global catastrophic risks as a class of risks, with the aim of:

- Identifying the particular challenges associated with risks of this magnitude - whether they be scientific, analytic, ethical, or to do with governance, coordination, planning, or perception.
- Identifying risk areas where insufficient attention has been paid to the most extreme scenarios.
- Examining how these different risks, and other global developments, may interact with each other.
- Identifying previously unidentified or potential future risks
- Trying to distinguish which extreme scenarios are plausible and worthy of further work, even if they may be low probability, as opposed to those that can be dismissed as science fiction.

III.1.3 RECENT GROWTH OF THIS COMMUNITY

There has been a lot of recent growth within the existential risk research community. The first dedicated centre established was arguably Nick Bostrom's Future of Humanity Institute (FHI; https://www.fhi.ox.ac.uk/) in Oxford in 2003. The FHI has focused on cross-cutting global catastrophic risk analysis, philosophical analysis on the global importance of reducing existential risk, and in the last decade has placed its strongest focus on characterising potential risks from artificial general intelligence and superintelligence; a lot of this has been in collaboration with the Machine Intelligence Research Institute in Berkeley. The Global Catastrophic Risk Institute (GCRI; http://gcrinstitute.org/) in the US was founded in 2011, focusing on risks including bioweapons, nuclear war, artificial intelligence, and natural

events, and employing risk analysis methodology. Under Seth Baum and Tony Barrett's leadership, it has played a key role in establishing links between the existential risk community and experts in these fields.

The Centre for the Study of Existential Risk (CSER; http://cser.org) was founded in 2012 by Martin Rees, Huw Price, and Jaan Tallinn, although its first research grants were secured, and first postdoctoral researchers hired, more recently in late 2015 and 2016. We now have a research team beginning work on biological threats, extreme climate change, ecological tipping points, catastrophic risks related to future advances in artificial intelligence, and analysis of emerging technologies such as geoengineering. We also have postdocs working on more cross-cutting themes such as horizon-scanning and foresight for extreme risk, responsible innovation in risky sciences and technologies, population growth and resource use.

The Future of Life Institute (https://futureoflife.org/) was founded in 2014, focusing on artificial intelligence, climate change, risks from biotechnology, and risks from nuclear weapons. It has organised two highly successful conferences on the future of artificial intelligence and potential risks it might bring, resulting in a widely signed and shared open letter on the responsible development of AI, a grants programme to support work on AI safety, and a set of principles aimed at promoting the beneficial development and application of AI within the research community and more broadly. It has also organised a conference on nuclear war, and has engaged in activities to encourage divestment from nuclear weapons.

Several more recent initiatives are underway within academia, including at Stockholm, Warwick (United Kingdom), and Australia National University, and other world-leading risk centres such as the Garrick Institute (UCLA) are increasingly including global catastrophic risk within their remit, indicating that a diverse range of new expertise will be brought to bear on these topics in coming years.

III.1.4 WHAT CAN WE LEARN FROM THIS RESEARCH, AND WHAT CAN WE DO WITH THAT KNOWLEDGE?

It is worth considering the insights we can gain from studying risks of this magnitude that may not be gleaned from bodies of work analysing and mitigating risk more broadly, or risks of a lower magnitude.

One source of value is simply to be able to rule out scientifically implausible scenarios, in order to better focus our attention on existing threats and plausible future threats. This is of particular value for low-probability, extreme impact events, especially those that may be entirely novel.

Many researchers argue the value of understanding existential risk on the grounds of their long-term moral significance. If we can demonstrate that certain threats have the potential for human extinction, or permanent collapse, then it is argued that the importance of mitigating them increases dramatically, due to the fact that they would not only harm current generations, but wipe out the potential for a huge number of future lives.

There may be specific actions we could take that are designed to mitigate the most severe versions of these threats in particular. These might include establishing a very strict ban on specific types of virus research or bioweapons, or on research on AI systems that could both come up with a realistic model of the world and engage in recursive self-improvement. They could also include seed banks, shelters and alternative foods, so that even in many otherwise extinction-level events we might increase the odds of continuation of the species.

In addition, there may be strategies we would avoid adopting unless we had reasonable cause to believe we were headed for a world in which a particular type of global catastrophe were likely. For example, it would be extremely foolish and unconstructive to call for a ban on artificial intelligence research at this point. At some point in the future however, a body of evidence might indicate that certain developments with catastrophic potential are likely within several years. This then might be reason to consider a temporary moratorium on progress to enhance the capability of AI systems, while various containment and safety measures

were being explored. This is not without precedent, even at lower-levels of risk: the US White House issued a temporary moratorium on gain-of-function influenza research several years ago to allow time for more risk-benefit analysis (Lipsitch & Inglesby, 2014)[22].

Or consider a world in which evidence indicated we were headed for climate change of 5 degrees or over, in the absence of drastic action. In these circumstances, we might consider deploying technologies, such as sulphate aerosol geoengineering, which we might be hesitant to deploy under less severe scenarios (Crutzen, 2006)[23]. Our reservations might be based on concerns over risks posed by the intervention itself, thorny global governance challenges that the intervention presents, or questions over public acceptance. These may be sound reservations that would rule out these strategies in all but the most exceptional of circumstances. A research community developing indicators that we may be approaching unusually dangerous global circumstances, and developing strategies for last-ditch solutions, may be of considerable value given the risks we may face in the coming century – even if many of these strategies are never needed or deployed.

Lastly, by studying the particular scientific, governance, ethical, and communication issues that arise when confronting one global catastrophic risk, we can learn valuable lessons to draw on for future challenges. For example, climate change in some ways exemplifies a lot of the issues that make existential and global catastrophic risk especially challenging. There is still a lot of scientific uncertainty over timelines, probability and pathways to the most severe impacts. The scale and impact are difficult to grasp, impossible to see, and the most severe consequences will fall on future generations. We still don't have broad public acceptance of the science, especially in the US. It involves countries around the world coordinating, each making near-term sacrifices in favour of the longer-term future. Yet the Paris Agreement was extremely encouraging. It involved 194 countries committing to make sacrifices in the interests of future generations in the face of these challenges. Despite the setback of the US's recent announcement of

[22] Lipsitch and Inglesby, "Moratorium on Research Intended To Create Novel Potential Pandemic Pathogens."
[23] Crutzen, "Albedo Enhancement by Stratospheric Sulfur Injections."

intent to withdraw, this remains an important achievement and a critical step towards progress on climate change. By learning from the successes and failures of this process, as well as from the history of nuclear non-proliferation and international diplomacy, norms and conventions prohibiting the use of biological weapons, and other global processes to manage and mitigate global risk, we can be better prepared for the future.

III.1.5 KEY CHALLENGES IN EXISTENTIAL AND GLOBAL CATASTROPHIC RISK

There is a large body of expertise from the risk sciences that is of relevance to the study of existential and global catastrophic risk. However, a number of characteristics make these risks particularly challenging to analyse using normal risk analysis approaches. These include the difficulty in estimating the probability and expected impact of rare or even unprecedented events, where there may be sparse data to draw on; and the changing nature of some risks, in particular those associated with rapidly developing technologies (especially those interacting with a rapidly developing infrastructure).

For some global catastrophic risks, probability is relatively straightforward to quantify. For asteroid impacts, for example, we can look to sources of evidence such as the earth's fossil record, patterns of impact craters on the Moon and on Mars, datasets of asteroids passing our field of vision, and use these to make a reasonable estimate of the frequency with which we might expect an asteroid of a given size to hit the earth. The pathways by which an asteroid impact would result in global catastrophe are also relatively straightforward; therefore, we can estimate the expected global harm expected from asteroids of different sizes.

However, a similar analysis is much less straightforward for other risks. This can be illustrated by the example of catastrophic climate change. For a start, there is great uncertainty about the sensitivity of the earth system to the effects of our activities. The possibility of severe climate change is affected by factors including to but not limited to the potential for methane release from beneath the melting arctic permafrost, and from the seabed; how much CO2 the deep ocean can absorb; the

possibility of collapse of Antarctic and Greenland ice sheets; the possibility that the gulf stream might halt. The most worrying scenarios involve a combination of these factors driving each other as part of a positive feedback loop. While ongoing research will help us understand these factors and their interactions more clearly, it is very difficult to assign a meaningful probability to an outcome such as >5 degree climate change in the 21st century under a certain emission scenario. It may be that attempting to assign strict probabilities is the wrong approach; an alternative would be to develop frameworks of 'safe operating thresholds' with wide error bounds, exemplified by the 'Planetary boundaries' framework put forward by Johan Rockstrom and colleagues at the Stockholm Resilience Centre (Steffen et al, 2015)[24]. The expected harm from long-term climate change will also depend very heavily on the extent to which various mitigation and adaptation strategies are adopted, and how successful they are.

Similarly, in the case of global pandemics, we can ask scientific questions about the possibility of the 'perfect virus' with high health impact, high infectivity, and long incubation time. However, the scale and severity of impact will be predicated by many other factors – movement of humans or other vectors, capabilities of health services, the response of the population, and more. It is plausible that the bulk of the damage might not even be caused by the virus, but instead by a broad infrastructure collapse as emergency services and hospitals are overwhelmed, just-in-time food delivery is disrupted, and other systems underpinning societal order collapse.

However, none of these challenges are insurmountable. Modelling and analysis of factors that contribute to these risks can deepen our scientific understanding. This can help us establish estimates for probability and impact, and in some cases, rule out concerns entirely. Where past examples are sparse, there is value in drawing on counterfactual examples of 'near misses', as described by Gordon Woo (Woo, 2016)[25] and others. Design and analysis of scenarios can help in identifying key considerations and interactions. This may help us identify key interventions that reduce risk significantly, even if in instances

[24] Steffen et al., "Planetary Boundaries."
[25] Woo, "Counterfactual Disaster Risk Analysis."

where it is difficult to assign tight probabilities to events. Other risk analysis techniques, and interventions to mitigate risks at scales smaller that global catastrophe level, also provide a range of useful insights for the analysis of global catastrophic risks, as shown by the work of GCRI and others.

By studying global catastrophic risks as a class of risks, we can identify shared characteristics of global catastrophe events, which may help us identify common strategies that aid us in becoming more resilient as a species against a broad set of risks. For example, a number of global catastrophic events (global nuclear war, super volcano eruption, asteroid impact) would result in large amounts of particulate matter being ejected into the atmosphere, resulting in a disruption of photosynthesis (Maher and Baum, 2013) [26]. The development of alternative foods that are not dependent on sunlight, and strategies to scale up production of these food sources rapidly, would be robust in the face of a broad range of catastrophe events (Denkenberger and Pearse, 2014)[27]. The maintenance of permanent seed banks, manned shelters suitable for lengthy use, and information vaults represent similar safeguards. Similar strategies, useful for reduction of a broad range of risks, are likely to be feasible at the level of national and international governance (Farquhar et al, 2017; Cotton-Barrett et al, 2016)[28].

III.1.6 CASE STUDY: POTENTIAL FUTURE GLOBAL RISKS FROM ARTIFICIAL INTELLIGENCE

In this final section, I aim to present this research community's work over the last decade on potential future risks from artificial intelligence as a case study on how the field has made progress on a particularly novel area of concern, focusing on the broad strategies involved. This is complemented by a more technical analysis of AI risk provided by Seth Baum.

[26] Maher and Baum, "Adaptation to and Recovery from Global Catastrophe."
[27] Denkenberger and Pearce, *Feeding Everyone No Matter What*.
[28] Farquhar et al., "Existential Risk -- Diplomacy and Governance"; Cotton-Barrat et al., "Global Catastrophic Risks 2016."

The theoretical nature of the risk: Early work, carried out mainly by the Machine Intelligence Research Institute and the Future of Humanity Institute, aimed to explore the possibility of artificial general intelligence of greater-than-human-ability. This was coupled with the aim of exploring and characterising the theoretical risk associated that such a development could pose, drawing on input from experts in computer science and other fields. Rather than focus on difficult-to-characterise questions like consciousness, sentience, and evil intentions, the work instead focused on foundational issues like:

- The difficulty of designing safe goals for very capable, powerful systems able to take a very wide range of actions in a wide range of environments, where the actions could have a wide range of consequences.

- Certain predicted behaviours that might be expected from a powerful optimizing agent, such as a drive to acquire additional resources, a drive to avoid being switched off prior to completion of its goal, or a drive to improve the system's own capability.

- The theoretical possibility and limits of recursive self-improvement. This refers to the possibility that a sufficiently capable system may be able to surpass human programmers in its ability to design the next generation of systems. It is hypothesised that this could in turn result in a 'chain reaction' of performance improvement (Yampolskiy, 2015)[29], rapidly leading to a system far beyond human capability in most cognitive domains.

- The challenge that for most of the relevant design and control processes, it may be necessary to solve a range of technical and theoretical issues ahead of time, before certain critical thresholds in capability are reached. Beyond these thresholds, it may be much more difficult to intervene effectively due to the level of capability and autonomy of subsequent iterations of the system. It is worth noting that in principle, it is possible that systems may be developed which have a much greater ability to engage in science, engineering,

[29] Yampolskiy, "From Seed AI to Technological Singularity via Recursively Self-Improving Software."

and manufacture, and a greater ability to manipulate the global environment than we humans have. This level of power raises concerns of global catastrophic or existential risk, unless very carefully developed.

These programmes of research were then published in a book, Superintelligence (Bostrom, 2014)[30], which received a lot of attention, as well as in many further academic papers.

These lines of argument are by no means uncontroversial – many experts disagree on various points. Most experts consider artificial general intelligence to be decades out of reach as a scientific milestone, and some expect hundreds of years of progress to be needed (Grace et al, 2017)[31]. Some are sceptical about the hypothesis that rapid progress, enabled by the engagement of the AI systems themselves in the research and development process, could occur once a certain level of capability and generality is reached (e.g. see Walsh, 2016)[32]. Yet others are sceptical that such systems would be likely to demonstrate the traits of agency, autonomy and goal-driven behaviour that may make the actions of such systems difficult to predict or intervene on. Some experts hold that these is a limit to how much meaningful work can be done at this point in time to ensure the safety and stability of future systems, given the limits that can be meaningfully predicted about the theoretical underpinnings and engineering design of these future systems, as well as the limits of our knowledge regarding the nature of intelligence. Others have raised the concern that a focus on risks from powerful future systems may distract focus from more near-term risks and opportunities associated with the current state of the technology.

However, a growing body of experts in AI consider many of these concerns plausible and worthy of further study (e.g. see Dafoe and Russell, 2016) [33] . While the level of capability warranting global catastrophic concerns is still decades away or longer, ongoing research

[30] Bostrom, *Superintelligence*.
[31] Grace et al., "When Will AI Exceed Human Performance?"
[32] Walsh, "The Singularity May Never Be Near."
[33] Russell, "Yes, the Experts Are Worried about the Existential Risk of Artificial Intelligence."

on the matter is warranted by the magnitude of the challenge, and the range of technical and governance questions in need of study in advance of such developments. It is the view of this author that these concerns should not be used as a rationale to propose slowing down progress on the technology at this point in time. Nor should this research take the place of necessary and valuable work on near-term opportunities and challenges posed by AI, but rather should complement it.

Broader scientific engagement: The next steps for the community were to engage more deeply with the scientific artificial intelligence community in academia and industry– to explore, discuss and debate these arguments, as well as other risks that may be associated with AI. In 2015, a landmark conference was held in Puerto Rico, with representatives of the leading companies working explicitly towards a vision of general AI, alongside experts in governance, law, economics, risk and other relevant fields. The conference resulted in an open letter calling for more research on AI that was safe, robust, and beneficial, and for ongoing attention to issues relating to the longer term. The letter was signed by research leaders in artificial intelligence across industry and academia, as well as leading experts in a range of different fields, and was accompanied by a paper outlining research priorities, and a grants programme to support relevant work. The conference has been followed by a programme of activities to foster collaboration with more of the machine learning community both near- and long-term issues relating to safe design and risk. This has including a series of workshops organised by CSER and others at the major machine learning conferences (ICML, NIPS, IJCAI, DALI). T In 2017 FLI organised a follow-on to the Puerto Rico conference in Asilomar, resulting in the endorsement by many leading researchers of a set of principles for the long-term development of AI.

Technical research: In parallel, much of the work in the last two years has focused on translating some of the more foundational questions raised by early work at FHI and MIRI and elsewhere into crisp technical research problems that can be worked on today. This includes approaches involving fundamental mathematical frameworks for agent decision-making and behaviour, as well as research programmes exploring how some of the behaviours that would be of concern in long-term systems may manifest in the near-term systems we are building

currently. A number of research agendas have been published in the last year, and several of the leading companies focused on general AI now have safety teams exploring these issues. In addition, there has been substantial growth in projects mapping progress and trajectories in AI in domains relevant to both narrow and general artificial intelligence.

Global engagement: Within the existential risk community, more thought is now going to the particular political and governance challenges that may emerge as we move closer towards more powerful and general AI systems, with programmes starting up at FHI, CSER, the Centre for the Future of Intelligence. Long-term impacts and risks have risen on the agenda for governments in Europe and the United States, for example being mentioned as worthy of further study in the US Office for Science and Technology Policy's recent report on preparing for the future of artificial intelligence. One priority that has emerged is the need for greater global engagement, particularly with research leaders and other stakeholders in China, but also India, Japan, and emerging hubs in Africa. Any global conversation around the future of artificial intelligence, and potential global benefits risks associated with it, needs to have global representation. From a pragmatic point of view, China is on course to emerge as a scientific leader and agenda-setter in AI research over the coming decade. If at some point in several decades we truly do approach a level of technological breakthrough with global risk consequences, we are unlikely to be able to achieve a safe transition without a strong level of global cooperation and trust, which is going to require a lot of dedicated work to achieve. Now is a good time to start laying the groundwork.

III.1.7 CONCLUSION

We are entering a century in which humanity will be confronted with unprecedented threats to global civilisation. Some of these may result from the manner in which our footprint as a species strains our global environment, such as the impacts of climate change and biodiversity loss. Some may result from the development and deployment of increasingly powerful technologies, whether due to malevolent use or unintended consequences. The challenge posed by the analysis and mitigation of these threats requires an interdisciplinary approach: a community that can draw on the best expertise from the risk sciences,

as well as the expertise of scientists, law and governance specialists, ethicists, and others. It also requires a community that can draw on expertise on different types and sources of risk, and consider both lessons that can be applied across risks, and the interactions that are likely to occur between different global developments. Many of the most useful tools in global risk analysis have been drawn from the risk science literature, and deeper collaborations between the existential risk community and the risk science community are likely to be increasingly important in the years to come.

III.2 TOWARDS AN INTEGRATED ASSESSMENT OF GLOBAL CATASTROPHIC RISK (SETH BAUM, TONY BARRETT)

Integrated assessment is an analysis of a topic that integrates multiple lines of research. Integrated assessments are thus inherently interdisciplinary. They are generally oriented toward practical problems, often in the context of public policy, and frequently concern topics in science and technology.

This paper presents a concept for and some initial work towards an integrated assessment of global catastrophic risk (GCR). Generally speaking, GCR is the risk of significant harm to global human civilization. More precise definitions are provided below. Some GCRs include nuclear war, climate change, and pandemic disease outbreaks. Integrated assessment of GCR puts all these risks into one study in order to address overarching questions about the risk and the opportunities to reduce it.

The specific concept for integrated assessment presented here has been developed over several years by the Global Catastrophic Risk Institute (GCRI). GCRI is an independent, nonprofit think tank founded in 2011 by Seth Baum and Tony Barrett (i.e., the authors). The integrated assessment structures much of GCRI's thinking and activity, and likewise offers a framework for general study and work on the GCR topic.

III.2.1 ETHICS

Ethics is an appropriate starting point because ethical considerations motivate much of the attention that goes to GCR. Interest in GCR commonly follows from support for an ethics of expected value maximization:

$$EV(a) = \sum_{\{c\}} P(c) \int_s \int_t V(c,s,t) \partial t \partial s \quad (1)$$

In Equation (1), $EV(a)$ is the expected value of an action a that an actor (individual, institution, etc.) could take; {c} is the set of possible consequences of a; P(c) is the probability of consequence c; and V(c,s,t) is the value of consequence c at spatial point s and temporal point t,

Seth Baum is Executive Director of the Global Catastrophic Risk Institute, a non-profit think tank that Baum co-founded in 2011. His research focuses on risk and policy analysis of catastrophes such as global warming, nuclear war, and future artificial intelligence. Dr. Baum received an M.S. in Electrical Engineering from Northeastern University and a Ph.D. in Geography from Pennsylvania State University.

which is integrated across all points in space and time. V(c,s,t) is in turn defined as:

$$V(c,s,t) = U(c,s,t)D(c,s,t,) \quad (2)$$

In Equation (2), U is utility, which is commonly interpreted as welfare, quality of life, or something along these lines; and D is a discount factor that can have values within [0,1]; U and D can both vary across consequences, space, and time.

Each term in Equations (1)-(2) represents a distinct ethics concept. $EV(a)$contains the idea that ethics should be based on actions aimed at achieving the best outcomes, accounting for uncertainty about outcomes. $\sum_{\{c\}} P(c)$embodies the claim that the importance of a possible outcome is directly proportionate to the probability of its occurrence. $\sum_{\{c\}} P(c) \int_s \int_t V(c,s,t) \partial t \partial s$ captures the general notion that actions should aim to make the world a better place. $U(c,s,t)$represents whatever it is about the outcomes of actions that is considered to ultimately matter, an irreducible intrinsic value. Finally, $D(c,s,t,)$ accounts for the possibility that some things—specifically some units of utility—may be favored over others.

This is not the space to review the nuances of and arguments for and against these ethics concepts, which are all quite standard. However, it is worth briefly considering the discount factor. A case can be made for not discounting utility, i.e. valuing all possible utility equally regardless of which consequence it is associated with and where it occurs in space and time. Such a case is often made and can find rigorous ethical support, though, as with most ethics questions, it is not without detractors. Mathematically, it involves setting D =1 ∀ (c,s,t), in which case the righthand side of Equation 1 simplifies to expected utility. Throughout this paper, we will assume D =1.

Valuing all utility equally leads quite directly to consideration of GCR. If all utility is indeed valued equally, that means equality across all points in space and time, including spaces and times that are quite distant. Expected value maximization then benefits from a perspective that is global or even cosmic.

He then completed a post-doctoral fellowship with the Columbia University Center for Research on Environmental Decisions. In addition to his scholarly work, he writes frequently for popular media and is a featured columnist for the Bulletin of the Atomic Scientists.

Figure 1 shows three possible long-term trajectories for human civilization. The vertical axis is the total human utility summed across the human population alive at any particular point in time. The horizontal axis is time. Starting from the left, the curve shows a gradually increasing total utility as the human population grows and per capita quality of life improves (Figure 1 box "Us Now"). One can imagine total utility eventually leveling off; indeed, the world population is expected to peak later this century, and per capita quality of life may likewise reach a cognitive satiety. The plausibility and likelihood of these prospects can be debated, but this is not central to the main argument. All that is required here is the idea of human civilization persisting into the distant future in a form more or less like its current form (Figure 1 box "Status Quo").

Barring any other major changes, the status quo would eventually end in approximately one billion years (Figure 1 box "Earth Becomes Uninhabitable"). Despite the long-time horizon, this is not a particularly speculative claim. The physics is fairly well understood: the Sun will gradually grow warmer and larger, rendering Earth uninhabitable to life as we know it in approximately one billion years. The exact timing is less certain—it could be in two or three billion years, or perhaps other amounts of time—but this detail is not important to the main argument.

A global catastrophe that happens in upcoming years, decades, or centuries (i.e., within the typical time horizons of societal planning) would prevent humanity from enjoying that billion or so years left on Earth (Figure 1 box "Global Catastrophe"). This is clearly a very large loss of value: the area between the global catastrophe trajectory curve and the status quo trajectory curve.

But the value may be even larger. If humanity avoids global catastrophe, it could go on to do something much greater than the status quo, enabling much larger instantaneous total human utility (Figure 1 box "Something Big"). One possibility is space colonization, permitting much larger populations than can be achieved within Earth's carrying capacity. Another possibility is radical technological breakthrough, permitting much larger populations and/or higher per capita utility on Earth or beyond.

The prospect for humanity accomplishing something along these lines raises the stakes for global catastrophe. The value lost could be astronomically large and possibly even infinite. Infinite value could accrue if it is possible to persist for an infinite time within this universe, to travel to a different universe, or to survive via some other route, perhaps one that contemporary physics has not yet imagined. The physics of the infinite is less well understood. As long as the possibility of infinite value cannot be ruled out, such that it has a nonzero probability, then the expected value (Equation (1)) is infinite. Thus, actions to reduce GCR are, at least arguably, of infinite expected value.

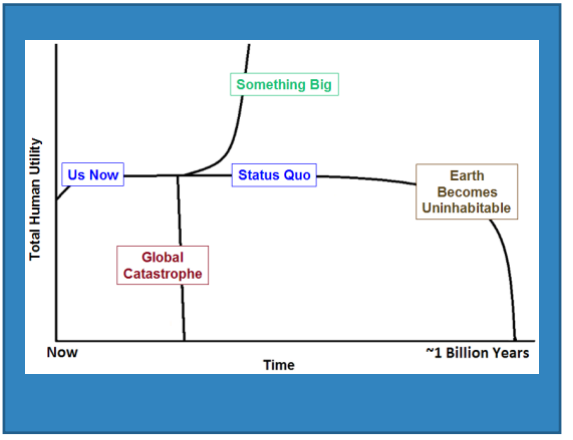


Figure 1: Possible long-term trajectories for human civilization. Adapted from Maher and Baum (2013).

What preceded is a simplified treatment of global catastrophe. Figure 2 shows more detail, depicting three different types of global catastrophes resulting in three distinct trajectories for human civilization. The first depicts global catastrophe quickly culminating in human extinction, after which total human utility is zero (Figure 2 box "Extinction"). This is the worst of the trajectories, in which all post-catastrophe utility is lost. There are even worse plausible scenarios in which a global catastrophe renders total human utility negative; these scenarios are beyond the scope of this paper.

The second trajectory shows some humans surviving the global catastrophe but in a diminished state, and then carrying on until Earth becomes uninhabitable (Figure 2 box "Survival Without Recovery"). This second trajectory can be thought of as the permanent collapse of human civilization. It likely involves large loss of population as well as a decline in per capita quality of life. The net effect is a large loss in total human utility relative to the status quo trajectory, comparable to but not quite as large as the extinction trajectory.

The third trajectory shows human civilization recovering back to the status quo after the global catastrophe (Figure 2 box "Recovery"). This is the most fortunate of the three global catastrophe trajectories. After a large initial decline, humanity makes it back to something along the lines of the large, advanced civilization that it currently enjoys. It could even go on to achieve something big, though likely with a delay relative to if no global catastrophe had occurred.

The lost value from the recovery trajectory depends on whether humanity goes on to achieve something big. If nothing more than the status quo would ever be achieved, with or without the global catastrophe, then the lost value from the global catastrophe is relatively small. To be sure, the "relatively small" here is still massive relative to most risks that get contemporary attention. The recovery curve in Figure 2 shows total human utility being reduced to a small fraction of the status quo level, which translates into billions of deaths and/or severe global immiserating.

Much more value would be lost from a delay in something big. Exactly how much depends on the relative long-term trajectories (the two curves labeled "Something Big" in Figure 2). Again, the physics here is not well understood. It is even possible that the no-catastrophe trajectory would remain larger than the catastrophe trajectory indefinitely, in which case the lost value would be infinite. Even if the loss is not infinite, it could still be astronomically large, though not as large as the losses in which humanity does not recover from the global catastrophe.

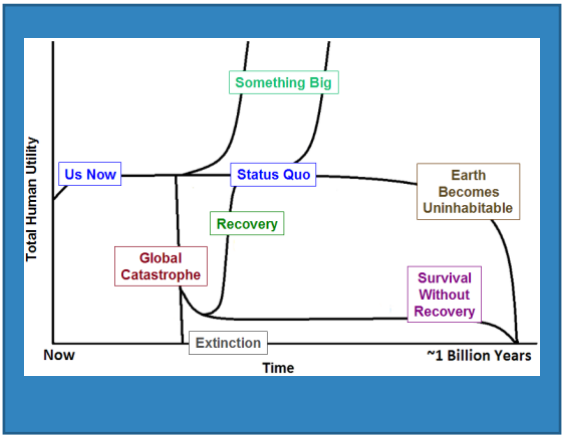


Figure 2: Possible long-term trajectories for human civilization showing different types of global catastrophe. Adapted from Maher and Baum (2013).

III.2.2 PRIOR LITERATURE

This is hardly the first scholarly analysis of GCR. The first were likely theological studies of Armageddon, end times, and related concepts. Perhaps the first scientific study came during the Manhattan Project. Prior to the first nuclear weapon test detonation, some of the physicists suspected that the explosion could ignite the atmosphere, killing everyone in the world. They conducted a study of the matter, finding that known physics rendered ignition very unlikely (Konopinski et al. 1946)[34]. Sure enough, they were correct, and that first nuclear explosion did not end humanity.

After World War II and especially with the buildup of nuclear arsenals, attention went to the prospect of nuclear war. It was commonly believed that a nuclear war with the large arsenals of the day would result in global catastrophe and possibly even human extinction. This led to some novel policy debates. One point of contention was the idea that it would be better to let the other side of the Cold War win than to let nuclear war end humanity. This debate took place in particular between

[34] Konopinski, "Ignition of the Atmosphere with Nuclear Bombs."

philosophers Sidney Hook and Bertrand Russell under the catchphrase "better red than dead" (Russell 1958a; 1958b; Hook 1958a; 1958b)[35].

In the 1980s, research on nuclear winter brought renewed attention to GCR. Nuclear winter is an environmental consequence of nuclear war, in which smoke from burning cities rises into the atmosphere and blocks incoming sunlight, disrupting agriculture and other important processes. Whereas the nuclear explosions of a nuclear war might only destroy the portion of the planet targeted in the war, leaving the rest of the world (including non-parties to the war) intact, the smoke of nuclear winter spreads worldwide, threatening populations everywhere. This prompted concerns that nuclear winter could cause human extinction. Carl Sagan cited the long-term significance of human extinction (essentially, Figure 2 box "Extinction") in arguing that nuclear winter made it much more urgent to address nuclear war risk (Sagan 1983)[36].

These discussions were not strictly academic. For example, at the height of the Cuban missile crisis, President Kennedy is said to have told a close friend, "If it weren't for these people that haven't lived yet, it would be easy to make decisions of this sort" (Schlesinger 1965/2002, p.819)[37]. Now, one can readily disagree with Kennedy: even if future generations are ignored, he was still facing an incredibly difficult decision. Or, phrased in terms of the underlying ethics, GCR can still be important even if one discounts future utility at a high rate, especially when one's actions can significantly affect the risk, as was clearly the case for Kennedy during the missile crisis. Still, it is notable that the ethics of future generations appears to have structured at least some of Kennedy's thinking during the crisis.

Another line of inquiry into GCR began during the 1970s with the rise of concern about environmental issues. This gave rise to an economics literature on environmental catastrophe (e.g., Cropper 1976)[38], which later led to literatures on the economics of catastrophic climate change

[35] Russell, "Freedom to Survive"; Russell; Hook, 26 May; Hook, 7-14 July.
[36] Sagan, "Nuclear War and Climatic Catastrophe."
[37] Schlesinger, *A Thousand Days*.
[38] Cropper, "Regulating Activities with Catastrophic Environmental Effects."

(e.g., Gjerde et al. 1999)[39] and on global catastrophes in general (e.g., Martin and Pindyck 2015)[40]. This economics literature brought a mathematical sophistication to the analysis of GCR, while continuing to emphasize issues of future generations, discounting, and significance for policy and decision making. However, the economics literature provides a rather crude treatment of the future, consisting mainly of simple mathematical assumptions extrapolated into the distant future with little regard for empirical considerations about what the future might actually look like.

Meanwhile, futurists from several disciplines have studied GCR with a greater attention to the nature of the future (Ng 1991; Tonn 1999; Bostrom 2002)[41]. This literature filled in empirical details such as the inhabitable lifetime of Earth and the long-term prospects for utility within the universe. Combining the mathematics from the economics literature with the empirical detail of the futures literature, one gets something along the lines of what is shown in Figure 2.

One common confusion in the GCR literature is to underestimate the importance of smaller catastrophes. An extreme case of this confusion is found in a much-cited passage of Parfit (1984, p.453-454)[42] that argues that human extinction is vastly more important than catastrophes killing 99% of the population, and indeed that the difference between extinction and 99% is much larger than the difference between 99% and 0 (i.e., no catastrophe). The problem with this logic is that it assumes that the surviving 1% would quickly recover back up to the status quo no-catastrophe state with no long-term loss in utility. However, as Figure 2 illustrates, this assumption does not necessarily hold, and indeed there is reason to believe that it often will

[39] Gjerde, Grepperud, and Kverndokk, "Optimal Climate Policy under the Possibility of a Catastrophe."
[40] Martin and Pindyck, "Averting Catastrophes."
[41] Ng, "Should We Be Very Cautious or Extremely Cautious on Measures That May Involve Our Destruction?"; Tonn, "Transcending Oblivion"; Bostrom, "Existential Risks."
[42] Parfit, *Reasons and Persons*.

not hold, in which case a 99% catastrophe could be of comparable loss as human extinction.

A similar and subtler case concerns smaller catastrophes involving "mere" millions or thousands of deaths. For example, Bostrom (2013)[43] dismisses the importance of the 1918 flu and the two world wars on grounds that they are not readily discernable when viewing the graph of total human population vs. time since 1900. The mistake here is to ignore the counterfactual: what matters is not whether these catastrophes are visible on a graph but whether they would have a long-term effect. Even a proportionately small loss can become extremely large or even infinite if it persists into the distant future. Such losses would still be smaller than the losses from larger catastrophes, but it would be a comparable loss, not something to dismiss as insignificant.

This last point raises the possibility that even small catastrophes involving just a few deaths could be comparable to the most extreme global catastrophes. Consider a decision between (A) a certainty of saving one human life, and (B) a one-in-ten-billion chance of preventing human extinction. Such a decision is quite plausible in the context of very low probability GCRs. The logic of Parfit (1984)[44] and Bostrom (2013)[45] point clearly in favor of (B). However, a complete consideration of possible consequences suggests that (B) is not obviously better and, depending on the details (e.g., which human life is to be saved), the decision could well fall in favor of (A). Exactly how this comparison should be resolved is has gone largely unexplored in the literature and remains an important open question.

III.2.3 TERMINOLOGY AND DEFINITIONS

Over the years, a large number of terms have been used to represent global catastrophe and related concepts. Table 1 provides a compilation.

[43] Bostrom, "Existential Risk Prevention as Global Priority."
[44] Parfit, *Reasons and Persons*.
[45] Bostrom, "Existential Risk Prevention as Global Priority."

Table 1: Terms used in the literature to represent global catastrophe and related concepts.

Term	Reference
Extermination	Russell (1958b)
Doomsday	Koopmans (1974)
Catastrophe	Cropper (1976)
Human extinction	Parfit (1984)
Oblivion	Tonn (1999)
Global catastrophe	Atkinson (1999)
Existential catastrophe	Bostrom (2002)
Survival	Seidel (2003)
Global megacrisis	Halal and Marien (2011)
Ultimate harm	Persson and Savulescu (2012)

At present, the two terms in widest use are "global catastrophe" and "existential catastrophe". A shortcoming of the term "existential catastrophe" is that it implies some sort of loss of existence, which could be the loss of the human species (i.e., human extinction) or the loss of human civilization. (The term is also found in other contexts, for example in business in reference to corporations that take on enough financial risk to threaten their ongoing solvency.) However, recalling Figure 2 and the surrounding discussion, what ultimately matters is not the existence of the species or the civilization but instead the long-term

trajectory. Indeed, Bostrom (2002)[46] defines existential catastrophe as an event that causes human extinction or permanently reduces its potential. Permanent reduction in potential captures some of the logic of long-term trajectories, though what matters is not the potential for long-term outcomes but the actual realization of them. Regardless, permanent reduction in potential is not "existential" in any meaningful sense of the word. Thus others (e.g., Tonn and Stiefel 2013)[47] have interpreted "existential risk" to refer strictly to human extinction risk. This is a more semantically sound interpretation, though, as discussed above, it excludes important risks.

The term "global catastrophe" does not suffer from the same semantic problem. The words can readily refer to the full range of catastrophes one might care about as per Figure 2. However, the term "global" is a spatial term that on its own does not capture the important temporal dimension of the consequences of catastrophes. Additionally, there is no clear threshold for what makes a catastrophe global. Even small catastrophes can be global—for example, a terrorist attack at a tourist venue killing one tourist from each continent is catastrophic to the deceased and their families across the globe. The GCR literature has assumed a higher severity for global catastrophe. Atkinson (1999) defines global catastrophe as an event in which at least one quarter of the human population dies; Bostrom and Ćirković (2008)[48] set a minimum threshold for global catastrophe in the range of 104 to 107 deaths or $109 to $1012 in damages. But these thresholds are arbitrary and do not signify any deeper reason for concern. Baum and Handoh (2014)[49] define global catastrophe as an event that exceeds the resilience of the global human system, resulting in a significant undesirable state change. This is a more meaningful definition, though it does not speak to long-term effects.

[46] Bostrom, "Existential Risks."
[47] Tonn and Stiefel, "Evaluating Methods for Estimating Existential Risks."
[48] Bostrom and Cirkovic, *Global Catastrophic Risks*.
[49] Baum and Handoh, "Integrating the Planetary Boundaries and Global Catastrophic Risk Paradigms."

Perhaps the most precise term would be "permanent catastrophe", defined as any event that causes a permanent reduction in instantaneous total utility. Such a term would capture the essential features of the expected utility calculus, including the possibility of nontrivial permanent effects of small catastrophes including single deaths. However, any of the terms in Table 1 should be fine. The GCR community is wise to avoid the contentious terminology battles that can be a major time sink for research fields. What ultimately matters is not which term is used but that the analysis is done correctly in order to accurately characterize the risks and the decision options for reducing them. It is to the analysis that the paper now turns.

III.2.4 INTEGRATED ASSESSMENT

The core questions to ask in GCR integrated assessment are: What are the risks? How big are they? What actions can reduce the risk? By how much? Answering these questions provides an understanding of the most important aspects of GCR. With answers to these questions, one can lay out the set of risks, the corresponding set of decision options, and an evaluation of it all in terms of expected value maximization (Equation (1)). This is the conceptual basis of GCR integrated assessment in simplest terms. (Some important refinements are discussed later in the paper.)

A complication for the expected value calculation comes from the extremely large magnitudes associated with the impacts of global catastrophes. As discussed above, the magnitudes could be astronomically large or even infinite. That makes the math more difficult. In response to this complication, Barrett (2017)[50] proposes a cost-effectiveness analysis of GCR reduction options. Adjusting slightly from the Barrett (2017) formulation, one can express GCR cost-effectiveness as follows:

$$ECE(a) = \frac{P_{gc}(*) - P_{gc}(a)}{C(a)}X \qquad (3)$$

[50] Barrett, "Value of Global Catastrophic Risk (GCR) Information."

In Equation (3), ECE(a) is the expected cost-effectiveness of action a; $P_{gc}(*)$ is the baseline probability of global catastrophe without the action; $P_{gc}(a)$ is the probability of global catastrophe with the action; $C(a)$ is the cost of the action, and X is the severity of global catastrophe. The Equation 3 formulation enables a simple comparison of different actions to reduce the probability of global catastrophe. Complications associated with the large severity of global catastrophe can be set aside because the variable X cancels out. Additionally, in including the cost of actions, Equation (3) enables consideration of budget constraints.

Some caveats are warranted. First, the variable X makes no distinction between global catastrophes of different severities. As discussed above, there can be important differences in the severities of different global catastrophes. Second, there is some debate about whether X does indeed cancel out if its value is infinite: whereas it is straightforward to state X / X =1 for finite X, it is not so simple for infinite X. A complete GCR analysis would account for both of these two issues, though they are beyond the scope of this paper.

If one accepts the Equation (3) formulation, the problem of selecting actions to minimize GCR takes the structure of a knapsack problem. In operations research and combinatorial optimization, the knapsack problem is the problem of selecting the highest value subset that fits within some constraint. One can imagine going on a trip and selecting items to put in a knapsack to take with. Should a large item be chosen, which is valuable but takes up all the space? Or should some combination of smaller items be chosen, which are each less valuable but may add up to something greater? Likewise, for GCR reduction, there are choices between actions of different cost and impact on the probability. Given a budget constraint (and budgets are in general constrained), the problem becomes one of selecting the subset of actions that minimizes the probability of global catastrophe while staying within the budget. This knapsack problem formulation provides a good starting point for understanding the analytical core of GCR integrated assessment.

III.2.5 RISK ANALYSIS

To begin filling in the details of the integrated assessment, the paper now turns to risk analysis. Table 2 lists some of the main GCRs, grouped into four broad categories: (1) environmental change driven by human activity, which is the generally unintentional side effects of large numbers of small actions in industry, agriculture, and other sectors; (2) technology disasters, which are the effects of misapplication of high-stakes technologies in which a small number of actions can have large global effect; (3) large-scale violence, in which harm is intentional; and (4) natural disasters, in which the source of the catastrophe is not human action. There are some GCRs that do not fit neatly into this categorization—for example, extraterrestrial invasion is sometimes considered as a GCR, which may not be caused by human action yet still may not qualify as "natural". That said, the categorization does cover most of the GCRs that are commonly considered.

Table 2: Four categories of GCRs and examples for each. Adapted from Baum (2015).

GCR Category	Examples of the GCRs
Environmental change	Climate change, biodiversity loss
Technology disasters	Artificial intelligence, biotechnology, geoengineering
Large-scale violence	Nuclear war, biological war, bioterrorism
Natural disasters	Pandemics, asteroid collision, solar storms

Identifying the GCRs is relatively straightforward; and the standard tools of risk analysis offer promise for analyzing them (Garrick 2008)[51], but

[51] Garrick, *Quantifying and Controlling Catastrophic Risks*.

fully quantifying them is not so easy. The GCRs are large, complex, and unprecedented, making for an unusually difficult risk analysis challenge (Baum and Barrett 2017)[52].

III.2.5.1 ASTEROID COLLISION

The challenge of GCR analysis can be seen clearly in the case of asteroid collision. Asteroid collision is perhaps the best understood and characterized of the GCRs. The underlying process is simple: a large rock hits Earth. The physical hazard is largely characterized via Newtonian mechanics. There is a substantial historical record of asteroid collisions, including the collision associated with the extinction of dinosaurs. There are also surveys of the current population of asteroids in the Solar System, thus far finding none on imminent collision course.

This corpus of empirical knowledge provides the foundation for asteroid risk analysis. Perhaps the most detailed study thus far is that of Reinhardt et al. (2016)[53]. Whereas most studies focus exclusively on asteroid diameter, this study considers the full range of physical parameters affecting collision severity: asteroid diameter, collision velocity, collision angle, asteroid density, and Earth density at collision point. Taking probability distributions across these parameters, the study calculates the probability of a "cataclysmic" collision, which it defines as a collision with energy of at least 200 megatons. Whereas prior studies found that cataclysm could only occur for asteroids of diameter one kilometer or greater, Reinhardt et al. (2016)[54] finds that cataclysm can occur for asteroids of diameter as small as 300 meters, and furthermore that most of the cataclysm risk comes from asteroids in the range of 300 meters to one kilometer, not from asteroids larger than one kilometer.

An important limitation of Reinhardt et al. (2016)[55] is that it uses a physical definition of event severity: the amount of energy released. The

[52] Barrett and Baum, "A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis."
[53] Reinhardt et al., "Asteroid Risk Assessment."
[54] Reinhardt et al.
[55] Reinhardt et al.

same limitation applies to many other asteroid risk analyses and analyses of other GCRs. (For elaboration in the context of environmental GCRs, see Baum and Handoh 2014.)[56] However, recalling the above discussion of ethics, what matters is not the physical severity but the human impacts.

It is not clear what the human impact of a 200-megaton asteroid collision would be, both in the immediate aftermath of the collision and for the long-term trajectory of human civilization. The same can be said for many other global catastrophe scenarios. Indeed, the aftermath of global catastrophes is the largest area of uncertainty in the study of GCR, as measured both in terms of how little is known and in terms of how important it is to the overall risk. The topic has also been poorly studied, with more research oriented toward the causes of catastrophes than toward their human effects. One should hope that humanity would quickly recover after even the most severe catastrophes, but this can hardly be guaranteed.

III.2.5.2 ARTIFICIAL SUPERINTELLIGENCE TAKEOVER

On the other end of the spectrum, a relatively difficult GCR to characterize is artificial superintelligence (ASI) takeover. ASI is AI with much-greater-than-human intelligence. Starting with Good (1965)[57], it has been proposed that ASI could use its intelligence to take control of the planet and the astronomical vicinity. Depending on the ASI design, this would cause either massive benefits or catastrophic harm, possibly including human extinction. The ASI does not need to be conscious or to have any formal intent with respect to humans—it just needs to act in ways that affect humans.

[56] Baum and Handoh, "Integrating the Planetary Boundaries and Global Catastrophic Risk Paradigms."

[57] Good, "Speculations Concerning the First Ultraintelligent Machine**Based on Talks given in a Conference on the Conceptual Aspects of Biocommunications, Neuropsychiatric Institute, University of California, Los Angeles, October 1962; and in the Artificial Intelligence Sessions of the Winter General Meetings of the IEEE, January 1963 [1, 46].The First Draft of This Monograph Was Completed in April 1963, and the Present Slightly Amended Version in May 1964.I Am Much Indebted to Mrs. Euthie Anthony of IDA for the Arduous Task of Typing."

ASI presents significant risk analysis challenges. No ASI currently exists, and there is no consensus on if or when it will be built. Technology forecasting is always a difficult proposition, all the more so for such a complex and unusual technology. The histories of AI and computing provide only limited insight, given their differences with ASI. Most extant AI is "narrow" in the sense that it is only intelligent within specific domains. For example, Deep Blue can only beat Kasparov at chess, not at the full space of problems. An ASI would likely be "general", with capabilities across a wide range of domains.

But these challenges do not render ASI risk analysis impossible. Indeed, established tools of risk analysis can be adapted to characterize ASI risk. Barrett and Baum (2017)[58] develop a fault tree model of ASI risk to identify the steps and conditions that would need to hold in order for ASI catastrophe to occur. This study looks specifically at ASI from recursive self-improvement, in which an initial AI makes a more intelligent AI, which makes an even more intelligent AI, iterating until ASI is built.

The fault tree contains two main branches:

(1) The ASI is built and gains capacity for takeover. This occurs if three subconditions all hold: (1a) ASI is physically possible, (1b) a "seed AI" is created and begins recursive self-improvement, and (1c) containment fails, meaning that there is a failure of efforts to either (1c1) prevent recursive self-improvement from resulting in ASI or (1c2) prevent the ASI from gaining the capacity for takeover.

(2) The ASI uses its capacity for takeover in a way that results in catastrophe. This occurs if three further subconditions all hold: (2a) humans fail in any attempts to design the goals of the ASI to not cause catastrophe, (2b) the ASI does not set its own goals to something that does not cause catastrophe, and (2c) the ASI is not deterred in carrying out its goals, whether by (2c1) humans, to the extent that human actions

[58] Barrett and Baum, "A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis."

might be able to deter an ASI, (2c2) another AI, including another ASI if this ASI is not the first, or (2c3) something else.

This distinction between 2c1, 2c2, and 2c3 is not in Barrett and Baum (2017)[59]. (The distinction between 1c1 and 1c2 is in the paper.) However, it could be readily added as an extension to the model. Indeed, one feature of this sort of model is that it enables a wide range of detail about ASI risk to be included in a clear and structured fashion. More generally, much of the value of the model comes from the process of laying out assumptions and seeing how they all relate to the risk. The graphical nature of fault tree models leads to clean visual depictions of the risk in order to help analysts and others make sense of it. (A graphic depicting the full model in Barrett and Baum (2017)[60] can be found online at http://sethbaum.com/ac/2017_AI-Pathways2full.png.)

While the model can also be used to quantify risk parameters as well as the total risk, such quantifications will often be uncertain due to the inherent ambiguity of ASI risk. This ambiguity poses a challenge for attempts to calculate optimal decision portfolios for minimizing GCR, such as in the knapsack problem described above. However, some of this challenge is attenuated by the details of the decision options themselves, to which the paper now turns.

III.2.6 RISK REDUCTION IN RESEARCH

Recalling the ethics of expected value maximization, what matters is not the risks themselves but the opportunities for reducing them. Large risks do not necessarily offer better risk reduction opportunities. Possible actions could have a small effect on a large risk, or they could be expensive, giving them a low expected cost-effectiveness. Likewise, GCR integrated assessment requires risk analysis, but it also requires analysis of risk reduction opportunities.

Table 2 lists some examples of actions that can reduce risk for each of the four GCR categories that were introduced in Table 1. These actions show the value of grouping the GCRs into these categories: the same

[59] Barrett and Baum.
[60] Barrett and Baum.

actions are often applicable across multiple GCRs within the same category:

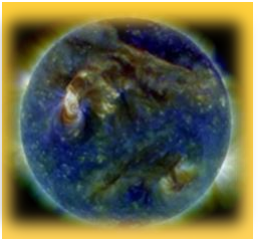
(1) A large portion of environmental change GCR is driven by energy and agriculture. This GCR can be reduced by via actions such as energy conservation, switching to energy with low carbon emissions, and shifting away from animal-based diets. This holds for risk from climate change, biodiversity loss, ocean acidification, depletion of freshwater and phosphate, among other global environmental risks. An exception is the global spread of toxic industrial chemicals, which derives mainly from other industrial processes.

(2) Technology disasters can often be avoided by making the technology design safer, for example by designing an ASI with safe goals (item (2a) in the ASI fault tree described above). These design details are specific to each technology. However, regimes for technology governance can cut across technologies. For example, Wilson (2013) [61] develops a proposal for an international treaty covering all GCRs from emerging technologies. The treaty would standardize precautionary decision-making principles, laboratory safety guidelines, oversight of scientific publications, procedures for public input, and other issues that cut across technologies.

(3) The risk of large-scale violence can often be reduced via arms control, i.e. via restrictions on the procurement and use of weapons. Some aspects of arms control are specific to certain weapons and/or certain actors, such as the New START treaty restricting nuclear weapons for the United States and Russia. Other aspects are more general, such as the Conference on Disarmament, an international forum for arms control and disarmament. Additionally, the risk of large-scale violence can be reduced by improving international relations and resolving conflicts without war. The same can also hold for terrorist groups and other nonstate actors, ideally so that they do not feel the need to cause or threaten violence in the first place. Progress in

[61] Wilson, "Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law Note."

improving relations and meeting needs peacefully reduces the risk of all types of large-scale violence.



(4) Some natural disasters can be prevented. For example, there are proposals to avoid asteroid collision by deflecting asteroids away from Earth. The prevention measures are generally risk-specific. When disasters cannot be prevented, the primary means for risk reduction is to increase society's resilience to the disaster, so that initial losses are relatively small and civilization can recover (as in Figure 2 box "Recovery").

Table 2: Examples of GCR reduction actions for each of the four GCR categories.

GCR Category	Examples of the GCRs
Environmental change	Climate change, biodiversity loss
Technology disasters	Artificial intelligence, biotechnology, geoengineering
Large-scale violence	Nuclear war, biological war, bioterrorism
Natural disasters	Pandemics, asteroid collision, solar storms

III.2.6.1 RISK-RISK SYNERGIES: SOCIETAL RESILIENCE

The risk reduction action of increasing societal resilience is an important one and worth discussing in further detail. It was brought up in the context of natural disaster risk, but it is applicable across a wide range of GCRs. Indeed, the only GCRs for which societal resilience is not helpful are those in which humanity goes extinct from the initial disaster. Only a small portion of GCRs would result in immediate extinction; these include physics experiment disasters, which could destroy the

astronomical vicinity, and ASI, which might kill all humans in pursuit of its goals regardless of any human resistance. But for most GCRs, the risk can be reduced by increasing societal resilience. Actions to increase societal resilience thus have strong risk-risk synergy for GCR: the same action can reduce multiple GCRs.

Broadly speaking, there are two ways to increase societal resilience to GCRs. The first is to enable human civilization to stay intact during the catastrophe. This includes measures such as increasing spare capacity in supply chains (as opposed to "just-in-time" supply chains with minimal spare capacity) and hardening critical infrastructure to withstand disasters. Some of these measures are specific to certain GCRs. For example, electric grid components can be hardened to withstand solar storms or nuclear electromagnetic pulse attacks, but this would not help against other GCRs. However, many of the measures are widely applicable across GCRs. For example, many GCRs could result in supply chain disruptions, due to some combination of damage to manufacturing facilities, suspension of shipping, and loss of labor. For all these GCRs, spare capacity in supply chains can enable the continuity of manufacturing and the provision of goods and services.

To develop measures for keeping human civilization intact during and after global catastrophes, it is important to have a systemic understanding of human civilization. There are often key nodes in the networks of physical infrastructure and human society that constitute human civilization. For example, transformers are key nodes within electricity networks; ports are key nodes within transportation networks. An emerging field of global systemic risk is mapping out global systems, assessing ways in which initial disturbances can propagate and cascade around the world, and identifying weak points and opportunities to increase resilience (Centeno et al. 2015)[62].

The second way to increase societal resilience to GCRs is to increase local self-sufficiency to aid survivors in the event that global human civilization fails. Again, the measures that can be taken often apply widely across GCRs. For example, several GCRs pose direct threats to global agriculture, including nuclear war, asteroid collision, and volcano

[62] Centeno et al., "The Emergence of Global Systemic Risk."

eruption, each of which block sunlight ("nuclear winter", "impact winter", and "volcanic winter"). Other GCRs threaten global food supplies in other ways, for example by disrupting supply chains. In the face of food supply catastrophes, local self-sufficiency can be enhanced via food stockpiles and alternative methods for growing food locally (Denkenberger and Pearce 2014; Baum et al. 2015)[63].

Both ways to increase societal resilience to GCRs feature extensive synergies across risks: the same action will reduce the risk of many different GCRs. And resilience measures are not the only ones to have this feature. Some other such measures (discussed above) are clean energy and agriculture, which reduce risk from several environmental GCRs. These synergies reduce some of the pressure on quantifying the risk: if an action reduces the risk for two different risks, the relative size of these two risks is less crucial. That said, the size of the risks remains important for comparing the value of different actions.

III.2.6.2 RISK-RISK TRADEOFFS: ARTIFICIAL SUPERINTELLIGENCE TAKEOVER

In addition to risk-risk synergies, in which one action reduces multiple risks, GCR reduction also often has risk-risk tradeoffs, in which an action reduces one risk but increases another. Evaluation of these actions is highly sensitive to risk quantification. Depending on how the risks are quantified, the action could even be found to cause a net increase in the risk.

An important example of risk-risk tradeoff in GCR involves ASI takeover. As discussed above, the ASI takeover itself could cause global catastrophe if its goals are unsafe. Alternatively, if its goals are safe, then it may help prevent other global catastrophes. Additionally, if the ASI is contained such that it does not (and cannot) take over, then the outcome could depend on how the ASI is used by whichever humans has it contained. It might be used malevolently, causing global catastrophe. Or, it might be used benevolently, avoiding other global catastrophe.

[63] Denkenberger and Pearce, *Feeding Everyone No Matter What*; Baum et al., "Resilience to Global Food Supply Catastrophes."

These possible outcomes should be factored into any decision of whether or not to launch an ASI, or a seed AI that could become an ASI. This means that the launch decision depends not just on the riskiness of the ASI itself, but also the extent of other risks—essentially, how risky it would be to not launch the ASI. Because ASI could provide unprecedented problem-solving ability across a wide range of domains, it might offer extensive reduction to a wide range of GCRs. This creates a great dilemma for those involved in the launch decision, the dilemma of whether or not it would be safer to launch the ASI (Baum 2014)[64].

III.2.7 SYSTEMIC INTEGRATED ASSESSMENT

The various interconnections between GCRs and actions to reduce GCRs suggest a refinement to the concept of integrated assessment. Instead of listing the risks and their corresponding risk reduction measures and analyzing each of them in isolation, it is better to analyze systems of risk and risk reduction measures. Thus, the core questions posed above can be rephrased: What are the systems of risk? How big are they? What suites of actions can reduce the total risk? By how much? Answering these questions provides a better understanding of GCR. These suites of actions can then be assessed in terms of their expected value or expected cost effectiveness.

III.2.8 RISK REDUCTION IN PRACTICE

Ultimately, what is of interest is not the analysis of GCR or the evaluation of GCR reduction measures—it is the actual reduction of GCR. In other words, GCR integrated assessment should be oriented towards risk reduction in practice; it should not just be an academic exercise. Broadly speaking, there are at least three approaches to GCR reduction: direct, indirect, and very indirect. Each of these is applicable in certain contexts.

III.2.8.1 THE DIRECT APPROACH

The direct approach involves presenting the results of risk analysis directly to decision makers, who then take the analysis into account in their decision making so as to reduce the risk. The direct approach is

[64] Baum, "The Great Downside Dilemma for Risky Emerging Technologies."

perhaps the most familiar one for risk management, and the most idealistic in the sense that it describes an ideal risk management process.

The direct approach does sometimes work. For example, Mikhail Gorbachev reports that he was influenced by research on nuclear winter to act to reduce nuclear weapons risk (Hertsgaard 2000)[65]. Gorbachev's case shows the potential for GCR research to speak to the highest levels of power. To be sure, the effort to draw attention to nuclear winter research was greatly aided by Carl Sagan at the height of his public popularity. Still, there are many other examples, some much more mundane but nonetheless important, of GCR research directly influencing decision making. Indeed, there are entire risks, climate change among them, that would be scarcely recognized if not for the efforts of research communities to study and present findings about the risk.

That said, the direct approach often does not work. One reason is differences in ethics. Simply put, not everyone agrees with the ethics of undiscounted expected utility maximization. The more people discount—the more parochial their concerns—the less they are likely to care about GCR. They may be even less likely to care about GCR if they are not trying to maximize value in the first place. Value maximization is associated with consequentialist ethics, yet moral philosophy recognizes other types of ethics, including deontology (ethics based on rules for which types of actions are required or forbidden) and virtue (ethics based on the character of the person). And many people do not pursue any formal set of ethics such as those found in moral philosophy. Unless people are seeking to maximize value, then the extremely large values associated with GCR may be less persuasive.

Another reason that the direct approach may not work is that people do not always want to hear the findings of risk analysis. People may be motivated by cultural, political, or economic factors to ignore risk analysis or reject its findings. Indeed, there is a growing cultural tendency to dismiss all types of expert analysis as elitist, unnecessary, or

[65] Hertsgaard, "Mikhail Gorbachev Explains What's Rotten in Russia."

otherwise unwanted (Nichols 2017)[66]. In the context of GCR, this phenomenon can be seen, for example, in the rejection of the scientific consensus on climate change, which is a major impediment to advancing climate policy, and in the rejection of expert advice to use vaccines, which could enhance the spread of pandemics.

III.2.8.2 THE INDIRECT APPROACH: MAINSTREAMING

When the direct approach does not work, one option is to go indirect via a technique called mainstreaming. The technique was developed by the natural hazards community in response to populations that could not be directly motivated to act on natural hazards even when they are quite vulnerable. The natural hazards community found that populations often had other priorities, such as those related to economic development. So, the natural hazards community integrated natural hazards into those other priorities. Thus, to mainstream is to integrate a low-priority issue into a high-priority issue, thereby bringing it more mainstream attention.

Mainstreaming has been successful for natural hazards, and it can also be successful for GCR (Baum 2015)[67]. For example, the 2014 Ukraine crisis brought increased interest in relations between the United States and Russia. This created opportunities to draw renewed attention to nuclear war risk. The risk was a major focus of attention throughout the Cold War, but since then had largely faded from the spotlight. It was commonplace to believe that nuclear war risk ended with the end of the Cold War, but sure enough, the weapons still exist in large number, and United States-Russia tensions had not been fully resolved. The Ukraine crisis exposed this, creating an opportunity for discussion of a wide range of nuclear weapons issues, including those not directly related to the United States-Russia relationship. Additional opportunity is created by the alleged intervention by Russia in the 2016 United States election. There is a growing sense that the Cold War is back, which, for better or worse, means improved opportunities to draw attention to nuclear weapons issues.

[66] Nichols, *The Death of Expertise*.

[67] Baum, "The Far Future Argument for Confronting Catastrophic Threats to Humanity," 201.

Another example involves AI. ASI remains more of a fringe topic, especially in policy circles, which tend to focus more on near-term technologies. However, AI is an increasingly important near-term policy issue. One of the most important AI policy topics is the unemployment that could be caused by the mass automation of jobs. Unemployment is commonly a top-priority policy issue. While much of the current political discourse on unemployment emphasizes globalization, immigration, and labor policy (e.g., minimum wage), automation is already a significant factor and is poised to become perhaps the dominant factor. Indeed, an ASI may be able to perform nearly any job, especially if paired with the robotics that it may be able to design. Of course, if the ASI kills everyone, then unemployment is a moot point. Still, it remains the case that ASI risk can be mainstreamed into conversations about unemployment.

III.2.8.3 THE VERY INDIRECT APPROACH: CO-BENEFITS

Another approach is even more indirect. It involves emphasizing co-benefits, which are benefits of an action that are unrelated to the target issue. For GCR, the co-benefits approach means emphasizing benefits of an action that are unrelated to GCR (Baum 2015)[68]. To execute this approach, one need not even mention GCR. Thus, the co-benefits approach can work even when there is complete indifference to GCR.

Perhaps the most fertile area for co-benefits is the environmental GCRs, where a plethora of co-benefits can be found. For example, quite a lot of energy can be conserved when people walk or bicycle instead of driving a car, which is also an excellent way of improving one's personal health. Diets low in animal products are also often healthier. Reducing energy consumption saves money. Living in an urban area with good options for walking and public transit enables an urban lifestyle that many find attractive, which in part explains the high real estate costs found in many high-density cities. Emphasizing these and other co-benefits can enable a lot of environmental GCR reduction, even when people are not interested in the environmental GCRs.

[68] Baum, "The Far Future Argument for Confronting Catastrophic Threats to Humanity."

Another important case for co-benefits is in electoral politics. It is often the case that a particular candidate or party would be better for reducing GCR. But the GCRs are often not priority issues for voters. Instead of trying to convince voters to care more about GCRs, it can be more effective to motivate them to vote based on the issues that they already care about. For example, in the United States, support for climate change policy often falls along party lines, with Democrats in support of dedicated effort to reduce emissions and Republicans opposed. But climate change is not typically a top issue for voters. Therefore, one could reduce climate change GCR by supporting Democrats based on the issues that voters care about. (Whether or not Democrats or other politicians should in general be supported depends on more than just their stance on climate change—it also depends on their stances on other GCRs, and perhaps on other factors as well.)

III.2.8.4 STAKEHOLDER ENGAGEMENT

A running theme across all three approaches to GCR reduction is stakeholder engagement: the process of interacting with stakeholders to share about GCR and hear their perspectives. The stakeholders are anyone who plays an important role in GCR decisions, including elected officials, citizens, business leaders, and technologists, among others.

Stakeholder engagement should be a two-way dialog. Results of GCR integrated assessment research should be shared with stakeholders so that they can be taken into consideration, as in the direct approach to GCR reduction. Additionally, it is important for researchers to listen to the stakeholders in order to learn their options, preferences, constraints, and perspectives on GCR in general and especially on the GCR reduction actions that they could take.

Insights from stakeholders should then be fed back into GCR integrated assessment research. If certain stakeholders are not able to take certain actions, for example due to institutional or cultural constraints, then those actions can be excluded from further analysis. Alternatively, if stakeholders can take the actions, but are less inclined to do so, then this increases the cost of the action by requiring extra resources (be it money, personnel time, or something else) to motivate them. This all factors back into the integrated assessment, and can be plugged directly

into the knapsack problem of identifying the suite of decision options that minimizes GCR.

III.2.9 CONCLUSION

Given the goal of expected value maximization, especially when value is defined as undiscounted utility, GCR reduction is an important priority. GCR integrated assessment can answer overarching questions about GCR, above all which actions or suites of actions can best reduce the total risk. This paper has presented a concept for GCR integrated assessment developed by the Global Catastrophic Risk Institute. It calls for quantification of GCRs and actions to reduce GCR in terms of expected value, accounting for systemic interactions, and conducted with two-way stakeholder engagement to factor in stakeholder perspectives and share assessment results. This integrated assessment concept aims to address GCR in a fashion that is both intellectually sound and practical.

III.2.10 ACKNOWLEDGMENTS

This paper was presented at the International Colloquium on Catastrophic and Existential Risk, held at UCLA during 27-29 March 2017. We thank colloquium participants and especially John Garrick for very productive discussion on this paper and related topics. Any errors or other shortcomings in this paper are the authors' alone.

III.3 NUCLEAR TERRORISM AND NUCLEAR PROLIFERATION (ALBERT CARNESALE)

Albert Carnesale is Chancellor Emeritus and Professor Emeritus at the University of California, Los Angeles (UCLA). He joined UCLA in 1997, and was Chancellor of the University through 2006 and Professor of Public Policy and of Mechanical and Aerospace Engineering through 2015.

The term "nuclear terrorism" encompasses several ways in which terrorists might make use of nuclear devices or radioactive materials. Most destructive among these is detonation of a nuclear weapon. (For purposes of this discussion, a nuclear weapon is any explosive device that derives its energy from nuclear reactions.) Far less destructive than terrorists' use of a nuclear weapon, but probably more likely, is dispersal of radioactive material over a wide area by sabotaging a nuclear facility or by setting off a "dirty bomb" (i.e., a device containing chemical explosives and radioactive material). This discussion focuses on the form of nuclear terrorism that is of greatest concern—detonation of a nuclear weapon.

A nuclear weapon must contain either highly enriched uranium (HEU) or plutonium. Uranium that is found in nature comprises two isotopes: U-238 and U-235. To be useful as fuel in a typical nuclear power reactor, the uranium must be "enriched" to increase the proportion of U-235 from 0.7% to about 3-5%. In weapon-grade uranium, HEU, the proportion of U-235 is on the order of 90%.

Plutonium is not found in nature, but can be produced in a nuclear reactor, separated from the remainder of the spent nuclear fuel by chemical reprocessing, and then either recycled as nuclear fuel or used for weapons. Enrichment and reprocessing facilities are dual-use facilities; that is, they can be used for both civilian and military purposes.

At present, only nations have the capacity to produce HEU or plutonium. Moreover, even if terrorists were to acquire some of a nation's HEU or plutonium, they would still face the formidable task of assembling a nuclear weapon. A far less difficult path would be to buy, steal, or be given a nation's nuclear weapon. Accordingly, the most promising avenues for preventing nuclear terrorism are stemming the proliferation of nuclear weapons to additional countries and enhancing the security of nuclear weapons and nuclear materials.

Nine countries now have nuclear weapons: the U.S., U.K., Russia, France, China, India, Pakistan, North Korea, and Israel (though Israel

neither confirms nor denies the existence of its nuclear arsenal). In addition, six non-nuclear-weapons-states have either enrichment or reprocessing capabilities: Argentina, Belgium, Germany, Iran, Japan, and the Netherlands.

The principal bulwark against the proliferation of nuclear weapons is the Nuclear Non-Proliferation Treaty (NPT). Almost all of the nations in the world are parties to the NPT; the exceptions being Israel, India, Pakistan, North Korea, and South Sudan (whose civil war has understandably delayed its ratification of the Treaty).

Key provisions of the NPT include: each nuclear weapons state (NWS) agrees not to help others acquire nuclear weapons; each non-nuclear-weapons-state (NNWS) agrees not to pursue nuclear weapons; a safeguards regime for verification of the Treaty is established; the "inalienable right" of all parties to pursue nuclear energy for peaceful purposes is acknowledged; each of the parties is obligated "to pursue negotiations in good faith on effective measures relating to cessation of the nuclear arms race at an early date…"; and a procedure for withdrawal from the Treaty after three months' notice is provided.

When the NPT entered into force in 1970, there were five (possibly six) NWSs: U.S., U.K., U.S.S.R. (Russia), France, China and (possibly) Israel. Today there are nine, with India, Pakistan, and North Korea having joined the club. Few, if any, observers in 1970 would have predicted so small an increase in nuclear weapons states over the ensuing 47 years! Hard work has paid off—at least thus far.

Where might terrorists get their hands on a nuclear weapon or on the materials to make them? The countries that readily come to mind are Iran, Russia, North Korea, and Pakistan.

The nuclear agreement between Iran, the U.S., Russia, China, France, the U.K., and Germany—formally known as the Joint Comprehensive Plan of Action, was implemented last year. It extends the time interval required for Iran to produce a nuclear weapon from two months (or less) to one year (or more). It also restricts Iran's enrichment activities and prohibits its production of plutonium. It appears that even the

Albert Carnesale's research and teaching continue to focus on public policy issues having substantial scientific and technological dimensions, and he is the author or co-author of six books and more than 100 articles on a wide range of subjects, including national security strategy, arms control, nuclear proliferation, domestic and international energy issues, and higher education.

harshest critics of the deal, including President Trump, have concluded that the agreement is worth keeping—for now.

Russia appears on the list of would-be sources of nuclear weapons or materials largely because it has about 6000 nuclear weapons, stockpiles of HEU and Pu large enough to produce about 100,000 more weapons, and a history of securing those weapons and materials at less-than-acceptable levels.

North Korea withdrew from the NPT in 2003, and since then has conducted five nuclear weapons tests. It is estimated that North Korea has tens of nuclear weapons, produces enough plutonium annually for another 1-2 weapons, and has substantial uranium enrichment capabilities. As former Secretary of Defense Robert Gates and others have observed, North Korea "will sell anything to anybody."

Pakistan is on the list not only because it has about 150 nuclear weapons and sufficient HEU and Pu to produce an additional 200 or so, but also because it is not a member of the NPT and is a relatively fragile state. The prospect of a failed state with so many nuclear weapons widely deployed is a clear cause for concern.

In light of all this, what is to be done to reduce nuclear risk? My high priority list is as follows:

- Minimize proliferation of nuclear weapons
- Minimize spread of enrichment and reprocessing
- Enhance security of nuclear weapons and materials
- Reduce stockpiles of nuclear weapons and materials
- Maintain the U.S. commitment to hold fully accountable any state or non-state actor that enables terrorist efforts to obtain or use weapons of mass destruction—and encourage other nations to make similar commitments.

III.4 BIOLOGICAL TERRORISM AS AN EXISTENTIAL RISK (PETER KATONA)

III.4.1 THE THREAT

Peter Katona is Clinical Professor of Medicine in Infectious Diseases at the David Geffen School of Medicine at UCLA and Adjunct Professor of Public Health at the UCLA Fielding School of Public Health. He has worked at the Centers for Disease Control and Prevention as an EIS Officer studying viral diseases and doing epidemic investigation, and at Apria, Corum and CVS as their Corporate Medical Director.

Whether from terrorism or antibiotic resistance or a biological accident, is deadly disease an existential threat, a catastrophic threat, or, at a minimum, a threat to national security? With earth's 7.4 billion people, 20 billion chickens, and 400 million pigs, we have the ideal scenario for creating and spreading deadly microbes from and between animals and plants. Trade and travel connect points of the globe in a matter of hours not days or weeks. More and more people are living in the microbe rich mega-city slums of the developing world such as Lagos or Mexico City where a small outbreak can spread into a pandemic even without a terrorist's hand.

Classical threats to humanity are abundant. They include climate change and environmental degradation, armed conflict, organized crime, terrorism of all types, poverty, weapons of mass destruction (WMD), corruption, energy concerns, emerging and re-emerging infectious diseases, fanaticism, fundamentalism, natural disasters such as earthquakes, pandemics and tsunamis, the consequences of poor leadership, and today's "fake" news, which is eroding our understanding of objective facts, especially in science.

The narrower category of existential threats includes a large asteroid colliding with the earth, eruption of the caldera under Yellowstone National Park, a global pandemic of a highly pathogenic and transmissible infectious agent, or thermonuclear holocaust among others. These threats can be intentional, accidental or naturally occurring events.

Biological agents are living organisms such as bacteria or viruses that may have disease-causing properties. Biological weapons are living biological agents or their toxins intentionally causing harm to humans, plants or animals. They may be easily obtained, and, with the right equipment and technical skills, grown relatively easily. They typically infect a host with a small amount of product. Biological weapons can be

contagious or non-contagious, and can be easily disseminated, resulting in high morbidity and mortality.

Natural emerging and re-emerging biological agents are constantly evolving. This may be caused or aggravated by climate change, antibiotic resistance, the overuse or misuse of antibiotics, new or changing disease vectors, and new habitats. For example, the suspected origin of the 2016 Ebola outbreak in West Africa might have been a hole in a tree where toddlers liked to play[69]. This may have precipitated contact with infected bats. SARS came from civet cats caged at an Asian meat market.[70] There is also the case of a woman who took a trip to Uganda in 2007 where she visited a bat cave and was subsequently diagnosed with Marburg virus, a relative of Ebola.[71]

There are different kinds of biological events: the intentional terrorist attack, the accidental release, and the natural outbreak. All may constitute public health emergencies. Intentional attacks can be biological crimes targeted against individuals, acts of bioterrorism making a violent political statement against a target audience, biowarfare between or by nations, or agroterrorism directed against animals or crops.

To accurately assess the importance of offensively used biological agents we have to look at the likelihood of occurrence regarding their transmissibility and destructive capabilities. Highest on this list are the CDC category A agents such as anthrax, smallpox, tularemia, the hemorrhagic fever viruses, and plague.[72] CDC Category B and C agents were considered less important. Influenza was not even on this list.

How much damage can a weaponized biological agent actually do? Can it wipe out humanity or wipe out much of our infrastructure? Can it cause massive direct and indirect damage to civil society with rampant

[69] Clouceff and Greenhalgh, "The Next Pandemic Could Be Dripping On Your Head."
[70] REUTERS, "Civet Cat Becomes SARS Scapegoat."
[71] Nebehay, "Avoid Caves in Uganda after Marburg Death."
[72] John Hopkins Bloomberg School of Public Health, "Centers for Disease Control and Prevention (CDC) Classification of Bioterrorism Microorganisms."

disease, civil unrest, anarchy, and a greater need to trade freedom for security? Is biological terrorism an existential threat, a catastrophic threat, a weapon of mass disruption, or a WMD? Is it a "black swan" low probability high consequence event or merely a weapon that temporarily overwhelms local and regional resources and responses, but then quickly recedes? It's hard to know when this has never happened.

III.4.2 WHAT THE US HAS DONE

In the US, we have spent a great deal of preparedness money, signed on to an international treaty, continued studying these agents for defensive purposes, and passed "select agent" laws.

In total, over $80 billion has been spent on biological defense initiatives. Yet there's still too few FDA cleared drugs or vaccines becoming available. Just one of thirteen viruses classified as category A have an approved therapy. It's been estimated that the cost of insuring adequate pandemic preparedness worldwide is greater than $3 billion a year. However, the projected annual loss from a pandemic could run as high as $570 billion. Acts of terror in total cost us over $50 billion across the globe per year[73]. With this in mind, how much spending is appropriate? John Mueller, former President of The Ohio State University once said, "if your chance of being killed by a terrorist in the United States is one in 3.5 million the question is how much do you want to spend to get that down to one in 4.5 million?" [74] Tom Clancy may have answered that question when he said, "the only difference between reality and fiction is that fiction has to make sense.[75]" But unlike his masterfully prepared stories, without good intelligence work we do not know how, where, or when the next attack will happen.

In 1972, we became signatories to the United Nations Convention on the Prohibition of the Development, Production and Stockpiling of

[73] Gates, "A New Kind of Terrorism Could Wipe out 30 Million People in Less than a Year — and We Are Not Prepared."
[74] Mueller, "Is There Still a Terrorist Threat?"
[75] Clancey, "Tom Clancy Quotes."

Bacteriological and Toxin Weapons and or Their Destruction (usually called the Biological and Toxin Weapons Convention or BTWC)[76]. By November 2001, there were 162 signatories with 144 ratifying the convention. Unfortunately, without established verification, many nations such as Russia and Iraq didn't abide by its terms.

Subsequent to the 2001 anthrax letters there have been strict laws enacted about possession and experimentation of what are called "select agents", or bugs mostly on the Category A list. This has made research with these agents more and more difficult. Scientists have even been jailed for merely misplacing a specimen. Publication of potentially dangerous findings is also a problem we will discuss later in this paper.

There are danger signs. The H7N9 flu virus responsible for China's 2017 epidemic is just two mutations away from spreading easily between humans. But as it infects more people some U.S. virologists who want to help are stuck on the sidelines because of a temporary 2012 U.S. ban on government funding for research that makes potentially pandemic viruses more dangerous or transmittable. The Department of Health and Human Services instituted the policy to counter bioterrorists, but researchers say those experiments may determine whether H7N9 can easily be spread in the air between mammals[77].

There have been dozens of countries with biological weapons programs. Best known are the very different Biopreparat program in the USSR[78] studying offensive technology, and the US program at the US Army Medical Research Institute for Infectious Disease (USAMRIID), which stopped offensive weapons experiments in 1969, but continued studying defensive technology.

III.4.3 MORE RECENT CONCERNS

Expert opinions differ on the plausibility of a bioattack. The U.S. Office of the Director of National Intelligence stated in 2008 that bioterrorism

[76] "Biological Weapons Convention."
[77] Jun. 20 and 2017, "Bioterrorism Rule Blocks Some U.S. Researchers from Studying Bird Flu."
[78] Alibek and Handelman, *Biohazard*.

is a more likely threat than nuclear terrorism. That same year, Director Mike McConnell stated that of all weapons of mass destruction, biological weapons were his personal greatest worry. Other defense experts and scientists insist that the possibility of any attack, especially a large-scale one, is small, given the immense challenges to obtaining, weaponizing, and deploying biological agents. Regardless, most biosecurity experts acknowledge that the potential of an attack should not be ignored. Moreover, specific preparations for a biological attack will likely benefit the response to other kinds of public health emergencies[79].

A chilling address was delivered at the 2017 Munich Security Conference by Bill Gates, who warned world leaders that a genetically engineered virus could kill more people than a nuclear weapon. His concern was supported by US and British intelligence agencies, which feel that "scientific terrorists" have access to the necessary tools to design biological weapons of mass destruction capability. In the face of this knowledge, governments around the world remain complacent and neither prepared or equipped to respond. Gates suggested preparing for epidemics "the way the military prepares for war," with more exercises and training. He added that "we ignore the link between health security and international security at our peril."[80]

III.4.4 AGENT, PERPETRATOR AND TARGET

Biological agents can be cheap and relatively easy to make, as production recipes have been available on the Internet. Detection of these agents is difficult as they are essentially hidden from our five senses. Their toxins are the most dangerous per weight of all substances. There is an incubation period between the time of contact with the agent and the development of any symptoms. This can range from a day for plague or a few weeks for smallpox. A well-planned weaponized aerosol disseminated through the air can have far-reaching

[79] "Biological Weapons, Bioterrorism, and Vaccines | History of Vaccines."
[80] Selk, "Bill Gates."

consequences, especially in enclosed spaces such as closed stadiums, subway systems and convention halls.

Having said all this, what does a terrorist need to get started? He or she needs basic knowledge of microbiology and specific weaponization know-how. They need equipment. They need the seed stock of the biological agent. Weaponization involves a number of chemical processes, including the important particle size of 1-5 microns. The product will need a coating with no electrostatic charge that is non-hydrophobic, and non-self-adhering. Other potentially manipulated factors are the virulence, the viability of the agent within each particle, and resistance to vaccines and antibiotics.

Put another way, you need an effective terrorist group or perpetrator and a target population. The selected biological agent has to be available in large enough quantities, weaponized, with a proper method of dissemination. If the attack is to occur outdoors, there are environmental conditions to consider such as inversions, wind speed, and ambient temperature.

Ideally the terrorist needs to do testing, which the US military did in the 1950s. Successful testing was done in Alaska to assess the influence of extreme cold weather. In Texas, an experiment was done utilizing flying aircraft to assess if there was sufficient energy to deliver a respiratory sized particle aerosol using liquid agent (there was). In New York City, experiments were done to see if subway cars generated sufficient wind currents to disseminate a dry powder product (they did). In San Francisco Bay experiments assessed the influence of wind speed and temperature. All this information is available to terrorists[81].

There are many routes of exposure to a biologic agent. There is ingestion, where the agent enters when the host eats, drinks, or puts contaminated fingers or other objects into his or her mouth. An example is the Salmonella salad bar attack by the Rajneesh in Oregon[82]. There is absorption, where the agent enters the host through mucosal surfaces

[81] Loria, "Over and over Again, the Military Has Conducted Dangerous Biowarfare Experiments on Americans."
[82] "1984 Rajneeshee Bioterror Attack."

such as eyes, nose, or throat, and rarely intact skin. T2 mycotoxin (a fungal byproduct) is an example. There is injection where the organism enters the host through the skin via a vector such as animal bites, hypodermic needles, any sharp object, or defects in the skin surface. An example is the 1978 incident where the Bulgarian Jorge Markov was poked with a ricin-laden umbrella from a spring-loaded gun[83]. The most important route is inhalation, where the agent enters when the host takes a breath. This occurred with the anthrax letters. The optimal particle size as mentioned is 1-5 microns, so the agent can easily establish itself in the lungs, cause local disease, or even enter the bloodstream. Inhalation exposure can occur with aerial bombs, spray tanks, carefully designed ballistic missile warheads, or even letters.

Who, besides the perpetrator, is involved in a biological attack? There are of course the ill, as well as the worried well. But there are also first responders, public health authorities, investigative agencies such as the FBI and local public safety, politicians, and the press – all needing to work together. To understand the interactions involved there is the "epidemiologic triangle", showing the relationship between the (1) biological agent, (2) the human (or animal) host, and (3) the environment. Changes in the environment may affect the transmissibility of an agent or the emergence of pathogenicity. Host behavior as well as changes to the biological agent itself may affect disease severity and transmissibility.

III.4.5 INDICATORS OF AN ATTACK

How will we know that we have been attacked? Clinical indicators may include unusual time of year or locale for a particular set of common symptoms. There may be an unusual number of sick or dying people or animals. There may be an unscheduled or unusual dissemination of sprayed material such as mosquito sprayers or crop dusters. We may even find abandoned sprayer dispersion devices.

There may be a "sentinel" case where an astute physician has diagnosed the disease in one of his or her patients. There may be arrival in large

[83] Edwards, "Poison-Tip Umbrella Assassination of Georgi Markov Reinvestigated."

numbers overwhelming emergency rooms. There may be gradual recognition by the medical community, an indicator from a public health surveillance system, a lab report or the opening of a suspicious letter. The best-case scenario is prevention by infiltration into terrorist organizations from intelligence gathering. The worst-case scenario is after mass chaos has occurred.

A biological attack will have three main phases. First, there's the pre-exposure planning phase where drug prophylaxis, immunization and training are done. Second, there is the incubation period phase, which may range from a few hours (toxins) or days to several weeks. This is the period between acquisition of infection and manifestation of disease. Finally, there is the period of overt disease where diagnosis, treatment, and multidisciplinary communication become paramount. The window period between the onset of illness in the first few cases and confirmation of the agent is very important. Vital decisions will have to be made at this time with limited data on hand, and the longer-term implications of these decisions are going to be important regarding resource allocation and public statements. After all this there is the recovery or resilience phase.

III.4.6 A HISTORY OF SELECTED BIOEVENTS

To understand the potential impact of a new biological weapon or pandemic we have to go back historically to naturally occurring pandemics. For perspective in the 20th century, 100 million people died from armed conflict, while 300 million people died from smallpox alone[84]. But go back to1346 to the battle of Caffa, where the Tartar army hurled the corpses of plague victims over the walls of the city, a seaport on the Crimean coast. Gabriel de Mussis, a notary, saw the Tatar attack on this well-fortified Genoese-controlled seaport (in modern Ukraine). He described how the plague-weakened aggressors catapulted dead victims of plague into the town: "[The Tatars], fatigued by such a plague and pestiferous disease, stupefied and amazed, observing themselves dying without hope of health ordered cadavers placed on their hurling machines and thrown into the city of Caffa, so that by means of these

[84] "Smallpox."

intolerable passengers the defenders died widely."[85] An epidemic of plague within the walls of the city followed, forcing a retreat of the Genoese forces. The exported disease continued to spread in Europe. Some speculate that this may have set off the European Black Death pandemic of the Middle Ages that resulted in 30-40 million deaths worldwide.

There is the population collapse in Mexico in the 1500s. There were three major outbreaks during that century, the first of which in1520 was thought to be caused by smallpox. Then there were outbreaks in 1545 and 1576 of "Cocoliztli", a disease that may have been Argentinean or Bolivian hemorrhagic fever, aggravated by unusual climate conditions. A consequence of these epidemics was the Mexican population collapse from close to 22 million to 2 million people within less than a century. The Mexican population did not recover until the 20th century[86].

The French and Indian Wars were not spared from acts of biological terrorism. British forces used plague-tainted clothing as a weapon in the 1785 siege of La Calle[87]. Earlier, British officers such as Sir Jeffrey Amherst had plans to intentionally transmit smallpox to Native Americans during Pontiac's Rebellion near Fort Pitt (present-day Pittsburgh, Pennsylvania) in 1763. It is not clear whether they actually carried out these plans. But, whatever its source, smallpox did spread rapidly among Native Americans in the area during and after that rebellion[88].

There was Japan's Unit 731 run by General Shiro Ishi between1937 and1945. Unit 731 was a biowarfare unit masquerading as a water purification facility. The Japanese, who controlled Manchuria, did

[85] Derbes, "De Mussis and the Great Plague of 1348."
[86] Al, "Megadrought and Megadeath in 16th Century Mexico - Volume 8, Number 4—April 2002 - Emerging Infectious Disease Journal - CDC."
[87] "Biological Weapons, Bioterrorism, and Vaccines | History of Vaccines."
[88] "Biological Weapons, Bioterrorism, and Vaccines | History of Vaccines."

experiments with plague dropped from airplanes by uniquely designed fragmentation bombs. Prisoners of war were also subjected

to aerosolized anthrax. In total, over 9,000 test cases eventually died. In one experiment, they had plague-infected fleas dropped with grain from airplanes. The grain attracted rats, and the rats became infected from the fleas. This brought the disease deeper into the human population[89]. I should note that General Ishi was never prosecuted as a war criminal.

In their experiments, the Russians produced biological agents in unconscionable quantities. For example, they produced greater than 1,000 metric tons of anthrax and plague, and 100 metric tons of smallpox. This was enough biomaterial to eradicate every living thing on the entire planet[90]. In 1979 the Soviet Union also had an anthrax accident at Sverdlovsk, a Soviet military compound. The accident resulted in 77 cases and 66 deaths, all downwind from the facility, and all from less than 1 gm of weaponized anthrax spores. Cases occurred within the incubation period of 4 to 45 days. The Russians, originally stating that this was naturally occurring intestinal anthrax, did not admit this was from a military accident until many decades later[91].

Iraq had biological and chemical weapons in the 1990s. They produced 19,000 L of botulinum toxin, 2,200 L of aflatoxin, and 8,500 L of anthrax. They filled 25 missiles and 166 bombs with anthrax and botulinum toxin. They also had delivery systems including SCUD missiles, bombs, and even a remote-controlled MIG-21 equipped with a 500-gallon sprayer[92]. The reason these agents were not used in Gulf War I remains unclear.

The first known biological attack on the US occurred in The Dalles, Oregon, in 1984. The Rajneesh, an Indian religious cult, contaminated the salad bars of nearby restaurants with a strain of Salmonella. Over 750 people were poisoned and 40 were hospitalized. None died. The purpose was to influence the outcome of a local election. What made

[89] Morris, "Forgotten Horrors."
[90] Ouagrham, "Biological Weapons Threats from the Former Soviet Union."
[91] Wampler and Blanton, "Anthrax at Sverdlovsk, 1979."
[92] Villar, Elliott, and Davenport, "Botulism."

this even more interesting was that it was not discovered as an intentional act until years later when members of the cult were indicted for another crime[93].

The 2001 anthrax letters resulted in 22 cases and five deaths. Initially, the letters were sent to NBC News, the New York Post, and AMI. Three weeks later two more letters were sent to Senators Tom

Daschle and Patrick Leahy. Dr. Steven Hatfill was the FBI's first person of interest, but eventually Bruce Ivins became the focus of the FBI investigation until his suicide just days before his indictment. The anthrax letters had multiple consequences. It was the largest investigation in law enforcement history, and at a very high cost. It led to numerous reports, commissions, meetings, books, articles, a building frenzy of high Biocontainment Level 4 labs, and many new federal programs[94].

III.4.7 US BIOPREPAREDNESS TODAY

Irwin Redlener, Director of the Earth Institute at Columbia University, said that preparedness was the right place on the continuum between mindless complacency and all-consuming paranoia[95]. In normal times, we look for medical affordability, quality and accessibility. We plan for dual use as well as plan stockpiles or caches. Disasters, when they happen, require a different approach, including functional resilience, the ability to surge, access to stockpiles, avoiding panic, and understanding that, with limited resources, we're dealing with a new type of ethics.

Should we be preparing for a conventional weaponized agent, or an exotic new threat: mixes of spores, bacteria, toxins, and viruses; altered antigenicity agents for vaccine invasion, antibiotic resistant organisms, chimera or fusion agents. There are gene drives and CRISPR technology with enormous implications for genetic manipulation. Cheap and simple

[93] Bovsun, "Guru of Poison."
[94] Lengel, "Little Progress In FBI Probe of Anthrax Attacks."
[95] Redlener, *Americans at Risk*.

CRISPR kits are now available on the Internet, allowing anyone to edit the genes of bacteria. The nightmarish prospect of engineered diseases looms[96]. And finally, there's the combination of a civilian, military and/or cyber targeted attack in conjunction with a biological attack.

Are "conventional" weaponized agents enough to worry about? Recently something very alarming happened in two laboratories. The highly pathogenic H5N1 influenza virus, that kills most of the people it infects, artificially acquired the ability to transmit easily in ferrets. The origin of this wasn't the naturally evolved strain, but one created in research laboratories in Holland and the US with the best of scientific intentions. Scientists had created a deadly strain of avian influenza virus that transmits easily in ferrets, who have a remarkably similar immune response to humans. Pressured not to publish its entire genetic blueprint, this heated up the debate about where you draw the publishing line. While no one wants a bioterrorist to get hold of such a recipe, researchers need to pool their knowledge, so as not to hamper their ability to give us the best chance of averting or surviving a pandemic of flu with a kill rate of 50% for example. US and the World Health Organization (WHO) committees disagreed about harm versus benefits of publishing this research[97].

Will offense or defense get ahead in the future, and will we have enough time to react quickly and efficiently? Disasters can happen in a matter of seconds. Budgets take years. Elected officials may have terms of only a few years. Strategic plans may take five plus years, but these problems will continue for decades.

Surveillance is paramount for preparedness, and falls into two categories: general routine non-specific (i.e. drug sales, absenteeism, syndromic), as well as the targeted surveillance that is intensified once we know that an outbreak has started. But we are living in a new world where there are prediction markets and social networking. There is Google trends, integrated diverse human and non-human surveillance

[96] Doudna and Sternberg, *A Crack in Creation*.
[97] Herfst et al., "Airborne Transmission of Influenza A/H5N1 Virus Between Ferrets."

systems, as well as big data mining and analytics. There is also uncertainty. Donald Rumsfeld, former US Secretary of Defense, once said "there are known knowns, there also known unknowns, but there are also unknown unknowns"[98]. How do you plan for an unknown unknown?

What about today's hospital preparedness system? In some ways, it's like independent fire stations without agreements for multiple alarm fires. There is no broad incident command (ICS) structure. The hospital ICS is a good start for facilities, but it is internal to each facility without a multi-hospital unified command, and preparedness exercises are rarely practiced. Electronic medical records systems, if used wisely, can help. Public health authorities have to get their message out. Unfortunately, in times of war today, hospitals and other healthcare facilities have been targets and are no longer safe havens, off-limits to bombardment.

III.4.8 BIO AS AN EXISTENTIAL THREAT?

Bugs constantly evolve to meet new challenges and move into new populations. Exposure opportunities increase as the environment changes. New niches open up that permit pathogens to expand and proliferate. And there will always be those with technical skills that want to cause us harm.

What are the factors that determine if a society, or a planet, ultimately fails or succeeds in today's rapidly changing world? Sixty years ago, Arnold Toynbee concluded in his "A Study of History" that the ultimate cause of imperial collapse was suicidal statecraft[99]. Jared Diamond, in his work Collapse addressed the significant influences: environmental changes, hostile neighbors, and trade partners — and considers the responses different societies have had to such threats[100].

[98] Graham, "Rumsfeld's Knowns and Unknowns."
[99] Toynbee and Somervell, *A Study of History. Abridgement of V. I-Vi by D.C. Somervell*.
[100] Diamond, *Collapse*.

In closing, let's look at the Drake equation and the Fermi paradox. The Drake equation is a probabilistic argument used to arrive at an estimate of the number of active communicative extraterrestrial civilizations existing in the Milky Way galaxy. There is an apparent contradiction between high estimates of the probability of the existence of extraterrestrial civilizations, such as in the Drake equation, and the lack of evidence for other civilizations[101].

This brings us to the Fermi paradox, which refers to modern science's "surprising" failure to detect extraterrestrial life, and provides evidence regarding the likelihood of humans surviving long enough to travel in space. Enrico Fermi, on a visit to Los Alamos, observed that given the age of the universe, the billions of stars in the galaxy, and the high probability that more than one of those stars ought to have developed intelligent life. Scientists should have found evidence of alien intelligence. Since this is not the case, we must consider several possible conclusions: (1) technologically advanced life might be extremely rare, (2) search methodologies are flawed, or (3) we are monitoring the wrong electromagnetic radiation band. Perhaps we simply have not been (4) searching long enough, or (5) intelligent life in the universe simply does not wish to contact Earth. Perhaps (6) civilizations such as ours are common in the universe but almost always perish before developing the capability for interstellar travel. If the universe contains a "trap" that usually destroys civilizations between when they split the atom and when they are able to colonize nearby star systems, we have more reason to suspect this trap than past civilizations at our level of development[102].

Nick Bostrom, the Oxford philosopher of existential risk, said "it might be that any sufficient technologically advanced civilization discovers some technology that causes its extinction. Given the increasingly volatile political situations around the globe, the existence of nuclear and biological weapons, a controversial disregard for and disagreement about the sustainability of our natural environment, and the

[101] "Drake Equation."
[102] Miller and Felton, "The Fermi Paradox, Bayes' Rule, and Existential Risk Management."

development of potentially destructive artificial intelligence, one can see outbursts might fall into the sort of doomsday trap" in the Drake equation[103].

The Israeli historian Yuval Harari noted that we are uneasily poised at the conjuncture of standard history and science fiction of sober analysis and mad prophecy of nightmare and Utopia [104]. With different backgrounds terrorists are prone to do different things. Osama bin Laden used airplanes as weapons. He was in the construction business. Ayman al-Zawahiri, who followed him, has not yet made his unique mark. He's a doctor. Abu Bakr al-Baghdadi, the leader of ISIS, is running out of options. Could use of a biological agent be more appealing to them?

What's the possible role of ISIS and other current terrorist actors? In 2014 an ISIS laptop was recovered containing a 19-page document on how to develop biological weapons, and late last year Kenyan authorities disrupted an anthrax plot by a medical student and associates affiliated with ISIS. Earlier this year, South Korea raised concerns that North Korea possesses biological weapons and could use drones to carry out attacks[105]. Even Russia may still be experimenting with bioweapons.

Bearing all this in mind, what's needed? This boils down to money wisely spent on:

- Multi-source pre-event intelligence
- Before and after disease onset disease surveillance
- Countermeasures such as vaccines, anti-toxins and antibiotics
- Multidisciplinary exercises, training, cooperation and communication

[103] Frye, "Life On Mars Bad?"
[104] Yang, "Is the 'Anthropocene' Epoch a Condemnation of Human Interference — or a Call for More?"
[105] Sell, "How Trump's Budget Makes Us All Vulnerable to Bioterrorism."

We need public-private partnerships in biomedical and inter-disciplinary research and collaboration. Public health needs to work well with clinical medicine, public safety and disaster management. We need a treatment coordination infrastructure with, for example, easily set up points of dispensing. There needs to be intelligence driven analytical threat mitigation, as well as good predictive modeling and the stockpiling of essentials.

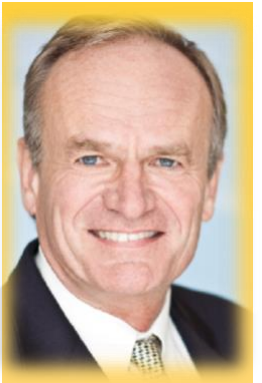
There are many relevant questions: How prepared are we? How prepared do we need to be? In a perfect world how prepared can we actually be? And finally, what training is or should be available to help us prepare and respond optimally? Access to data and the ability to quickly analyze it is the new source of authority.

There have been 10 naturally occurring pandemics in the last 300 years, and like acts of terror, they will come. We won't know how they will play out, or what the unknown unknowns will be. We do know the next time will be different and we will probably get through it. The Chinese depicted the word “crisis” with two pictograms, one was the symbol for danger and the other was the symbol for opportunity.

Lawrence Gostin, the director of the WHO’s Collaborating Center on Public Health Law and Human Rights, notes “the next weapon of mass destruction may not be a bomb. It may be a tiny pathogen that you can’t see, smell, or taste, and by the time we discover it, it’ll be too late.”[106]

[106] Baumgaertner, “Trump’s Proposed Budget Cuts Trouble Bioterrorism Experts.”

III.5 BIOLOGICAL AND NUCLEAR TERRORISM RISK ASSESSMENT (DETLOF VON WINTERFELDT)



Detlof von Winterfeldt is the Tiberti Chair for Ethics and Decision Making at the Viterbi School of Engineering of the University of Southern California (USC) and a Professor of Public Policy of USC's Sol Price School of Public Policy. In 2003 he co-founded the National Center for Risk and Economic Analysis of Terrorism Events (CREATE), the first university-based Center of Excellence funded by the US Department of Homeland Security.

III.5.1 INTRODUCTION

Terrorists pose many threats in the US and worldwide. These include explosives attacks on airlines, cars driven into large crowds, attacks with guns and knives in public places, attacks using chemicals or biological agents, and attacks using nuclear or radiological materials. Among these attack modes there are only two that pose potentially existential threats to the US and the world: Attacks with biological agents that could cause a major worldwide epidemic and nuclear terrorism that could kill hundreds of thousands of people and cause enormous economic impacts.

From its inception, the Department of Homeland Security (DHS) has been concerned with assessing these threats. In particular, the Science and Technology Directorate (S&T) of DHS has developed several tools for terrorism risk assessment, including the biological terrorism risk assessment (BTRA), the chemical terrorism risk assessment (CTRA), and the radiological/nuclear terrorism risk assessment (RNTRA). In this summary, we trace the origins of the terrorism risk assessments at DHS, discuss some of the early results and criticisms, and suggest some useful areas of future research,

III.5.2 ORIGIN OF THE TERRORISM RISK ASSESSMENTS (TRAS)

Soon after the creation of the DHS, the Science and Technology Directorate at DHS grappled with the issue of how to quantify the risk of terrorism using weapons of mass destruction, including biological and nuclear weapons. Of particular urgency was a request by the White House to provide bi-annual assessments of the threat that terrorists might use biological weapons (White House, 2004)[107]. In 2004 S&T established a team to investigate alternative methodologies to quantify

[107] The White House, "Homeland Security Presidential Directive / HSPD-10: Biodefense for the 21st Century."

these risks. At the time, there were many options for risk assessment, among them a standard probabilistic risk analysis (PRA) approach, an approach using a new method for quantifying risks using "possibility theory" (PT), and an approach using a risk scoring method based on multiattribute utility analysis (MAU). Each approach had received some funding to demonstrate that they could add value to the bioterrorism risk assessment. These demonstrations were conducted at different National Laboratories (PRA at Battelle Columbus, PT at Los Alamos National Laboratory, and MAU at Sandia National Laboratories).

At the time, the author of this summary was asked, as the Director of the National Center for Risk and Economic Analysis of Terrorism Events (CREATE) of the University of Southern California (USC), to provide an assessment of the three approaches. Other experts from universities were involved in this assessment as well. Their conclusion was that the risk scoring methodology using the MAU approach was too qualitative to be useful; that the PRA approach was appropriate, but lacked sufficient depth in modeling the subject matter of epidemiology of infectious diseases; and that the PT approach, while based on solid subject matter expertise, used an obscure mathematical theory that would not likely be acceptable by the broader scientific community or by policy makers. The review panel recommended to combine the best of the PRA and the PT approaches, in essence using the best of the subject matter expertise at LANL with the probabilistic modeling at Battelle, shedding possibility theory in the process. The panel also provided some caveats that it would be exceedingly difficult to assess probabilities of rare event involving terrorist attacks with biological agents.

For reasons that were not entirely based on this recommendation the S&T Directorate decided to only use the PRA approach and asked Battelle Columbus to develop it further for the first report to the President, which came to be known as the "Bioterrorism Risk Assessment" (BTRA), the first of several TRAs that were to follow. Battelle Columbus used a straightforward application of a nuclear power plant risk assessment as a template to model bioterrorism. In essence, they built a huge event tree that mapped out terrorist actions, targets, choices of biological agents, and eventual health consequences. They eventually developed event trees with millions of end points.

Detlof von Winterfeldt's research interests are in the foundation and practice of decision and risk analysis as applied to the areas of technology development, environmental risks, natural hazards and terrorism.

S&T continued to engage CREATE in the development of BTRA, primarily on the aspect of expert elicitation of probabilities of branch probabilities in the event tree (e.g., which agent are terrorists most likely to use?). CREATE researchers developed a protocol and computer tools to aid this elicitation, developed training materials for the subject matter experts (intelligence analysis and biological scientists) to express their judgments quantitatively, and provided some demonstration elicitations and protocols for elicitation. An important part of this elicitation was that experts were able to express their uncertainty about their probability judgments as secondary probability distributions. The actual elicitations were highly classified and it is not clear, if these protocols were followed or the computer tools were used.

The results of the Battelle Columbus PRA were published in the January 2006 Presidential Report on Biological Terrorism Risk Assessment (BTRA), which showed a ranking of biological agents in terms of the risk they posed to the country with associated uncertainties based on the uncertain inputs provided by the SMEs. A stylized example of the results is shown in Figure 1. For obvious reasons the x – axis (the biological agents) and the y-axis (risk=probability times consequences) are classified, but the general message of the BTRA was: Here are five agents that we should worry about (and take action like stockpiling vaccines), then there are several others, and some that we really should not worry about.

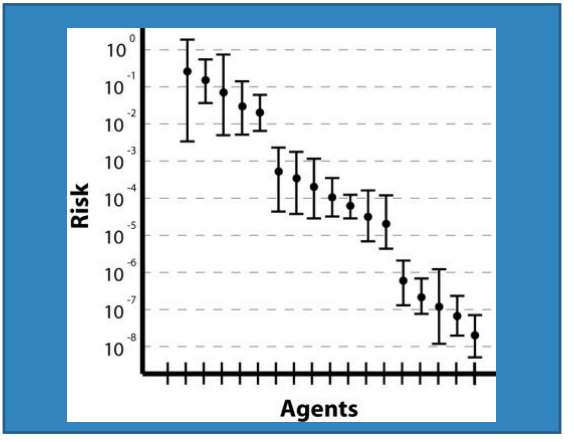


Figure 3: Illustrative Results of the First Terrorism Risk Assessment

III.5.3 BTRA CRITICISM BY THE NATIONAL ACADEMIES

Soon after the first BTRA was published, the National Academy of Science was asked by the S&T Directorate to conduct a review. This review was highly critical of the PRA approach (see National Research Council, 2008)[108]. Among other things, the NAS committee criticized that defensive actions by our government were modeled as events instead of decisions and that terrorist actions were modeled as uncertainties instead of the strategies that terrorists would use to achieve their own goals. One statement in the NAS report said that terrorist actions should not be treated as probabilistic inputs of an analysis but rather as outputs of an analysis. The committee report even stated that BTRA in the then current form should not be used for decision making – at the White House level or elsewhere.

The risk analysis community had several academic exchanges about the value and limitations of using PRA for terrorism risk analysis (see, for example, Garrick et al., 2004; Ezell et al., 2010)[109]. The main competitor was game theory and many game theorists argued that PRA, which was developed for natural disasters and technological accidents, was not suitable for situations involving an adaptive adversary. Decision analysts also chimed in, including researchers from CREATE, to argue that a mix of decision analysis, game theory and PRA was the best approach (Garcia and von Winterfeldt, 2016)[110].

III.5.4 LIFE AFTER BTRA

In spite of the criticism of the National Academy committee, BTRA survived and underwent several changes and improvements. It also spawned similar PRAs – the Chemical Terrorism Risk Assessment (CTRA), the Radiological and Nuclear Terrorism Risk Assessment (RNTRA) and the Integrated Terrorism Risk Assessment (ITRA). These risk assessments

[108] Council et al., *Department of Homeland Security Bioterrorism Risk Assessment*.

[109] John Garrick et al., "Confronting the Risks of Terrorism"; Ezell et al., "Probabilistic Risk Analysis and Terrorism Risk."

[110] Garcia and von Winterfeldt, "Defender–Attacker Decision Tree Analysis to Combat Terrorism."

are increasingly used within DHS and its components. Concerns are often raised that they are too complex and not sufficiently linked to the real issues that policy makers at the agency face. Research at CREATE is currently underway to make the TRAs more useful for decision making

III.5.5 BIOLOGICAL AND NUCLEAR TERRORISM: ARE THEY EXISTENTIAL THREATS?

Due to the classified nature of the TRAs, it is not possible to answer this question in quantitative terms, but there is sufficient unclassified data that, coupled with some common-sense insights, can lead to conclusions. Regarding biological threats, several attacks have occurred in the past using biological agents like bacillus anthraxis (anthrax), clostridium botulism or similar agents (botulism, salmonella), ricinius communis (ricin). None of these has proven to be existential threats. An existential threat is posed by yersinia pestis, the biological agent than causes the pneumonic plague or the variola major virus that causes small pox. However, it is extremely unlikely that terrorists would use these agents, even if they were able to cultivate them, because the "blow back" would likely kill more of their own than in the western countries that they are trying to attack. The most effective action against these threats is the US strategic stockpile of vaccines that would, in the event, reduce the number of infections and deaths.

Radiological attacks with a dirty bomb have very limited health effects (see Rosoff and von Winterfeldt, 2007)[111]. They can cause a few fatalities due to acute radiation exposure and, at the high end, hundreds of latent cancers due to exposure to a radioactive plume or persistent ground shine. They can cause massive fear and economic consequences, primarily due to the concerns with low-dose radiation and the extremely expensive decontamination costs. The main defensive action to counter radiological attacks is to secure the sources of radiological materials – in hospitals, food and blood irradiation plants, and nuclear facilities.

This leaves nuclear terrorism as the ultimate threat. Unclassified documents suggest that a terrorist attack in New York City with a 10-

[111] Rosoff and Von Winterfeldt, "A Risk and Economic Analysis of Dirty Bomb Attacks on the Ports of Los Angeles and Long Beach."

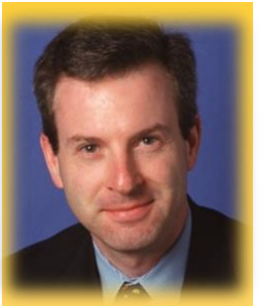
kiloton improvised nuclear device can kill hundreds of thousands people immediately due to the blast and fire storm and at least one hundred thousand additional fatalities due to latent cancers. The US has developed the Global Nuclear Detection Architecture, which, in addition to securing nuclear materials at US and foreign facilities, tries to prevent special nuclear materials like enriched uranium and plutonium to enter the US territories.

III.6 THE TRAGEDY OF UNCOMMONS: PSYCHOLOGY, POLITICS AND POLICY (JONATHAN WIENER)

III.6.1 AUTHOR AND LINK

Author: Jonathan Wiener

Link: http://onlinelibrary.wiley.com/doi/10.1111/1758-5899.12319/full



Jonathan Wiener is Perkins Professor of Law, Environmental Policy and Public Policy, at Duke University, where he co-directs the 'Rethinking Regulation' program. He is also a University Fellow of Resources for the Future, and a member of the scientific committee of the International Risk Governance Council.

III.6.2 ABSTRACT

The 'tragedy of the commons' is a classic type of problem, involving multiple actors who face individual incentives to deplete shared resources and thereby impose harms on others. Such tragedies can be overcome if societies learn through experience to mobilize collective action. This article formulates a distinct type of problem: 'the tragedy of the uncommons', involving the misperception and mismanagement of rare catastrophic risks. Although the problem of rare and global catastrophic risk has been much discussed, its sources and solutions need to be better understood. Descriptively, this article identifies psychological heuristics and political forces that underlie neglect of rare catastrophic 'uncommons' risks, notably the unavailability heuristic, mass numbing, and underdeterrence. Normatively, the article argues that, for rare catastrophic risks, it is the inability to learn from experience, rather than uncertainty, that offers the best case for anticipatory precaution. The article suggests a twist on conventional debates: in contrast to salient experienced risks spurring greater public concern than expert concern, rare uncommons risks exhibit greater expert concern than public concern. Further, optimal precaution against uncommons risks requires careful analysis to avoid misplaced priorities and potentially catastrophic risk–risk trade-offs. The article offers new perspectives on expert vs public perceptions of risk; impact assessment and policy analysis; and precaution, policy learning and foresight.

III.7 SOCIETAL AND ETHICAL ISSUES RELATED TO CATASTROPHIC AND EXISTENTIAL RISK (ANDERS SANDBERG)

Anders Sandberg's research at the Future of Humanity Institute at University of Oxford centers on management of low-probability high-impact risks, societal and ethical issues surrounding human enhancement, estimating the capabilities of future technologies, and very long-range futures. He is currently senior researcher in the FHI-Amlin industry collaboration on systemic risk of risk modeling.

Existential risk poses unusual societal and ethical issues. From an ethical standpoint, the big issue is the nature of the moral disvalue of existential risk: most value systems assign it a enormous importance. But the rational level of concern depends on how the future is discounted, the value of future generations, extreme uncertainties, and how to handle the apparent paradoxes that extreme actions appear rational due to the overwhelming value at stake. From a societal perspective, existential risk can be caused or worsened by maladaptive societal organisation. However, resiliency and policy responses are also societal properties: many of the key challenges in mitigating existential risks are related to how societies react to unusual or unprecedented extreme risks.

III.7.1 INTRODUCTION

Humanity, like any species, is subjected to risk. This ranges from global disasters such as pandemics that could kill a large fraction of the world population over permanent stagnation or dystopias to extinction (Rees 2003; Bostrom & Ćirković 2008)[112]. These big risks are not theoretical. It has been suggested that a supervolcanic eruption 73,000 ago reduced human populations to near extinction (Ambrose 1998)[113], and the Cold War certainly had numerous frighteningly close calls (Schlosser 2013)[114]. 99.9% of all species that have existed are extinct (Raup 1986)[115] and all related hominin species have gone extinct, H. neanderthalensis just 40,000 years ago.

Human extinction comes under the umbrella term of existential risk, defined as "premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable

[112] Rees, *Our Final Century*; Bostrom and Cirkovic, *Global Catastrophic Risks*.
[113] Ambrose, "Late Pleistocene Human Population Bottlenecks, Volcanic Winter, and Differentiation of Modern Humans."
[114] Schlosser, *Command and Control*.
[115] Raup, "Cohort Analysis of Generic Survivorship."

future development" (Bostrom 2002, 2013) [116]. This covers both standard extinction and possible degenerate states. In general, these risks have a scope that is global and often transgenerational (they have consequences that matter for many or all future generations) and a severity that is catastrophic or fatal. There is no strict definition for global catastrophic risks (GCRs), but one suggestion is an event that would lead to the death of approximately a tenth of the world's population or have a comparable impact (Cotton-Barratt et al. 2016)[117].

Such risks may have low probability but their extreme severity makes managing them important challenges for the long-term wellbeing of humanity. However, due to their extreme nature they also pose special challenges to how we evaluate risks ethically and socially that can make mitigation harder.

The aim of this essay is to outline some of the ethical problems, such as: What is the moral disvalue of existential risk? What is the rational level of concern? How should we take the value of future generations into account? How to handle the extreme uncertainty and apparent paradoxes of extreme action in the face of extreme stakes? It will also look at some of the social problems, such as: How does social organisation increase or decrease extreme risk? How can responses be made more appropriate?

Anders Sandberg is currently senior researcher in the FHI-Amlin industry collaboration on systemic risk of risk modeling. He is senior Oxford Martin fellow, and research associate of the Oxford Uehiro Centre for Practical Ethics, the Oxford Centre for Neuroethics, the Center for the Study of Bioethics (Belgrade), and the Institute of Future Studies (Stockholm).

III.7.2 HISTORY

Most, if not all, human cultures have imagined an end of the world. The role of apocalypses in culture involve a satisfying ending to history, a promise of revelation, an end of suffering, a new start, or a sense of shared purpose. Occasional millenarianist outbreaks channel social anxieties or the need for reform, sometimes with destructive effects. However, except for these rare outbreaks, this apocalypticism is typically framed in terms of hope, theology and cultural history rather than as a practical risk.

[116] Bostrom, "Existential Risks"; Bostrom, "Existential Risk Prevention as Global Priority."
[117] Cotton-Barrat et al., "Global Catastrophic Risks 2016."

It was in the early modern era the end of the world became a scientific proposition rather than a story. This came from several different sources. Astronomy recognized that the heavens could change (e.g. by observing supernovae), that comets cross planetary orbits and that meteors are extra-terrestrial impactors: it naturally suggested that Earth could be hit by comets or asteroids, and that the stability of the sun was not a given. Evolutionary biology made it clear humans were a species among others and hence could potentially go extinct or degenerate. Thermodynamics implied the eventual heat death of the universe. These perspectives implied a bleak future or ongoing, hard to prevent threats. While many of these worries were confined to scientific speculation and fiction such as H.G. Wells' The Time Machine (1895) they occasionally produced scares such as the concern in 1910 that cyanogen from the tail of Halley's Comet might poison the Earth's atmosphere (Bartholomew & Radford 2011, ch. 16)[118].

In the 20th century the end of the world went from a hypothetical possibility to a frightening reality. The threat of nuclear weapons and environmental degradation created a focus on anthropogenic existential risk that could happen within one's lifetime or the near future (Kuznick 2007)[119]. Surveys show that a large fraction of the public regard an end of our existing way of life within a century as a likely possibility (Randle & Eckersley 2015)[120]. The crucial difference from earlier periods was that the risks were due to human actions and technology rather than natural disasters. The shift in focus away from natural risk also led to the realisation that human agency could affect and reduce extinction risks through appropriate policy.

[118] Bartholomew and Radford, *The Martians Have Landed!*
[119] Kuznick, "Prophets of Doom or Voices of Sanity?"
[120] Randle and Eckersley, "Public Perceptions of Future Threats to Humanity and Different Societal Responses."

III.7.3 WHAT IS THE MORAL DISVALUE OF EXISTENTIAL RISK?

The most quoted and classic introduction to the issue of how bad existential risk is can be found in (Parfit 1984)[121]:

> I believe that if we destroy mankind, as we now can, this outcome will be much worse than most people think. Compare three outcomes:
> (1) Peace. (2) A nuclear war that kills 99% of the world's existing population. (3) A nuclear war that kills 100%.
> (2) would be worse than (1), and (3) would be worse than (2). Which is the greater of these two differences? Most people believe that the greater difference is between (1) and (2). I believe that the difference between (2) and (3) is very much greater. … The Earth will remain habitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilized human history. The difference between (2) and (3) may thus be the difference between this tiny fraction and all of the rest of this history. If we compare this possible history to a day, what has occurred so far is only a fraction of a second.

Parfit opened the conversation on what moral disvalue existential risk has (and hence, the moral importance of preventing it). Since then other thinkes has attempted to analyse it (Leslie 1998; Bostrom 2003, 2013; Posner 2004; Matheny 2007; Beckstead 2013)[122].

In the case of human extinction, the most obvious loss of value is the remaining lifespan of existing humans: it would at least be as bad as 7.3 billion deaths.

---

[121] Parfit, *Reasons and Persons*.

[122] Leslie, *The End of the World*; Bostrom, "Existential Risk Prevention as Global Priority"; Bostrom, "Astronomical Waste"; O'Malley, "Catastrophe"; Matheny, "Reducing the Risk of Human Extinction"; Beckstead, "On the Overwhelming Importance of Shaping the Far Future."

However, it would also imply the loss of value of all human culture and artefacts that would be destroyed or become meaningless. All effort that has gone into building cathedrals, writing music or caring for one's grandchildren would be for naught. This includes thwarted preferences of past generations who hoped for various things, such as being remembered and having descendants, as well as our own hope for a glorious future.

A more serious loss may be the loss of future generations. Beside the present generation all future generations and all the good they could experience and achieve will never come to exist.

One way of putting this issue is to argue that just as we should not discriminate against spatially remote human lives, so we should not discriminate against temporally remote human lives: they all have the same (or similar) value. This to some extent follows as an ethical consequence of special relativity theory if we make the (common-sense?) assumption that ethics must be Lorenz-invariant like physics is. This means that allowing harm to come to future generations is as bad as allowing harm to existing people. Indeed, they are in a particularly vulnerable position since they cannot perform any action or make any plea to influence our behaviour.

Another line of reasoning discounting prices of commodities makes economic sense it may not make sense to discount fundamental goods such as well-being or value itself (Broome 1994)[123], which would plausibly include the value of human lives. Indeed, if we were to discount future human lives at some fixed discount rate there will be a future time when all of subsequent history would be worth less than a fraction of one present human life: it would be rational to set events in motion that would lead to inevitable eventual human extinction for the sake of giving even trivial help to someone in the present.

The sheer magnitude of the future at stake is often underestimated. Even fairly modest calculations of the number of human lives that could come about give a tremendous weight to future generations. If, for example, we assume a sustainable population of just 100 million people

[123] Broome, "Discounting the Future."

each living for 100 years over a span of 800,000 years (giving H. sapiens a total lifespan of one million years, typical for a mammalian species) the total number of future human lives would be 800 billion, more than 100 times the present human population. If humanity spreads into space and persists over longer periods the numbers quickly become astronomical – even conservative models give rise to estimates above 1030 lives. Even though we may be highly uncertain about just which scenario can and will come to pass, the median result is still enormous and robust to model changes. This in turn implies that the value of reducing risk to the entire future is very large, even when the reduction of hazard probability is tiny by ordinary standards (Bostrom 2003)[124]

Whether there is an intrinsic value in the existence of the human species (or any species) beyond the value to or of its members is debated in environmental ethics (O'Neill 1992; Bradley 2001)[125]. But if there is one, for example because the unique human perspective on the world has value, or because our ongoing history and human potential matter for their own sake, then clearly existential risk matters. Conversely, another position might be that if humans are the sole conscious observers in the universe and value is conferred by valuing observers, then the loss of observers may imply the loss of value itself: without anybody to experience the universe there is no point in it.

Finally, and perhaps trivially, existential risk represents a massive loss of options. Whatever the good is, we will not be able to figure it out or achieve it if we are extinct, degenerate or have lost the ability to achieve an open future.

A more rarely asked question is the moral disvalue of global catastrophic risks. It is not unreasonable to believe that at some level of maturity humanity would be relatively safe from both anthropogenic and external risks. Even a non-terminal disaster can matter beyond the direct impact: a "mere" global disaster causes global loss of time, human capital and opportunity beside the active harm, prolonging the time humanity remains in our currently vulnerable state. Events that reduce

[124] Bostrom, "Astronomical Waste."
[125] O'Neill, "The Varieties of Intrinsic Value"; Bradley, "The Value of Endangered Species."

the chance of reaching maturity or delay it hence increase existential risk even if they are not in themselves lethal.

III.7.4 DOUBTS ABOUT EXTINCTION

One can agree with the existence of a vast future having positive value without accepting it as having overwhelming value if one thinks that extra lives have diminishing value. It does not have to be the ruthless "One death is a tragedy, a million statistics" perspective, just the view that once there are trillions of human beings the extra value of adding another one is less than in the case where there were just a million humans. But this does not imply that extinction does not matter: as the population dwindles, every life becomes more precious.

The most common doubt is the person affecting view: something needs to be good or bad for someone in order to matter, so the loss of all future generations does not matter since they do not exist yet, and will not exist in this case. If one holds this view then the far future does not matter beyond what extent people care about it. Present-day existential risk may still matter, since existing people may find their lives cut short by a near-term disaster.

Not everyone agrees that human extinction is a negative thing (other risks such as eternal dictatorship or permanent loss of value or potential may still count strongly).

Antinatalists argue that existing or at least coming into being actually has a negative value, and that we have moral reasons not to want this to continue (e.g. (Benatar 1997))[126]. From this perspective extinction would merely cause negative value to existing people and prevent an enormous amount of disvalue. While antinatalist views have been expressed widely across the history of philosophy they are not particularly popular (something adherents may attribute to human evaluation biases).

These arguments are somewhat related to Lucretius' arguments in De Rerum Natura that death is not to be feared since one does not exist

[126] Benatar, "Why It Is Better Never to Come into Existence."

while dead, and hence one cannot be harmed by being dead. A species-level version of the argument is analogous, and has the advantage that there are no mourning relatives who could provide some disvalue of one's absence.

Even more pessimistic views view life as having unacceptable levels of pain and suffering. Were we to accept them, we should try to euthanize ourselves and maybe even the biosphere (since wild animals suffer significantly). Needless to say, these views are even less popular. However, they might hypothetically motivate well-meaning people to seek to cause existential threats (see below).

Moral uncertainty arguments consist of recognizing the existence of different well-reasoned moral or axiological theories and that we are uncertain about which are correct. We should hence try to act in such a way that the most likely outcome is right or acceptable. In the choice between theories that assigns tremendous disvalue to existential risk and theories that are neutral about it (like the person-affecting view and Lucretian views) acting to reduce existential risk is OK even to the neutral theories, so we should reduce existential risk. However, were we also to take antinatalist and pessimist views into account the evaluation now must somehow balance the potential harm with the potential gains caused by bringing humans into being. This depends sensitively on what arguments one accepts.

III.7.5 JUSTICE AND EXTREME RISK

In many ethical discussions justice is a key consideration, yet it is more rarely invoked when discussing existential risk. Arguing that humanity should not be saved because of (say) giving higher priority to justice (a very literal take on fiat justitia, et pereat mundus) or individual autonomy is self-defeating since justice and autonomy require that there are agents.

However, an intervention to avert the total loss of the future can still be more or less moral. If there are two options and one is morally unproblematic while the other one is immoral then obviously humanity ought to select the unproblematic one. But what if there is normative

disagreement? Or the different options are not equal in their feasibility or who gets saved?

In philanthropy, there is sometimes a tension between helping the global poor and reducing existential risk. This is partially an issue of person-affecting or non-person affecting ethics. But one can also argue that it is an allocation between needy people now and protecting people in the future.

Avoiding existential risks in the first place is great from a fairness perspective: everybody benefits equally. Recognizing the extreme importance of existential risk may have beneficial effects in that current conflicts and struggles are recognized as petty compared to the shared threat. There might however be a fairness issue in who pays for the necessary interventions or pre-disaster resiliency. For example, only a few nations can at present perform asteroid deflection missions: could they demand support from other nations, or should they bear the costs as they typically represent the most fortunate and prosperous nations?

When interventions try to slow or halt the damage from a hazard, unfairness can creep in. Even if the interventions are eventually successful certain groups may have borne a significant amount of damage either because of practical factors (closeness to the source, isolation, higher vulnerability) or decisions made by others (prioritization, mitigation strategy).

Even more salient is who is imposing risks on others. For anthropogenic risks, many come about through actions by particular groups, yet other groups will bear the harm if the risk occurs. This may place a valid and very strong claim on groups generating potential risk for everybody: they have a responsibility to offset the imposed risk. Since compensating for existential risk is extremely hard the only activity that might have a chance to offset the imposition of existential risk is existential risk reduction (this is the claim in the gain-of-function influenza research controversy, where researchers produce novel strains of pandemic flu – introducing new risk – in the hope of mitigating the pandemic risk). Hypothetically, “existential hope” situations where good outcomes add

significant value to the future (Cotton-Barratt & Ord 2015)[127], could make some significant risks worth taking. But arguing the case beforehand would require extraordinary evidence or plausibility given the high stakes.

The GCR case has far more scope for unfairness, since existential threats affect humanity as a whole while GCR will affect some subset. Empirically (or almost by definition) disasters tend to harm the poor and vulnerable the most. That means that fairness as a moral motivation to reduce risk may become far more active for large but not terminal risks.

III.7.6 WHAT IS THE RATIONAL LEVEL OF CONCERN?

Assigning a tremendous value to mitigating existential and GCR risk is not unproblematic. Problems occur because mitigation seem to take precedence over anything else. Surely giving up non-essentials like art or entertainment, or allocate resources from non-existential risk (such as flood and earthquake) mitigation in favour of existential risk mitigation is too extreme? One might argue that this is a reductio ad absurdum of existential risk, or bite the bullet and give total priority to existential risk. More likely we have not yet understood how to balance the extremes of these risk, theoretically and practically.

One problem is that when utilities become very large actions that are unlikely to succeed may still have a very large expected utility. This is sometimes called the "Pascal's mugging" problem after a paper by Bostrom (2009)[128] where a robber convinces Pascal to hand over his wallet in exchange for many times its value next day: even though the bargain is unlikely to be upheld, given a sufficiently great reward it becomes rational. The fact that we are uncertain about our own reasoning and probability estimates makes it hard to argue that the probability is so low that it is not rational to hand over the money. One way around this is to use bounded utilities: the reward cannot be arbitrarily large (in the case of existential risk, this means strongly discounting the value of future or present people). Another is to not naively use expected utility but instead apply Bayesian probability to

[127] Cotton-Barrat and Ord, "Existential Risk and Existential Hope."
[128] Bostrom, "Pascal's Mugging."

weigh options based on how good their evidence is. However, this is problematic for highly uncertain risks where quantitative likelihoods are not forthcoming.

Existential risks are particularly problematic for precautionary approaches, which has been dubbed the "black hole challenge": there does not seem to be any limit to the amount of time and resources that should be spent on investigating them (Munthe 2016)[129]. This is made worse by the potential infinitude of possible existential risks: Munthe suggests that the existential risk researcher has to deal with all of them. Yet some risks clearly have higher priority and epistemic support than others: we know mass extinction from asteroid impacts is more likely than from supernovas (due to empirically based probability estimates in astronomy), we have credible arguments for nuclear war being more likely to end civilization than asteroids, there exist logical but controversial arguments for the possibility of superintelligent AI being an existential threat, while there are no arguments for the possibility that dropping pencils will cause planetary devastation. This produces a rough hierarchy of credence and research priorities. However, there might not be any simple answer to the correct method of handling low-probability theories that predict large risks (Ćirković 2012)[130].

As mentioned above discounting the future, reducing its value at some fixed percentage per unit of time, has the effect of cutting off essentially all of the future beyond a certain horizon from consideration. Even if the value of all future generations is (say) 1050 lives, at 5% discounting we should ignore them if they occur more than 3000 years in the future. Hyperbolic discounting is more sensitive for remote futures, but has its own problems with time-inconsistency.

If disaster magnitudes have a heavy tail distribution with a low index (i.e. distributed as $\Pr[X > x] \propto x^{-a}$ or similar, with $a < 1$) there is a somewhat similar problem: the expectation diverges as time goes by, and the set of events that have happened tends to be totally dominated by the largest event. Hence effective mitigation aiming at reducing the

[129] Munthe, *The Black Hole Challenge*.
[130] Ćirković, "Small Theories and Large Risks—Is Risk Analysis Relevant for Epistemology?"

probability times the impact should presumably focus on reducing the risk of the largest events rather than the smaller ones. Yet the median event will be of a very different character than the dominant ones, and mitigation focusing on the large ones without effect on small ones will look inefficient. In many practical situations, this is politically impossible.

III.7.7 SOCIETAL ISSUES

III.7.7.1 RISKS TO AND FROM SOCIETY

While some risks to humanity are natural hazards that are not due to human agency in any form most scenarios of global catastrophic and existential risk involve at least some form of human agency. This might consist of creating preconditions for the hazard (e.g. global travel making pandemics riskier, new technology enabling new risks such as powerful, autonomous AI) and human actions triggering and participating in the risk (e.g. war) (Häggström 2013)[131]). The most human-dependent risks involve threats where human activity is at the core of the problem (e.g. breakdowns of societal stability, systemic risks, devolution to degenerate states).

Actual disasters rarely involve a simple chain of cause and harm but emergent patterns of contributing effects, especially when human society is involved. A model of how compound risks can act is the synchronous failure model of Homer-Dixon et al. (2015)[132]. Multiple stresses (such as climate change, resource shortages, or conflicts) can interact and accumulate in a social-ecological system, pushing it towards a state where its coping capacity is diminished. Different subsystems become coupled because they require support from each other to function in the stressed state. When a crisis occurs (either externally triggered or because of an internal component finally gives up) it rapidly cascades through the system, spreading between subsystems and causing the whole to fail. Simultaneous damage is often multiplicative in

[131] Häggström, *Here Be Dragons*.
[132] Homer-Dixon et al., "Synchronous Failure."

severity. This suggests that it is beneficial to look at pathways and stages of the risk evolution.

Cotton-Barratt and Sandberg (2017)[133] classified existential risk based on the three stages of its unfolding and how they break through the existing mitigation processes. The motivation for this classification is to analyse the policy implications rather than the natural/technological processes they involve or what life support systems they undermine. The first stage is the origin: what entity/entities originate the hazard, and why does normal risk avoidance mechanisms fail? The second is the process of hazard unfolding, and why responses to it fail. The third one is the harm: how does it cause permanent damage, and why does resilience fail? (Figure 4)

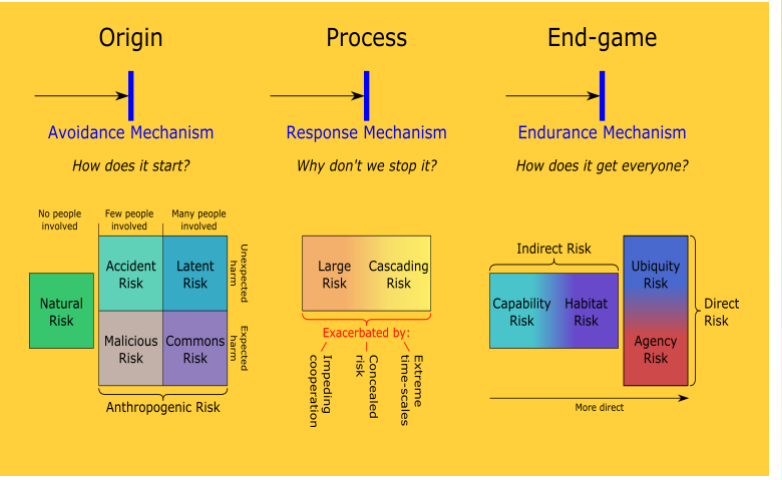


Figure 4 : Classification of existential risks by origin, process and end-game after (Cotton-Barratt & Sandberg 2017).

The first stage relates to the question of how the risk bypasses avoidance mechanisms. Either it is a natural risk we fail to anticipate or lack the ability to head off, or it is anthropogenic. In the latter case, some risks have a localized origin in a small group while others are due to societal or large-scale activities. Avoidance can fail because the harm is unexpected (an accident or a latent risk) or expected but hidden (malicious risk) or expected but not stopped due to group coordination

[133] Cotton-Barrat and Sandberg, "Classifying Risks of Human Extinction."

failure (commons risk). Note that "commons risk" also includes conflicts as a subset.

The second stage deals with why the response is insufficient. There are two main reasons: the effect may be extremely large, or it is a cascading effect that grows in size (e.g. a pandemic). Certain properties impede response: coordination may be impeded (e.g. by a communications breakdowns or information concealment (Chernov & Sornette 2015)[134]), the risk may be under-recognized (e.g. a long incubation infection), or it occurs on extreme timescales (both very fast and very slow).

The third stage deals with how all or most humans are harmed. This can occur directly through an effect affecting every location or individual (a ubiquity risk) or an effect deliberately seeking out people (agency risk). It can also indirectly cause extinction or long-term reduction in potential by permanently reducing human capability (e.g. by knocking society down to a primitive or limited state or damaging cognition), or by harming the environment humans live in (habitat risk).

Societal effects strongly influence many of these possible pathways. Insufficient oversight enables malicious risk, errors in human coordination increases commons risk, and insufficient shared insight latent risk. Commons risks are entirely due to social factors. Risks impairing cooperation clearly interact with existing cooperation levels. Vulnerable social capabilities (e.g. trustworthy sharing of scientific information or fragile global supply chains) make capability risks worse. Agency risk obviously is tightly linked to societal structures and incentives that modulate what human or machine agency is used for.

Society can clearly contribute to impairment or improvement of risk identification/prevention, preparedness/response, local and global endurance, as well as risk management itself. While object level interventions such as vaccine stockpiles, underground refuges and emergency plans are necessary, the processes leading to them being

[134] Chernov and Sornette, *Man-Made Catastrophes and Risk Information Concealment*.

created, maintained and used are also essential for global risk mitigation.

III.7.7.2 DANGEROUS GROUPS

There exists a "myth of mass panic" that sometimes impair planning for disasters. It is the belief that exaggerated or irrational fear would spread through contagion in an affected group, leading to unrestrained escape behaviour and breaking social rules. This has led to decisions to restrict information to the public about disaster preparation or even imminent risks, as well as justifying military rather than humanitarian responses to disasters. This can exacerbate disaster effects because it undermines collective resilience and fails to make use of the positive abilities of involved people (Drury, Novelli & Stott 2013; Chernov & Sornette 2015; Dezecache 2015)[135].

Another likely non-problem is doomsday cults. While millennialism is not uncommon and groups welcoming an expected end of the world have always been present, it is rare for such groups to take active steps to practically bring about the end of the world. To deliberately seek to cause the end of the world requires (1) a reason to think this would be a positive outcome, (2) that this is achievable, (3) that one needs to act to actually achieve this, and (4) that one can acquire the means to achieve it. Religious groups supporting (1) still largely reject (2) and (3) since they regard the end as part of a divine plan outside of human agency. They may pursue (4) through fulfilment of prophecy (or, in some cases, magic) which likely does not pose an actual risk. In order to actually fulfil all four conditions a rather peculiar theology where the divine plan presupposes active human assistance using concrete means is required; the closest so far may have been Aum Shinrikyo. Apocalyptic thinking can cause harmful behaviour and short time horizons, but rarely lead to active globally harmful measures (with some exceptions

[135] Drury, Novelli, and Stott, "Psychological Disaster Myths in the Perception and Management of Mass Emergencies"; Chernov and Sornette, *Man-Made Catastrophes and Risk Information Concealment*; Dezecache, "Human Collective Reactions to Threat."

such as policymakers supporting increases of tension in the Middle East in the hope to fulfil apocalyptic prophecy).

It is worth noting that secular groups accepting (1) may pose a greater risk than religious groups, since they do not have to assume the theological complications and may be more realistic in pursuing (4). Such groups might include or negative utilitarians that believe human life has net negative value, or deep ecologists that think Earth would be better off without humans. While so far achieving the end of the world has been outside individual or group abilities, more powerful technology is making causing large-scale harm more feasible. It is hence rational to be somewhat concerned about malicious risk.

However, historically, the largest deliberate impositions of risk have been from reasonable people doing the wrong thing for good reasons. The Cold War mutually assured destruction nuclear balance was a deliberate creation of significant risk to the survival of the species in order to achieve the goal of national security.

III.7.7.3 COGNITIVE AND CULTURAL LIMITATIONS

Cognitive biases, the systematic deviations from rational thinking most people exhibit, have important effects on planning and responding to extreme risks (Yudkowsky 2008)[136]. While these biases have largely been helpful heuristics for good-enough planning in our natural environment they are badly adapted to the extreme risk domain. The availability heuristic makes us underestimate the likelihood of unprecedented events (or overestimate it if it is often reported in fiction or media). Scope neglect makes us insensitive to the number of lives involved, making willingness to help scale sublinearly with the size of the problem. Without rich context information people are bad at judging differences between low probability events (Kunreuther, Novemsky & Kahneman 2001)[137]. These factors are amplified in social settings since individuals affected by them then transmit biased information and behaviour to

[136] Yudkowsky, "The Value Loading Problem."
[137] Kunreuther, Novemsky, and Kahneman, "Making Low Probabilities Useful."

each other, potentially creating a false consensus. This makes constructing risk management strategies that are resistant to behavioural biases extra important for extreme risks (Kunreuther & Heal 2012; Wiener 2016)[138]

A peculiar category of risks is warning shots: disasters so large that they cannot be ignored and cause protective action ("never again"), preventing future even larger threats. Such disasters would be useful in the long run despite their direct disvalue. In a sense, they are making the availability heuristic work for us. Unfortunately, when people are not harmed, they attribute this to safeguards being adequate and subsequent downgrade estimates of probability and harm (Dillon & Tinsley 2008; Dillon, Tinsley & Cronin 2011)[139]. There is hence a real possibility that global disasters, if considered the wrong way, can harm further disaster readiness. It also only works for extreme risks that have smaller instances than terminal ones (e.g. pandemics, wars, terrorism rather than supernovas or AI takeover).

Current existential risk research is about risks regarded as immanent: actual or potential risks. However, the cultural concept of apocalypse is transcendent: it is about meaning and how we choose to view history. We project our hopes and fears onto the future, often using fiction and cultural narratives. This imagery can both help and hinder risk response. While public understanding of certain risks such as major asteroid impacts likely have been helped by disaster movies (although giving an overly optimistic idea of intervention options) and depictions of the consequences of nuclear war strongly supported calls for disarmament (Kuznick 2007)[140], less concrete risks are often transformed into concrete and personal stories with convenient solutions that may be misleading (e.g. the persistence of Terminator imagery as a response to serious discussion about AI risk).

[138] Kunreuther and Heal, "Managing Catastrophic Risk"; Wiener, "The Tragedy of the Uncommons."
[139] Dillon and Tinsley, "How Near-Misses Influence Decision Making Under Risk"; Dillon, Tinsley, and Cronin, "Why Near-Miss Events Can Decrease an Individual's Protective Response to Hurricanes."
[140] Kuznick, "Prophets of Doom or Voices of Sanity?"

III.7.8 CONCLUSIONS

Mitigating extinction risk is an undersupplied global public good. Since it is non-excludable and non-rivalrous there is a free-rider problem (non-participants gain the benefit without having to pay) and each producer of risk reduction would only gain a fraction of the total benefit. This is amplified by the transgenerational nature of risk reduction: most of the benefit will accrue to future generations. In principle, the value to them of our present preventing extinction is near-infinite, but they cannot pay us any compensation (Matheny 2007; Bostrom 2013)[141].

Yet a moral case can be made that existential risk reduction is strictly more important than any other global public good.

We clearly need to improve our capacity to handle existential and globally catastrophic risks. This would include improving collective insight into the problem, technology foresight, and coordination ability. We need to construct institutions, norms, and principles for extreme risk mitigation that help us overcome the undersupply of risk mitigation and the problems of choosing rational actions. We need to encourage a long-term, hopeful attitude to get the societal motivation to handle the problem.

It might be that building humanity's capacity to meet large future existential risks may be more important than marginal reductions today. But we need to survive long enough to become wise!

[141] Matheny, "Reducing the Risk of Human Extinction"; Bostrom, "Existential Risk Prevention as Global Priority."

III.8 RISKS AND RISK MANAGEMENT IN SYSTEMS OF INTERNATIONAL GOVERNANCE (CATHERINE RHODES)

Catherine Rhodes is Academic Project Manager at the Centre for the Study of Existential Risk, working across its research projects. Her work has broadly focused on the interactions between, and respective roles of, science and governance in addressing major global challenges.

III.8.1 INTRODUCTION

International governance systems will have an important role in management of existential and global catastrophic risks, which by their nature have global impacts, and require coordinated international responses. This paper outlines some of the main contributions international governance systems can make to risk management and draws out some of the ways in which they can fail and be a source of risk. This motivates work to better understand and analyse international governance systems' capabilities and deficiencies in relation to specific sources of existential and catastrophic risks, and so the paper is framed by an effort to outline steps through which this can be done as a contribution to building broader research agendas on managing these classes of risk.

III.8.2 EXISTENTIAL AND GLOBAL CATASTROPHIC RISKS

Other presentations at the Colloquium provided detailed outlines defining and describing sources of existential and catastrophic risks. What follows is a brief outline of how they are understood for the purposes of this paper.

Existential risks are those that threaten the survival of humanity in its entirety, or civilizational collapse to a point from which it cannot recover; global catastrophic risks are those that are a threat to the survival of at least 10% of the global population. These risks can be of natural origin or emerge from or in combination with human activity and technology. While some may be associated with single events, such risks are more likely to result from events in combination and with cumulative and cascading effects, particularly because of the level of interdependence and connectedness of much of our critical infrastructure, along with the capacity of human systems to cope at times of severe social disruption.

As short hand, the term extreme risks is generally used in the paper to refer to both existential and global catastrophic risks.

III.8.3 INTERNATIONAL GOVERNANCE

International governance is formed of rules, norms, institutions and other mechanisms designed to influence state behavior in the absence of supranational government. International governance serves to coordinate state action in areas of high international interdependence, where separate action by individual states will be insufficient to address issues of common concern or achieve common interests.

In this paper, the term 'international' is used to refer to governance components that are potentially universal in scope, that is open to any state to subscribe to or participate in without restriction on, for example, geographic or economic grounds (so it does not cover e.g. regional laws or bilateral agreements).

States are a key focus in international governance because they remain the main actors in this arena. They create and are the subjects of international law; they form the membership of international organizations, and retain decision-making authority within them. This emphasis is not meant to imply that other actors – such as corporations, NGOs, scientific and technological communities – are unimportant. Indeed, some of these actors may be more significant than states in driving future technological change, and they may be difficult to reach through traditional, state-focused forms of governance. The governance systems outlined in this paper are likely to retain some importance, but a shift to other forms of governance will be a necessary complement in coming years.

III.8.3.1 KEY COMPONENTS OF INTERNATIONAL GOVERNANCE

The main components of international governance include: norms, rules, institutions – such as international organizations - and a range of other coordination mechanisms. Two main categories of rules are distinguished in international law as 'hard' and 'soft' law – the former referring to international treaties and conventions, which are legally-binding on those states that sign and ratify them; and the latter referring to voluntary standards, guidelines and codes. Both types of rules influence state behavior, and soft law does not necessarily have a weaker effect.

In the context of extreme technological risks, Catherine Rhodes is interested in understanding the intersection and combination of risk stemming from technologies and risk stemming from governance (or lack of it). Her expertise is in international governance of biotechnology, including biosecurity and broader risk management issues.

Norms include certain core principles, referred to as customary international law that emerge through state practice, and are considered universally applicable (whether or not a state has subscribed to a particular treaty). One example are principles from the Geneva Conventions, such as the prohibition on torture, inhuman and degrading treatment. There are other normative principles outlined in various declarations and appearing in the preamble to many treaties, which do not prescribe particular actions but which may guide behavior.

International organizations are formed by states, frequently in association with particular rules, to which they are assigned oversight and administrative responsibilities. Their member states form their governing bodies, retain decision-making powers, and are generally their main source of funding. Day to day operations tend to be run by a secretariat (led by a secretary- or director- general), and a range of standing and ad hoc bodies may be set up to assist their operation.

Other coordination mechanisms are established to support the functioning of international rules and organizations, for example through information exchange and reporting mechanisms, advisory bodies, surveillance and response systems, and expert networks.

III.8.3.2 INTERNATIONAL GOVERNANCE SYSTEMS

International governance systems can be conceptualised as sets of rules, norms, organizations and other coordination mechanisms that together govern behavior in relation to a particular area of concern – for example a field of technology, or a particular global challenge. (Such systems are also referred to in international relations literature as regimes.)

While these governance components can act in combination to shape action in a particular area, they were often not originally designed to do so, having been developed at different times, for different (though not necessarily incompatible) purposes, and often in separation from one another. For example, in the area of biotechnology governance, rules for disease control, environmental protection, arms control, trade and human rights are all applicable, but were developed for different purposes and in different historical (and political) contexts.

Because so many issues now require international responses, there is rarely a single rule to look to for addressing a particular issue or technology. Understanding the interactions between components is important when assessing the capabilities of governance systems to manage particular risks. Risk management will often be only one objective of a particular rule or governance mechanism. When analysing governance systems, if we concentrate solely on components clearly identifiable as risk management focused, important interactions are likely to be missed.

III.8.3.3 STEPS IN UNDERSTANDING / ANALYSING GOVERNANCE SYSTEMS

Initial steps to take when seeking to analyse international governance systems include:

- Identifying the goal that is of interest – in this case management of extreme risks.
- Identifying the general area of interest – for example a particular technology or area of human activity.
- Matching the potential impacts of this technology or activity (for health, the environment, etc.) with areas of international concern.

This will indicate the areas to search in order to identify relevant components of the governance system.

Applying these steps using biotechnology as the area of interest:

- Potential impacts include: benefits and risks for human, animal and plant life and health; benefits and risks for ecosystems and biodiversity; potential application for hostile purposes; etc.
- Areas of international concern that these match with include: health and disease control; conservation of biodiversity; arms control; etc.

Relevant governance components can then be searched for within those areas.

After identification of relevant components, further steps in analysing international governance systems will then involve: understanding how those components interact; establishing what they can contribute to risk management; and identifying any gaps and deficiencies.

III.8.4 KEY COMPONENTS OF RISK MANAGEMENT IN INTERNATIONAL GOVERNANCE SYSTEMS

Here, some of the key components of risk management in international governance systems are described with illustrative examples from the area of biological risks. A more extensive – though not exhaustive – list of examples is provided in Table 3 toward the end of the paper.

The main sources of biological risks include:

- Those that are naturally occurring, such as pathogens of natural origin;
- Those that are human induced, such as antimicrobial resistance; and
- Those more directly caused by human activity, including through:
    - accidents (for example accidental release of material from laboratory containment facilities);
    - deliberate actions with benign intent and unintended consequences (for example release of a biological control agent that becomes an invasive species); and
    - deliberate actions with malign intent such as warfare, terrorism or oppressive uses.

Components of risk management include:

III.8.4.1 SPECIFIC RISK ANALYSIS PROCEDURES

Some international rules incorporate specific guidance on the conduct of risk analysis. Providing standardised methods of risk analysis promotes harmonization of procedures among states, and enables informed decision making. The import risk analysis procedure detailed in chapter 2.1 of the Terrestrial Animal Health Code, for example,

facilitates decision making on whether to import animals and animal products from one territory into another. It involves a four-stage process of hazard identification, risk assessment, risk management and risk communication; with risk assessment being further broken down into the stages of entry assessment, exposure assessment, consequence assessment, and risk estimation. This allows for consideration of the environment into and routes through which the hazard might be introduced and have impact, as well as of the hazard itself.

III.8.4.2 REPORTING REQUIREMENTS

Reporting requirements and associated information systems can be utilised to support the implementation of rules and the broader work of international organizations, for example they can facilitate timely responses to events and act as the trigger for certain actions to manage risks. In the biological risk area, for example, reporting should capture disease outbreaks at an early stage so that action can be taken to prevent their international spread. Reported information – for example on the disease status of an exporting country – can form a key input for risk analysis.

III.8.4.3 SURVEILLANCE AND MONITORING SYSTEMS

Surveillance and monitoring systems can perform a similar role in providing advanced warning that can trigger risk management activities. There are several systems used by international organizations to identify and track pest and disease outbreaks and inform responses. (These would probably capture deliberate disease events, but there is no dedicated system for this.) Some of the systems are general, for example the World Health Organization's Global Outbreak Alert and Response Network for human disease outbreaks, and others more specific, such as its Global Influenza Surveillance and Response System .

III.8.4.4 EXPERT NETWORKS

Surveillance, monitoring and response systems and processes are often supported by expert networks. These can play a role in, for example, the development of diagnostics and treatments, and give guidance on appropriate risk management responses to particular events. A Roster of Experts can be drawn on by the World Health Organization (WHO)

Director-General for advice when a potential 'public health emergency of international concern' is notified under the International Health Regulations. Expert networks have also been formed as collaborations between international organizations, such as OFFLU – a World Animal Health Organization (OIE) and Food and Agriculture Organization (FAO) joint network convened to facilitate research and provision of advice to international organizations and states, for the prevention, diagnosis, surveillance and control of animal influenzas.

III.8.4.5 CAPACITY BUILDING

International rules often include capacity building commitments and requirements, and international organizations play a supporting role for capacity building efforts. Capacity building may directly relate to risk analysis and response capabilities, for example training of local personnel in risk analysis, or boosting the capacity of emergency response systems; it can also involve more general scientific, technological, administrative and legislative capacity building.

In the biological risk area, core capacity building requirements outlined in the rules relate to local and national detection, assessment and response to disease events. The International Health Regulations, for example, require states to achieve a substantial list of core capacity requirements for surveillance and response, and at airports, ports and border crossings. Mechanisms to support capacity building include some specific to one international organization such as the laboratory twinning program of the World Animal Health Organization, which helps to build veterinary expertise across regions, and collaborative efforts such as support the Standards and Trade Development Facility. International scientific cooperation and technology transfer are also strongly promoted by several of the rules and organizations.

III.8.4.6 SCIENCE ADVISORY PROCESSES

Science advisory and review processes associated with rules and organizations can help identify developments in science and technology with implications for risk management, including developments that can support risk management (e.g. improved biosensors and diagnostics), and those that may increase risk. The World Animal Health Organization's specialist commissions provide for regular review and

development of its international standards. This is an area in which soft law instruments can have an advantage in being more quickly and easily updated in response to new scientific information – in 2016, for example, chapters were added to the Terrestrial Animal Health Code on control of anti-microbial resistance, including recommendations on risk analysis methodologies (chapters 6.6 – 6.10).

III.8.4.7 PROHIBITIONS

Some activities are banned completely by international treaties; the normative force of such prohibitions has strong value, even where enforcement is challenging. These prohibitions are often quite general in nature, which has value when dealing with areas of scientific and technological advance – where more specific prohibitions may be rapidly overtaken by developments. In combination, the 1925 Geneva Protocol and the Biological Weapons Convention prohibit development, production, stockpiling and use of biological agents and toxins and associated equipment for non-peaceful purposes. States have reaffirmed that this "applies to all scientific and technological developments in the life sciences and in other fields of science relevant to the Convention".

III.8.4.8 NORM DEVELOPMENT AND PROMULGATION

International governance systems can also promote dissemination of norms. Over the last decade, the international community has placed increasing emphasis on the responsibility of life scientists to safeguard their research from misuse, and to minimise safety and security breaches through good laboratory practices. Examples include: the WHO's promotion of a biorisk management approach combining laboratory biosafety, biosecurity and bioethics, and promoting local responsibility for selection and implementation of appropriate controls; and encouragement of the development of codes of conduct, and education and training activities to promote cultures of responsible scientific practice from states parties to the Biological Weapons Convention.

III.8.5 GOVERNANCE FAILURES

There are a range of ways in which international governance systems can respond to risks, but even where there are a substantial range of components in place, identifying them does not give the full picture. Rules are not all implemented by all states to the same extent, nor in the same ways; they may at times be in tension or conflict with each other, and states may then prioritise compliance in different ways. There are also significant disparities in capacity to participate in the creation of international rules and in ability to effectively implement them. These are some of the reasons why it is important to understand the ways in which different governance components can interact, and the implications this can have, particularly where these cause problems for risk management.

In this section, some major contributing factors to governance failures are introduced; two examples from the biological risk area are then used to illustrate where several of these failures occur in practice.

III.8.5.1 POWER RELATIONS DOMINATING OUTCOMES

All states are nominally equal under international law. Of course, this is far from the case in practice and powerful states dominate many international processes, including which issues reach the international agenda in which forums; negotiation of treaties; and what actually gets implemented once agreed.

III.8.5.2 SHORT-TERMISM AND PURSUIT OF ECONOMIC COMPETITIVENESS

Short-term interests also tend to dominate the development and implementation of international governance, despite the long-term nature of many of the challenges and necessary responses. Pursuit of economic competitiveness is a strong driver in this, and a particular problem when trying to govern emerging technologies.

Reversion to pursuit of a narrowly conceived version of national interest in times of crisis

There is a tendency to revert to pursuit of a narrowly conceived version of national interest in times of crisis, with protection of short-term and immediate interests of a state's population prioritised over an effective global response, which can negatively impact risk management leaving everyone more vulnerable.

Large disparities in vulnerability to risk; capacity to manage risk; and capacity to access benefits of scientific and technological advances

The frequent commitments to address capacity building through support by developed states for activities such as technology transfer, provision of additional financial resources, and sharing of expertise, are inadequately fulfilled, and have been for several decades. This is becoming a barrier to international cooperation, particularly where agreement is sought for the control of particular technologies. International management of extreme risks may be severely damaged by the resulting lack of trust and reduced incentives for cooperation in governing emerging technologies.

III.8.5.3 COMPLEXITY

As the number of issues reaching the international agenda grows, there is an increasing overlap between governance activities in different domains and in the range of interactions between them, which means that outcomes can be difficult to predict and functionality harder to understand.

III.8.5.4 POOR COORDINATION

Some of the problems caused by complexity can be addressed through better coordination of overlapping governance domains, for example through cooperative activities between international organizations, but these activities face various constraints, particularly in terms of resourcing and political barriers.

III.8.5.5 BURDEN OF REGULATORY PROLIFERATION

All the different components of international governance require related activity at the national level – for example in creating and enforcing national implementing legislation, collecting data for

reporting, etc. They also create a need to participate in a growing number of governance forums and negotiations. This presents significant burdens for many states, weakening their participation in international governance activities and reinforcing inequalities in power relations.

III.8.5.6 INADEQUATE RESOURCING

Many international organizations are severely resource constrained and are mostly dependent on financial contributions from their member states, limiting the scope of their activities. This can mean that even when a governance need is recognised and an organization mandated to perform related activities, they are not able to do so. Likewise, there is limited support available to promote fuller participation by developing countries, despite explicit recognition of the problems this might cause, and funds to support national implementation activities are also generally under-resourced.

III.8.5.7 POOR DESIGN OF RULES

Poor design of rules can arise from issues of poor coordination – for example, where there is a lack of awareness of related governance processes and no understanding of how a new rule might impact them – it can also stem from inappropriate transfer of principles from one area to another, or from failure to consult with affected groups in order to understand whether rules will function as intended.

III.8.5.8 KEY COMPONENTS ABSENT

Sometimes key components needed for effective risk management have simply not been established; this may be due to lack of awareness of their importance, but it can also happen when a need is recognised but political agreement cannot be reached.

III.8.5.9 INABILITY TO KEEP PACE WITH DEVELOPMENTS IN SCIENCE AND TECHNOLOGY

An inability of governance systems to keep pace with developments in science and technology is more obviously a problem where technological risk is a focus of governance, but good quality scientific input is frequently needed in other risk areas too. In some cases, there

are advisory or review processes in place, but these are not always adequate, can be politically dominated or captured by interest groups, or again may simply be absent.

The following examples demonstrate how these governance failures often occur in combination and the types of impact they can have.

III.8.5.10 PANDEMIC INFLUENZA

Pandemic influenza is a well-recognized severe global risk. As well as the rates of illness and deaths associated with an influenza pandemic, they are likely to have other disruptive effects. The UK Government, for example, in the 2015 edition of its National Risk Register for Civil Emergencies, when outlining the potential consequences of a pandemic for the UK suggested that "in the absence of early or effective interventions to deal with a pandemic" consequences would include "significant social and economic disruption, significant threats to the continuity of essential services, lower production levels, and shortages and distribution difficulties".

The handling of the 2009 H1N1 influenza pandemic, and earlier problems in the international system for sharing influenza viral samples, highlight several failures of international governance.

Among other significant disparities, there is huge variance in countries' vulnerability to outbreaks, capacity to manage them effectively, and ability to access diagnostics, vaccines and treatments. A study in the Lancet highlighted some of the implications of these gaps, projecting that in the event of an influenza pandemic with similar mortality rates to that of the 1918-1920 pandemic, there would be approximately 62 million deaths (based on world population levels in 2004), 96% of which would be in developing countries.

During the 2009 pandemic, all vaccine supplies went to a few developed countries that had advanced contracts with manufacturers. They sourced these supplies for their own populations, which were not in the most affected areas, nor the most vulnerable to the impacts of the pandemic. Fortunately, that pandemic was not particularly severe. However, prioritising the immediate interests of their own populations above an appropriate global response (e.g. deploying a vaccination

strategy most likely to contain the outbreak) could have lengthened the duration of the pandemic and contributed to its international spread. A vaccine stockpile (initially of 150 million doses) with centralized distribution has been established by the World Health Organization, but this does not appear to be at a sufficient level (given likely size of infected population), and there seems to be little in place to prevent recurrence of the 2009 situation in future outbreaks.

It is notable that this happened despite WHO member states acknowledging a breakdown in trust in the international influenza surveillance system (then known as the Global Influenza Surveillance Network) in 2006 and 2007, when Indonesia blocked sharing of samples from human cases of H5N1 avian influenza on the basis that these were being released to companies who were producing products unaffordable to its own population.

Indonesia drew on principles from the Convention on Biodiversity to justify its action. Acceptance of this approach in the World Health Organization's 2011 Pandemic Influenza Preparedness Framework has inappropriately introduced principles for the conservation of biodiversity to the governance of pathogens, raising serious concerns about delays and obstructions to time urgent global public health efforts.

III.8.5.11 THE BIOLOGICAL WEAPONS CONVENTION

The Biological Weapons Convention has widely acknowledged deficiencies affecting management of risks from scientific and technological advances.

Capacity building commitments in Article X of the Convention have been inadequately addressed and this is a regular point of complaint during meetings of states parties, and may be a disincentive for developing countries to engage in further cooperation to control emerging technologies.

One of the components most notably absent from the Biological Weapons Convention is a verification mechanism to check that activities are compliant with its provisions. This is a serious weakness given that it can be difficult to distinguish peaceful from non-peaceful activities, and

a lack of confidence in compliance can erode trust and promote arms-racing. There is also no guidance on risk analysis associated with the Convention, for example for judging whether particular lines of research carry a significantly heightened risk of diversion to non-peaceful purposes.

The following quote is from a report of the InterAcademies Panel on science and technology developments with implications for the Biological Weapons Convention, based on the five-year period between the Convention's 7th and 8th Review Conferences (2011-2016):

> *"Recent advances could also facilitate almost every step of a biological weapons program, and technological barriers to acquiring and using a biological weapon have been conspicuously eroded since the Seventh Review Conference.*
>
> *The sometimes-formidable challenges associated with the synthesis of existing agents and the development of novel agents have been overcome in some cases by using gene transfer and other biosynthetic engineering approaches.*
>
> *Modification of biological agents enables them to be more easily optimized for specific purposes… Developments in scale-up and production technologies have changed production signatures. Less space and time are needed…*
>
> *It is also now easier to deliver a biological agent given advances in areas such as nanoparticles and sophisticated modeling of dispersal patterns using the techniques of aerobiology."*

This highlights why it is a serious concern that the Convention has limited ability to keep pace with such developments. There are very limited opportunities to review science and technology developments within the Convention. The current system, relying on states to voluntarily provide background documents to review conferences, is

neither frequent nor in depth enough to adequately capture and address potential implications for the Convention.

Inadequate resourcing is also a major problem for the Biological Weapons Convention. States parties are behind on their financial contributions for its meetings' budget and the budget for its small Implementation Support Unit. Its very limited institutional structure is itself recognised to be inadequate; further funding is needed for its expansion and for any improved science and technology review process.

III.8.6 CONCLUDING POINTS

International governance systems' role in and capabilities for risk management warrant further research and there are several areas in which it is clear that improvement is needed. Some improvements, for example in avoiding pursuit of narrowly conceived short-term national interests above effective risk management, will be generalizable across risk areas, but it is also necessary to carefully analyse the governance landscape of specific risk areas, to identify particular instances and impacts of deficiencies. As well as undertaking the steps listed in

Table 4, which can usefully form part of a research agenda on management of existential and global catastrophic risks, there is also some more focused work that might be done to analyse existing risk analysis processes.

This includes research addressing the following questions: Are these processes robust enough and appropriate to existing needs? Are they readily adaptable and responsive to changes in e.g. technology? And are they widely and consistently implemented? And there is a broader research question that needs to be addressed on how to move beyond traditional state-centric governance activities in order to manage global risks more effectively.

III.8.7 TABLES

Table 3– Risk management components in international governance

Component	Examples
Specific risk analysis procedures	Import risk analysis in chapter 2.1 of the Terrestrial Animal Health Code
	International Standards for Phytosanitary Measures (ISPM) No.2 Framework for Pest Risk Analysis
	Decision instrument on public health emergencies of international concern in International Health Regulations
	Risk assessment for transboundary movements of living modified organisms in Annex III the Cartagena Protocol
	Safety assessment for foods produced using / derived from recombinant DNA microorganisms, plants or animals under the Codex Alimentarius
Reporting requirements	Notification of potential public health emergencies of international concern under the International Health Regulations
	Notification of listed diseases under the Terrestrial Animal Health Code
Surveillance and monitoring networks	Global Outbreak Alert and Response Network (human disease outbreaks)
	Global Influenza Surveillance and Response System

	World Animal Health Information System
	Emergency Prevention and Response Systems of the Food and Agriculture Organization
Expert networks	World Health Organization Reference Centers
	World Health Organization Collaborating Laboratories
	International Health Regulations' Roster of Experts
	World Animal Health Organization Reference Centers
	World Animal Health Organization Collaborating Laboratories
	United Nations Secretary General's Mechanism for Investigating Alleged Use of Chemical and Biological Weapons Roster of Experts
	World Animal Health Organization / Food and Agriculture Organization collaborative network on animal influenzas (OFFLU)
Capacity building commitments and requirements	International Health Regulations' core capacity requirements
	Standards and Trade Development Facility
	Article X of the Biological Weapons Convention

	Article 22 of the Cartagena Protocol on Biosafety
Science advisory / review processes	Submission of background documents and consideration of science and technology developments with implications for the Convention at Review Conferences of the Biological Weapons Convention
	Specialist Commissions of the World Animal Health Organization
	Subsidiary Body on Scientific, Technical and Technological Advice of the Convention on Biological Diversity
Prohibitions	Article I of the Biological Weapons Convention; 1925 Geneva Protocol
Norm development and promulgation	Recommendations on the development of awareness-raising, education and training, and codes of conduct for life scientists by states parties to the Biological Weapons Convention
	World Health Organization's biorisk management approach
	Responsible Life Sciences Research for Global Health Security

Table 4–STEPS FOR ANALYSING INTERNATIONAL GOVERNANCE SYSTEMS

Identify the goal
Identify the general area of interest

Match the potential impacts of this technology / activity with areas of international concern to find out where to look for components

Identify relevant governance components

Understand how they interact

Establish what they can contribute to risk management

Identify any gaps and deficiencies

Establish whether they are effectively implemented and what impacts they have in practice

III.9 CYBER, NANO, AND AGI RISKS: DECENTRALIZED APPROACHES TO REDUCING RISKS (CHRISTINE PETERSON, MARK S. MILLER, AND ALLISON DUETTMANN)

"If men were angels, no government would be necessary. If angels were to govern men, neither external nor internal controls on government would be necessary. In framing a government which is to be administered by men over men, the great difficulty lies in this: you must first enable the government to control the governed; and in the next place oblige it to control itself."

—James Madison

Christine Peterson writes, lectures, and briefs policymakers and the media on coming powerful technologies - especially nanotechnology and artificial intelligence. She is Co-founder and Past President of Foresight Institute, the leading nanotech public interest group. She serves as an Advisor to the Machine Intelligence Research Institute, the Global Healthspan Policy Institute, the National Space Institute, and Ligandal, Inc.

III.9.1 ABSTRACT

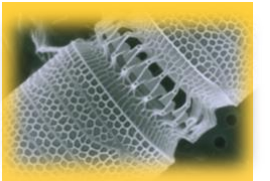
The aim of this paper, rather than attempting to present one coherent strategy for reducing existential risks, is to introduce a variety of possible options with the goal of broadening the discussion and inviting further investigation. Two themes appear throughout: (1) the proposed approaches for risk reduction attempt to avoid the dangers of centralized "solutions," and (2) cybersecurity is not treated as a separate risk. Instead, trustworthy cybersecurity is a prerequisite for the success of our proposed approaches to risk reduction.

Our focus is on proposing pathways for reducing risks from advanced nanotechnology and artificial general intelligence.

Nanotechnology can be divided into stages; even at the most advanced stage, society should be less worried about biology-style accidents than deliberate abuse. Development of nanotech weapons would not be detectable by current weapons monitoring techniques. An automated monitoring system, if based on sufficiently secure software foundations and physical arrangements, could serve as the basis for an arms control enforcement mechanism. Design needs to be open and decentralized to build the required public trust.

Civilization, taken as a whole, is already a superintelligence. It is vastly more intelligent than any individual, it is already composed of both human and machine intelligences, and its intelligence is already increasing at an exponentially accelerating rate. Civilization as a whole does not "want anything"; it has no utility function. But it does have a tropism—it tends to grow in certain directions. To the extent that its dominant dynamic emerges from non-coercive, non-violent, voluntary interactions, it is already shaped by human values and desires. It tends, imperfectly, to climb Pareto preferred paths. Society would address the value alignment problem more effectively by strengthening this dynamic rather than trying to replace it. No designed utility function would clearly serve human happiness better, and no replacement for civilization's dynamics is likely to be adopted anyway.

While still controversial, there is increasing concern that once artificial general intelligence fully surpasses the human level, human skills will have little or no economic value. The rate of economic growth will be extraordinary, but humans will have little comparative advantage. Will civilization still serve human preferences? When growth is extraordinary, so are returns to capital. The least-disruptive approach may be a one-time, gradual distribution of tradeable rights to as-yet unclaimed resources in space.



Christine Peterson serves as an Advisor to the Machine Intelligence Research Institute, the Global Healthspan Policy Institute, the National Space Institute, and Ligandal, Inc. She is credited with coining the term 'open source software.'

III.9.2 INTRODUCTION

The long-term goal can be defined as human survival in the face of various existential risks, including those posed by both advanced nanotechnology and artificial intelligence that exceeds human intelligence. Many risk-oriented organizations, e.g., Center for the Study of Existential Risk (CSER), Future of Humanity Institute (FHI), and Future of Life Institute (FLI), and many future-directed philanthropists, e.g. Elon Musk, Jaan Tallinn (Tallinn, 2012)[142], and Peter Thiel, rate artificial intelligence as one of the most important issues facing humanity, if not as the single most important issue to work on for enabling a positive future.

[142] Tallinn, "CSaP Distinguished Lecture: The Intelligence Stairway - Networks of Evidence and Expertise for Public Policy."

Definition Nanotechnology

Nanotechnology is a term used to describe precise control of the structure of matter. Four levels are defined, which describe the production of progressively more capable atomically-precise (AP) (and partially AP) nanosystems (Drexler, Pamlin, 2013)[143]. Level 1 includes chemical synthesis, nanomaterials synthesis, nanolithography, and biotechnology; Level 2 includes AP macromolecular self-assembly and AP mechanical manipulation; Level 3 includes biomimetic or machine-based productive nanosystems; and Level 4 includes high-throughput atomically-precise manufacturing. While defensive military use scenarios have been described for the earlier stages of nanotechnology (Kosal, 2009)[144], these levels have not been seen as involving catastrophic or existential risks. More substantial risks are expected at the highly advanced Level 4 stage, which is substantially longer-term (Drexler, 2007)[145].

Definition AI

Artificial intelligence (AI) that exceeds human-level intelligence in all intellectual tasks is described using a variety of terms, including superintelligence (Bostrom, 2014)[146], advanced AI (Russell, Norvig, 2009)[147], smarter-than-human AI (Soares, 2016)[148], strong AI (Muehlhauser, 2013)[149], and Artificial General Intelligence (Goertzel, 2014)[150]. For simplicity, this paper will use the term Artificial General Intelligence (AGI hereafter). While AGI is sometimes used to describe an AI whose intelligence level is merely equal to that of a human, it is widely assumed that once AI reaches human level intelligence, it will soon

[143] Drexler and Pamlin, "Nano-Solutions for the 21st Century."
[144] *Nanotechnology for Chemical and Biological Defense | Margaret Kosal | Springer*.
[145] "Interview."
[146] Bostrom, *Superintelligence*.
[147] Russell and Norvig, *Artificial Intelligence*.
[148] Soares, "The Value Learning Problem."
[149] Luke Muehlhauser, "When Will AI Be Created?"
[150] Goertzel, "Artificial General Intelligence."

surpass this level (Tallinn, 2012)[151]. With potential risks associated with the development of AGI becoming a greater concern, the new field of AI safety research, while still in its infancy, is growing fast and with high quality (e.g., OpenAI's launch in 2015, DeepMind's safety board established by Demis Hassabis in 2016). For an overview of different safety approaches represented in the field, see Mallah, 2017[152].

Biotech and cyber risks arrive earlier than, and are subsets of, nano and AGI risks

In the case of both advanced nanotechnology (physical technology) and AGI (software technology), there are closely-related concerns that arise earlier in time. Both advanced nanotech and biotechnology are based on systems of molecular machines, with biotech restricted to molecular machines similar to those which nature has discovered; thus, it is theoretically a subset of the broader category of molecular machine systems referred to as nanotechnology. (Biotech is sometimes referred to as "nature's nanotechnology.") Similarly, the cyber attack risks of today are a small subset of what will be possible in the future from AGI. Eventually, cyber attacks will be performed by AGIs, while today's cyber attacks are often performed by today's existing superintelligences, such as corporations and nation states.

Even these earlier risks are very challenging: in fact they can seem harder to address than the long-term ones, because they seem more real and concrete. Biotech dangers and cyber attack dangers relate more closely to the current world, so it is easier to see why they are so challenging, while Level 4 nanotechnology and AGI are still relatively abstract, so it is harder to see why they are difficult. However, a world safe against a level 4 nano-attack would be a world already safe against biotech attack. Likewise, a world safe against AGI would also be a world already safe against cyber attack. Biotech dangers and cyber attack are risks worth addressing in regard to making the world a safer place, both for the practical value of solving these very real problems and the additional benefit of learning strategies and designing institutions

[151] Tallinn, "CSaP Distinguished Lecture: The Intelligence Stairway - Networks of Evidence and Expertise for Public Policy."
[152] Mallah, "The Landscape of AI Safety and Beneficence Research."

applicable to the related longer-term challenges. For example, addressing cyber attack issues now will head off substantial concerns regarding cyber risk from the society's increasing vulnerability to attack on networked consumer products including threats to the control of self-driving cars (Kornwitz, 2017)[153].

Focus on scenario of sophisticated attacker

With regard to biotech attack and cyber attack, there are two types of sophisticated attack scenarios: (1) by nation states that prepare attacks as weaponized systems for potential use in war, and (2) by the iteration of open attacks becoming more sophisticated over time, with the information needed for the attacks getting commoditized, eventually enabling a wide variety of players to engage in attacks. An example for the second attack scenario, which is also known as the "script kiddie problem" in the cyber world, is the Stuxnet virus. This virus was one of the most elite pieces of malware before its workings became understood, commoditized, and reused by less-advanced attackers who could not have constructed the attack originally. Therefore, in both cases, for current purposes we can consider sophisticated attackers as the primary threat.

III.9.3 ESTABLISHING THE COMPARISON

III.9.3.1 BIOTECH ATTACK AND NANOTECH ATTACK

III.9.3.1.1 BIOTECH ATTACK

Biotech attack and nuclear attack both have physical component to monitor

In contrast with cyber attacks, biotech attack at least has a physical component involved, although the physical aspects can be very small, e.g., a small lab with a small number of people. Because there is a physical component, there is in principle something physically observable in the process, which suggests that one possible place to look for a useful precedent is the response to nuclear proliferation. The current nuclear weapon situation is still very concerning, but humanity

[153] Kornwitz, "The Cybersecurity Risk of Self-Driving Cars."

has survived since World War II without these weapons being used in battle again, and this is partly due to non-proliferation treaties backed by monitoring regimes, amongst other reasons.

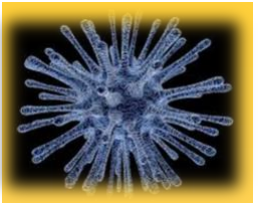
Biotech attack prevention requires far higher level of monitoring than nuclear, but similar to nano

The challenge ahead is that the physical objects that must be monitored in order to detect hostile nuclear weapons activity are relatively large-scale and easy to verify—with useful data available even from satellites in space—compared to what will be needed to monitor for offensive biotech use. Monitoring styles can be divided into three broad categories:

1. Traditional top-down, Big Brother-style surveillance, which tends to lead to abuse;
2. A symmetric system of surveillance plus sousveillance (upward-looking monitoring), advocated in The Transparent Society, which would be destructive of privacy but could hold abuses in check (Brin, 1998)[154]; and
3. A decentralized automatic network of agents with "confinement," in which information is not revealed to humans unless clear, pre-agreed, tripwire criteria are triggered; this would be difficult to implement but avoids abuses of category 1 and the loss of privacy from category 2.

An early experiment with category 3 was made by William Binney, at that time a senior official at the NSA (U.S. National Security Agency), who initiated a monitoring system called Thinthread to perform a legal, Constitutional version of surveillance based on filtering and automatic encryption. Data on individuals was only unencrypted "if a judge found probable cause to believe the target was connected with serious crime, including terrorism." Unfortunately, the program was cancelled just before 9/11 and eventually replaced by a similar system without the filtering and encryption protections (O'Cleirigh, 2015)[155].

[154] Brin, "The Transparent Society."
[155] Fiona O'Cleirigh, "Bill Binney, the 'Original' NSA Whistleblower, on Snowden, 9/11 and Illegal Surveillance."



The degree of monitoring required to prevent biotech attack would be comparable to a level of monitoring corresponding to a pervasive surveillance state. While finding an acceptable and working level of monitoring to detect hostile biotech would be very challenging (Omohundro, 2014)[156], it is very much like the level of monitoring required to detect hostile nanotech weaponization, so it is at least very similar with regard to problem domain. Since both involve systems of molecular machines, both require verification at the molecular level: a daunting challenge.

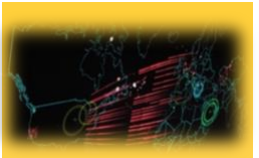
As societal transparency and surveillance have been increasing over time, it has become increasingly difficult for independent third parties such as terrorists to hide a secret research program to develop biotech weapons. However, the ongoing global illegal drug trade demonstrates that the level of transparency and surveillance now in place is not effective at preventing even a large-scale illegal societal dynamic. The very substantial financial flows connected with the illegal drug trade constitute an extra-legal institutional framework operating in secret. Unauthorized bioweapons labs could take advantage of these same extra-legal mechanisms; the illegal drug trade in this sense preserves areas of secrecy usable by non-governmental weapons efforts. For decades, governmental programs to address this issue have involved attempting to reduce the illegal drug flow, with marginal success and high costs in terms of corruption, similar to that seen from alcohol prohibition in the U.S. during the last century. The question arises: if drugs were legalized, how much of this extra-legal financial institutional framework would survive? If society becomes sufficiently concerned about terrorists developing bioweapons, this still-controversial approach to undermining terrorism may be tried (Miller, 1980s-90s)[157].

III.9.3.1.2 LEVEL 4 NANOTECHNOLOGY ATTACK

Desirability of Quantitative Risk Assessment

[156] Omohundro, "Autonomous Technology and the Greater Human Good."

[157] Miller, "A Series of Private Communications to Christine Peterson in the 1980s and 1990s."



Currently, Level 4 nanotechnology remains decades in the future. All of the other risks explored here are more urgent. But as with other technological risks, it is desirable to carry out a Quantitative Risk Assessment (QRA) as early as possible. This process, even when done very early in a risk field, helps clarify areas needing further work. For Level 4 nanotech, "it can sort out some of the more predictable risks" (Garrick, 2008)[158].

Nanotech security entails biotech security; both require high-level monitoring

The reason why a world that is safe against nanotech attack is one that is already safe against biotech attack is that all of the attack vectors that are concerning for biotech are forms of attack that nanotech could engage in (i.e., highly advanced nanotech could engage in a number of other forms of attack, but it certainly includes all of those biotech attacks). So to defend against such nanotech attacks, one has to be able to defend against biotech attacks. Likewise, if there is going to be the degree of monitoring that prevents those hostile nanotech attacks from happening in the first place, then there will need to be monitoring of activities in small-scale labs with small numbers of people performing small-scale manipulation of generally widely deployed synthesis mechanisms that are otherwise general-purpose.

III.9.3.2 CYBER ATTACK AND AGI ATTACK

AGI security requires widespread cybersecurity

In testimony before the U.S. Senate Subcommittee on Space, Science, and Competitiveness Committee on Commerce, Science, and Transportation, OpenAI Co-Founder and CTO Greg Brockman stated, "The Internet was built with security as an afterthought, rather than a core principle. We're still paying the cost for that today, with companies such as Target being hacked due to using insecure communication protocols. With AI, we should consider safety, security, and ethics as early as possible, and bake these into the technologies we develop."

[158] Garrick, *Quantifying and Controlling Catastrophic Risks*.

Researchers have already noted the importance of cybersecurity for AGI, but in a different context than will be discussed here (e.g., Yampolskiy, 2016; Bostrom, 2017)[159]. In order for the world to be safe against AGI, why must it already be safe against cyber attack? Even if the first AGI is confined within an impenetrable virtual box, we should expect knowledge of how to build AGIs to proliferate rapidly. Some will grow AGIs independently and release them. When encountering systems that are vulnerable to possible attacks, AGIs will often be able to discover these vulnerabilities and invent attacks. AGIs in the wild will only be limited by enforcement mechanisms that other systems use to limit interactions to agreed rules. The integrity of these enforcement mechanisms relies on the security of their underlying platforms.

III.9.3.2.1 CYBER ATTACK

Operating systems are the most vulnerable level currently

Computer systems have multiple levels of vulnerability to cyber attack, including hardware, firmware, operating systems, and users. Of these, currently the easiest pathway for attack are the operating systems, therefore our discussion starts with this most urgent vulnerability.

Problem for cybersecurity is social constraint that could be overcome via genetic takeover

With regard to cyber attack, it is widely believed that improvements to safety are a matter of technological discovery or need for new research. However, most of the techniques required to build systems that are largely secured from cyber attack, with a few exceptions, have already been known since the 1960s and 1970s, e.g., capability-based security (Miller, 2003)[160]. These techniques would actually be adequate if society could somehow reconstruct the computational world, from its beginning, on top of those techniques. The problem is that a multi-trillion dollar ecosystem is already built on the current insecurable

[159] Yampolskiy and Spellchecker, "Artificial Intelligence Safety and Cybersecurity"; Future of Life Institute, *Interactions between the AI Control Problem and the Governance Problem | Nick Bostrom*.
[160] Miller, Yee, and Shapiro, "Capability Myths Demolished."

foundations, and it is very difficult to get adoption for something that needs to rebuild the entire ecosystem from scratch. Thus, researchers have been exploring strategies to bridge from current systems to new secure ones, in what is in other contexts known as "genetic takeover," a term derived from biology (Cairns-Smith 1982)[161]. In a genetic takeover, the new system is grown within the existing system without directly competing with the existing system. The new system can coexist with the existing system, work in a world dominated by the existing system, and be competitive in that world. Once the new system comes to be widespread enough, one can start to shift over to the new system, and the previous system eventually becomes obsolete.

A real-world analogy can be drawn with how society has adapted to earthquake risk. Faced with an installed base of existing, unsafe building infrastructure, instead of requiring an immediate demolition and reconstruction, building codes are written to require earthquake reinforcement to be done on a gradual basis as other renovations take place. Over time, the installed base becomes much safer.

Genetic takeover was possible in the past and there are some hopeful examples today

The computer industry has had genetic takeovers; for example, the move from mainframes to personal computers. The entire ecosystem of mainframe software rested on a few mainframe platforms, which thereby seemed to be permanently entrenched. The new personal computing ecosystem initial grew alongside, complementing rather than competing with the old one at first, but eventually displacing it. So the attempt to replace today's existing, entrenched software ecosystem is not hopeless; but it is very difficult. Currently, there are several promising efforts to grow securable infrastructure smoothly within the entrenched infrastructure, such as Capsicum (Watson 2012a)[162], seL4

[161] Cairns-Smith, *Genetic Takeover*.
[162] Kennaway, "A Taste of Capsicum."

rehosting Linux (Nordholz 2011)[163], Secure EcmaScript (Miller 2013)[164], Sandstorm on Cap'n Proto (Filardo 2016)[165], Monte (Simpson 2017)[166], and CHERI (Watson 2012b)[167]. However, funding remains low in comparison to the urgency and importance of the challenge, and none of these projects has yet achieved widespread industry adoption.

The adoption barrier is often ignored, but critical to success and hard to overcome

This adoption barrier to making the world a safer place is ignored in most abstract discussions of AGI and nanotech attack, perhaps because one imagines that once humanity is faced with these dangers urgently, society will do what needs to be done. If there's a known technological solution for dealing with the dangers, it is natural to assume those most concerned will be able to get a majority to build, adopt, and deploy these technological solutions fast enough to avert disaster. In the case of massive cyber attacks, one would hope that government and industry would invest in rebuilding infrastructure on more securable bases. However, after seeing how weakly the world has reacted to cyber attacks that reveal massive vulnerabilities, this now appears to be unrealistic wishful thinking. The more likely reaction to the panic following a major breach will be to direct even more effort into entrenched techniques that do not and cannot work, because those are seen as recognized best practices. Techniques that actually could work will be seen as experimental and outside established best practices, best avoided in an emergency.

U.S. electric grid highly vulnerable to cyber attack today

As an example of a serious attack that could happen at the present time, the U.S. electric grid is vulnerable today to cyberattack (McLarty, Ridge,

[163] Nordholz and Seifert, "Efficient Virtualization on Hardware with Limited Virtualization Support."
[164] Miller, Cutsem, and Tulloh, "Distributed Electronic Rights in JavaScript."
[165] Filardo, "Research Report."
[166] Simpson, *Monte*.
[167] Watson et al., "CHERI."

2014)[168], with damage estimates by Lloyd's ranging up to $1 trillion (Rashid, 2015)[169]. Damage to the electric grid via cyber attack can include physical as well as software damage, and would take months (arguably, years) to repair, leaving an entire multi-state region without power. Lloyd's, as an insurance company, focused on estimating financial damages rather than fatalities. While plans have been made at the federal level in the U.S., they were prepared under a previous Administration, and it is as yet unclear whether these or similar plans will be carried out (Executive Office of the President, 2016)[170]. Of the recommendations in this paper, making capability-based upgrades to grid software is by far the most urgent.

III.9.4 IMPLEMENTING A SAFETY APPROACH

III.9.4.1 SAFETY AGAINST BIOTECHNOLOGY AND NANOTECHNOLOGY ATTACKS

Monitoring and multilateral deployment

Advanced nanotechnology can and likely will be simulated well before it can be implemented; this has been termed "exploratory engineering" (Drexler, 1988)[171]. This time gap between knowing what is buildable and carrying out the actual construction creates a possible strategy to increase safety. This ability could combine with open source approaches; as examples, today there is an active world of open source activities including impressive efforts such as the OpenWorm Project (Szigeti, 2014)[172]. A deployed defense system that could actually defend against a nanotech attack—which is much more achievable than attempting to prevent the attack to occur—would consist of a deployed fabric of systems that could detect and react based on trustworthy

[168] McLarty, Thomas F, and Ridge, "SECURING THE US ELECTRICAL GRID."
[169] Fahmida Y. Rashid, "Cyber Attack on Power Grid Could Top $1 Trillion in Damage: Report | SecurityWeek.Com."
[170] Executive Office of the President, "National Electric Grid Security and Resilience Action Plan."
[171] Drexler, "Nanotechnology and Exploratory Engineering."
[172] Szigeti et al., "OpenWorm."

mechanisms; this proposal has been termed an "active shield" (Drexler, 1986)[173].

Such a system would need to be based on both a high level of computer security and the decentralized form of monitoring described in section 1.1 above: a mutually-watching system of watchers. This complete system of decentralized defensibility, once deployed, would create and maintain an effective monopoly of force, enforcing rules of voluntarism and taking immediate physical action against malfunctioning watchers, to end noncompliance at early stages. Thus, the goal of the system would be mutually assured survival, rather than mutually assured destruction, or even the mutual deterrence enforced by the threat of using nuclear weapons.

In order for such a system to be considered trustworthy, it would need to be designed in an open source, open manner, and be on record as requiring a simultaneous multilateral release of deployment when such deployment eventually becomes possible. This ability to model systems well before actual construction is feasible creates a potentially useful time gap: a window in which it is possible to "design-ahead" (Drexler, 1986)[174].

The design-ahead window

The design-ahead window, however, is an opportunity that seems unlikely to be successfully exploited in a safety effort. Even in a best-case scenario—a system is designed that, if deployed, would monitor for offensive use and take action to prevent that use—the danger remains that one side might get to deployability before its competitors and decide to carry out a first-strike attack.

Hobbesian Trap

Given the uncertainties involved in conflict, it would appear that all parties have a lot to gain from simultaneous multilateral deployment of a mutual defense system. Unfortunately, the technical designs resulting

[173] Drexler, *Engines of Creation*.
[174] Drexler.

from sophisticated design-ahead also create a first-strike instability. This results in a "Hobbesian Trap" (Pinker, 2011)[175], such that even if no party involved desires to start a conflict, the fear that another party would do so gives an incentive to perform a first strike. We see no simple answer to this challenge.

A multilateral deployment is the scenario that, if it can be arranged, would be the most trustworthy, given that it would require the least degree of trust in the non-corruption of any one institution.

III.9.4.2 SAFETY AGAINST AGI AND CYBER ATTACKS

III.9.4.2.1 AGI

Dominant AI arrival scenarios

There is a particular safety scenario of AGI discussed specifically in the circles around Nick Bostrom (Bostrom, 2014)[176] and those around Eliezer Yudkowsky (Yudkowsky, 2015a)[177] that has become sufficiently dominant that it is worth explicitly contrasting with another perspective on the issue. The following is simplified and mostly focused on Nick Bostrom's scenario as outlined in his book Superintelligence: Paths, Dangers, And Strategies (Bostrom, 2014)[178]. Bostrom considers two scenarios for AGI ramping up: the slow takeoff scenario and the hard takeoff (fast) scenario.

III.9.4.2.2 SLOW TAKEOFF SCENARIO

Slow takeoff scenario would be safer but is less likely

Let us first consider Bostrom's slow takeoff scenario. In some sense, this is a straw man, because Bostrom believes that while a slow takeoff scenario would be safer, a hard takeoff scenario is more likely and more dangerous, thus more worthy of concern. We agree on this point, but discuss it first for reasons that will become clear. In a slow takeoff, AGI gradually emerges in a likely naturally multilateral environment.

[175] Pinker, *The Better Angels of Our Nature*.
[176] Bostrom, *Superintelligence*.
[177] Eliezer, "Rationality."
[178] Bostrom, *Superintelligence*.

Consider a scenario in which the slow takeoff is happening in a world in which secure computing technologies—techniques such as capability-based security—have become a worldwide general adoption success, so that the world has become generally much safer against cyberattacks.

Civilization as relevant superintelligence

For this scenario, one of the most relevant observations is that civilization as a whole is already a superintelligence, composed of both human and machine intelligences, serving a great variety of different interests. Granted, as machines become more intelligent, the fraction of the intelligence of civilization contributed by machine intelligence will come to be greater than the fraction contributed by human intelligence. However, in some sense this is irrelevant, because the greater intelligence is the intelligence of civilization as a whole, so we can consider that to be the relevant superintelligence. While corporations, industries, and even nation states do not meet some of the criteria that are sometimes assumed for the idealized portrayal of superintelligence (e.g., they are limited by human speed on some non-parallelizable tasks, Yudkowsky, 2016a)[179], the set of criteria they do fulfill is sufficient to merit describing them in that way with regard to possible risks.

Just as the intelligence of humans is often judged by their ability to achieve certain goals set by an intelligence test, one could measure society's intelligence by its ability to achieve the goals set by individuals using resources provided for this purpose. Miller and Drexler suggest this thought experiment: "One can imagine putting a person or an ecosystem in a box and then presenting problems and contingent rewards through a window in the box. A box full of algae and fish will 'solve' a certain narrow set of problems (such as converting light into chemical energy) and will typically pay little attention to the reward. A box containing an intelligent person will solve a different, broader range of problems. A box containing, say, an industrial civilization (with access to algae, fish and Bell Labs) will solve a vastly greater range of problems.

[179] Yudkowsky, "Corporations vs. Superintelligences."

This ability to solve externally posed problems can be taken as a measure of that ecosystem's 'intelligence' (Miller, Drexler, 1988)[180].

Thus, in any slow takeoff scenario, in which AGI is gradually emerging, the intelligence of civilization is the superintelligence that is relevant.

Civilization as networks of entities making requests of each other

Civilization as a whole is largely composed of networks of entities making requests of other entities (Miller, Tulloh, 2016)[181]. Some of those entities are humans, some are software, and in this scenario, some of those software entities are machine intelligences. The making of requests consists primarily of the mutually voluntary interaction of the party making the request and another party responding to the request. The response to the request might not be to serve the best interests of the request-making entity. However, human institutions, having evolved over many thousands of years, tend to shape interactions to be mutually voluntary and in the interests of both parties.

This definition resembles Minsky's societal definition of intelligence in which adaptive intelligence arises from a system being conflicted, rather than perfectly aligned. For humans, Minsky defends the multiple self-view in which "a part of me wants this, a part of me wants that" (Minsky,1985)[182]; because humans that are guided only by hunger will soon die, it is only the interaction of hunger and other desires (pain avoidance, etc.) that enables the organism to survive. Civilization as a whole is the most complex known system of adaptive intelligence with conflicted parts, thus the relevant superintelligence.

[180] Miller and Drexler, "Comparative Ecology: A Computational Perspective."
[181] Miller and Tulloh, "Decision Alignment (Extended Abstract)."
[182] Minsky, *The Society of Mind*.

Civilization encourages voluntary interactions

Civilization emerges from voluntary and involuntary interactions between individuals, with the balance continuing to shift towards the voluntary (Pinker 2011)[183]. Voluntary interactions happen when all participants expect to benefit, or they would not participate. A Pareto preferred change makes at least someone better off and no one worse off is (Freudenberg et al.,1991)[184]. Voluntary interactions tend, imperfectly, to move the world in Pareto preferred directions—to benefit their participants without involuntarily harming non-participants. Thousands of years of evolution of norms, laws, and institutional frameworks enable humanity to arrange ever more complex patterns of cooperation. Civilization is thus, imperfectly, largely shaped by human preferences already. It is not that civilization has a utility function, but it has a tropism. Civilizations tends, imperfectly, to grow in Pareto preferred directions. Civilization is an entrenched working system that is already superintelligent and already serves human interests.

Civilization as relevant superintelligence that serves human interests

Imagining that a new, better system can be designed to take over the world and displace this entrenched system of civilization is rather unrealistic. Instead, the goal should be to amplify the existing process of civilization and to defend it, to increase the likelihood that it is not displaced. If the current system is entirely displaced, it seems unlikely that a new system with a more beneficial utility function would actually be implemented. Human effort would be better invested in working to prevent any such unitary revolutions, because it appears unlikely that their result will serve the interests of massive numbers of people.

III.9.4.2.3 HARD TAKEOFF SCENARIO

Hard takeoff scenario involves sudden unitary takeover

Bostrom's main concern regards the prospect of a hard (i.e., sudden) takeoff, in which one particular AGI instance reaches AGI first, performs

[183] Pinker, *The Better Angels of Our Nature*.
[184] Freudenberg and Tirole, "Game Theory."

a strategic takeover, and pursues its utility function. According to Bostrom, the most important strategy that humanity can use to make AI safe in that scenario, apart from setting up the initial conditions correctly, is to shape the AI's utility function so that it serves human interests, by selecting the right top-level goal. Bostrom states that "our entire future may hinge on how we solve these problems" (Bostrom, 2003)[185].

Prevent hard takeoff using the technological knowledge that would make it possible

In the case of Bostrom's hard takeoff scenario, the AGI would displace human civilization as the overall framework of relevance for intelligence and come to dominate the world in a sudden manner. We argue that to the extent that this is the concern, but it is believed that humanity will have the ability to constrain what the AGI does (e.g., by giving it the correct top-level goal), then any abilities that humans have to constrain such an AGI should instead focus on setting up an alternative, decentralized distribution of AGIs with a system of checks and balances, rather than trying to constrain one AGI to act in human interests.

Make superintelligence part of fabric of civilization

If humans are in a position to design what the initial breakout technology is able to do, then they should also be in a position to prevent it from performing a unitary strategic takeover. Instead, our efforts can focus on directing the technological ability that the breakthrough represents to itself become widely deployed as non-coercive entities in the world. These non-coercive entities can then take part as interactive agents in the fabric of civilization, deployed by different parties simultaneously to serve many different ends (Miller, 2016) [186] . This proposal has some similarity to Drexler's technical proposal to distill superintelligent machine intelligence to apply only to specific problem domains, while avoiding the creation of one agent that has general intelligence, at least until a solution to AI safety is reached

[185] Bostrom, "Ethical Issues in Advanced Artificial Intelligence."
[186] Miller and Tulloh, "Decision Alignment (Extended Abstract)."

(Drexler, 2015)[187]. This would enable the use of many targeted, general-but-restricted AIs without requiring or entailing a unified AGI (Drexler, 2017)[188].

The safety of civilization rests on its lack of a utility function, i.e., it is a negotiated compromise using an institutional framework that accommodates a great diversity of different ends. Thus, the safety relies on the fact that the simultaneous deployment of many instantiations of such a superintelligence would occur with the many instantiations serving many different ends, and no one entity being in a position to dominate. Additionally, most of those goals should be best served by cooperating with other entities, in extensions of the cooperative framework of civilization, just as most human goals are today. This game-theory-style approach has been described more generally: "The examples of memes controlling memes and of institutions controlling institutions also suggest that AI systems can control AI systems" (Drexler, 1986)[189].

Civilization is already tested against AGIs

As mentioned earlier, civilization has already demonstrated its accommodation of superintelligences, in that large institutions themselves are already superintelligences with diverse interests that are interacting in a mostly mutually voluntary fashion. Thus, the stability of civilization has not only been tested by humans, it has also been tested by multiple interacting superintelligences, and has survived largely successfully.

Bostrom places hard moral philosophy between humans and safety

A difficulty with the approach pursued by Bostrom, Yudkowsky, and others (Armstrong, 2014) [190] is that in attempting to construct a powerful entity that acts in human interests, it is necessary to ask some deep philosophical questions about what is it that humans want or

[187] Drexler, *MDL Intelligence Distillation*.
[188] Drexler, "Reframing AI Prospects (Index Page)."
[189] Drexler, *Engines of Creation*.
[190] Armstrong Stuart, "Smarter Than Us."

should want and assumes that this question can be answered satisfyingly by the designers (Duettmann, 2014)[191]. For instance, Bostrom notes the danger of the Paperclip Maximizer Scenario, in which humans want to give the AGI an apparently peaceful goal such as maximizing paperclips, and the AGI executes the literal command and maximizes paperclips by converting most of the matter in the solar system (humans included) into paperclips (Bostrom, 2003)[192]. While the types of concerns expressed in this thought experiment are valid, these are deep philosophical questions about what humans really want or even, as Yudkowsky states, what humans would really want "if we knew more, thought faster, were more the people we wished we were, and had grown up farther together" (Yudkowsky, 2004)[193].

Yudkowsky views the most important issue regarding AGI as "constructing superintelligences that want outcomes that are high-value, normative, beneficial for intelligent life over the long run; outcomes that are, for lack of a better short phrase, 'good.'" (Yudkowsky, 2015b)[194]. Even Yudkowsky's less ambitious suggestion to construct a "Task AI," that is less sovereign than a full AGI, still relies on constructing partial normative theories. Yudkowsky calls this suggestion "insanely difficult" (Yudkowsky, 2016b)[195]. We agree. Bostrom refers to these as value-loading problems and acknowledges that AI safety must be "philosophy with a deadline" because focusing on human philosophical exploration into areas such as metaphysics doesn't contribute to solving the value-loading problems (Bostrom, 2014)[196]. However, even contemplating the extremely complicated value-loading problems, and attempting to construct the perfect goal, might well result in a completely different outcome, because the technological breakthrough will occur before philosophers have arrived at any satisfying answers to these questions. It is likely that human designers simply "do not possess the full wisdom needed to implement and grow

[191] Duettman, "The Reflective Equilibrium as Ethical Standard for AI."
[192] Bostrom, "Ethical Issues in Advanced Artificial Intelligence."
[193] Yudkowsky, "Coherent Extrapolated Volition."
[194] Yudkowsky, "The Value Loading Problem."
[195] Yudkowsky, "Task-Directed AGI."
[196] Bostrom, *Superintelligence*.

a flawlessly benevolent intelligence" (Steunebrink et al., 2015)[197], not least because the AI research community lacks the diversity required to represent a wide enough range of different interests well (Li, 2016)[198].

Rather than positioning the answers to philosophical questions that have caused disagreement for thousands of years between humanity and safety, it seems advisable to construct potential solutions which avoid moral questions that are this unanswerable.

Avoiding Benevolent Dictator scenario

A unitary takeover, whether fast or slow, is a "Benevolent Dictator" scenario at best. For much of human history, the central question of political philosophy was "Who should rule?". Political philosophy finally advanced once society realized that this was the wrong question, and to question instead whether there must be a unitary ruler (Popper 1945)[199]. Although Yudkowsky and Bostrom seek to construct the perfect dictator rather than to find one, this quest does recapitulate many of the problems of this old framing.

Ideal safety strategies would work despite uncertainty in timeframe

In current discussions of AGI safety, attempts are often made to estimate a median, average, or otherwise most-expected timeframe for the arrival of the technology. However, timeframe estimates vary by at least one order of magnitude and sometimes more, from relatively near-term (Kurzweil, 2012)[200] to very long-term (Ng, 2015)[201]. Tools such as prediction markets (Hanson, 2003)[202] and reputation-based prediction sites such as Metaculus (Aguirre, 2017)[203] may be of some help in clarifying timeframes, but currently uncertainty remains high. In this

[197] Steunebrink, Thórisson, and Schmidhuber, "Growing Recursive Self-Improvers."
[198] Patel, "Computer Vision Leader Fei-Fei Li on Why AI Needs Diversity."
[199] Popper, Ryan, and Gombrich, *The Open Society and Its Enemies*.
[200] Kurzweil and Lane, *How to Create a Mind*.
[201] Ng, "Andrew Ng."
[202] Hanson, "Combinatorial Information Market Design."
[203] Aguirre, "Metaculus."

situation, attempting to make a useful estimate of expected timing is overly optimistic; the error bars are too large. It appears advisable to develop AGI safety strategies that are robust against both early-arrival scenarios and late-arrival scenarios. A similar point has been made about the timeframe of risks from advanced nanotechnology (Drexler, 1986)[204].

III.9.4.2.4 IMPLEMENTING SECURE COMPUTING

The world is not yet hostile enough to incentivize secure computing systems today

The AGI safety and cyber attack safety strategies above require secure computing infrastructure. The adoption of secure computing is being delayed because the overall software ecosystem is not currently "hostile enough," i.e., companies and institutions can be too successful when they build systems that are very high-quality on many dimensions but are implemented in architectures that are insecurable.

Small projects can now free-ride on larger projects' being more attractive targets

In today's world in which primarily large-scale, entrenched software projects get attacked, most damage to early-stage software projects is due to dangers other than security. Therefore, for most early projects, investing in costly security is less important than investing in other areas, e.g., assembling the product and receiving feedback from user experience. Additionally, when hiring employees, a small company considers the additional value of the person to the project, so with regard to security, companies generally minimize the education burden that their team has to take on by following what are widely viewed as current best practices, rather than more unusual (and more secure) techniques. Consider the maxim that "to escape from a bear, one doesn't have to outrun the bear, but merely the other guys"; if a small project engages in the same allegedly best practices as bigger projects, it can escape attack because other projects are bigger targets. By the time the small project becomes large, it would then be a serious target, but by that point it has enough capital to manage the security problem

[204] Drexler, *Engines of Creation*.

without truly fixing it. Currently, all large corporations are managing their pervasive insecurities rather than fixing them.

Current system is only sustainable because attacks are not very sophisticated yet

This situation is only survivable because nation-states are developing the most sophisticated attacks but not yet deploying them seriously. Additionally, the attacks that nation-states are developing are probably much less sophisticated than the attacks that the most advanced organizations could be engaging in by making better use of bleeding-edge early technologies combined with static analysis technologies. For instance, the strategies that are known from the Snowden revelations include gathering Zero-Day Attacks, i.e., entities wanting to take over others' computers accumulate Zero-Day Attacks, to prepare for a future day when that entity will use them against those target computers owned by others (Wikileaks, 2013)[205]. However, rather than gathering known Zero-Day Attacks, one can imagine software that is able to analyze the software being attacked and find entirely new, previously unknown Zero-Day Attacks. Having the best state-of-the-art software for discovering vulnerabilities built into the deployed attacking system would enable the system to discover vulnerabilities and exploit them while it is in active contact with the target, rather than just launching built-in attacks against previously known vulnerabilities. This level of attack software is one that the currently entrenched architectures are not going to survive, and it is likely to precede AGI.

The launch of a sophisticated attack would make the world hostile enough to end fragile systems, but would also severely disrupt it

On the positive side, at the point that this higher level of attack gets deployed, the world's software ecosystem will become hostile enough that the relative safety through obscurity of smaller, earlier projects will end because now insecurable systems of all sizes will be punished early on. The downside of this situation is that it comes with the danger of widespread destruction of the existing software infrastructure. If a certain threshold of the world's installed software base is destroyed, it

[205] "WikiLeaks - The Hackingteam Archives."

could be difficult to transition to a safer situation without having gone through a serious downturn in overall functionality of the world's computation systems, not to mention the world economy.

seL4 microkernel as example of code safe against attacks

Combining various state-of-the-art research has led to some impressive results at finding vulnerabilities in software targets. One example is research on combining Machine-Learning sophisticated AI with sophisticated static analysis of programs to find vulnerabilities (Brooks 2017)[206]. This level of sophistication is not accidentally going to be part of an attack system. However, if it is built in as part of an experiment run on a platform that is believed to be secure but that is not air-gapped (i.e., is not isolated from the internet), such an experiment would be very good at detecting flaws. The seL4 microkernel is our best example of an operating system kernel that seems to be secure, due to its formal proof of end-to-end security and its track record of having withstood a Red Team Attack (a full-scope, multilayered attack simulation) which no other software has withstood (Fisher, 2014) [207]. One hopeful development is increased funding of seL4 by the U.S. Department of Defense. Nevertheless, its security rests on some counterfactual assumptions, such as that the formal model of the underlying hardware is accurate.

III.9.4.2.5 A MODEL OF DECISION ALIGNMENT

A model of software object security can be combined with a model of human-to-human security

Many complex systems can be described as networks of entities making requests of other entities. In economics, there are principal-agent relationships, in which a principal sends a request to an agent. The principal uses various techniques to try to align the decision of the agents with the interests of the principal to increase the likelihood that the request is fulfilled.

[206] Brooks, "Survey of Automated Vulnerability Detection and Exploit Generation Techniques in Cyber Reasoning Systems."
[207] Fisher, "Using Formal Methods to Enable More Secure Vehicles."

Economics, for instance, studies principal-agent relationships among humans and examines both hazards, such as divergent interests and asymmetric information, and techniques for addressing those hazards. Software engineers deal with principal-agent relationships among computational objects and examine hazards and techniques such as object design patterns. Human Computer Interaction (HCI) deals with human-object interactions and examines hazards and techniques, such as user confusion or request expressiveness.

The techniques principals use to align agent decisions with the principal's intent can be divided into six categories: Select agent (admission control), Inspect internals (static analysis), Allow actions (least authority), Explain request (abstraction design), Reward cooperation (incentives), Monitor effects (reputation feedback); see Table 5 (Miller, Tulloh 2017)[208]. A unified view that looks at the different techniques in relation to each other can provide important insights. Reasoning across both rows and columns of Table 5, and combining techniques, allows reaping the payoff of having different techniques reinforce each other. Thus, Table 5 is not simply about reasoning by analogy, but instead reasoning about a single integrated network spanning multiple systems.

Table 5–The Elements of Decision Alignment (Miller, Tulloh, 2017)[209]

	Human to Human	Human to/from Object	Object to Object
Select agent	Trademark Chain of Custody	App stores White and black lists	Trusted developer same origin
Inspect internals	Accounting controls	Trusted path URL bar	Types, Verification Open source eyeballs
Allow actions	Law, Contracts	App permissions Powerbox	Security Protection patterns

[208] Miller and Tulloh, "The Elements of Decision Alignment."
[209] Miller and Tulloh.

Explain request	Language	User Interface	Abstraction
Reward cooperation	Economics Incentive Alignment	Objective functions	Machine learning Agorics
Monitor effects	Reviews, Complaints Word of mouth	Bug reports	Contracts, Testing Backprop

For example, computer security (Allow actions) taken alone misses some differences among agent actions that cause harm to the principal, such as when the agent benefits from misbehavior (Reward cooperation). Instead, principal-agent arrangements can be designed such that each technique fills in for weaknesses in the others, creating greatly increased structural strength built out of individually breakable parts.

While perfect security might ultimately be unattainable (Yampolskiy, 2016) [210], this approach has the possibility of delivering adequate security and is a great deal more secure than any of the insecurable security systems that are now widely in use. Moreover, it is not only applicable to today's computer security but also is independent of the intelligence of the agent and therefore can be applied to AGI safety as well.

III.9.4.2.6 THE PROBLEM OF SUPPLY CHAIN RISK

Formal security proofs rest on assumption that hardware is safe but it might not be

After insecure operating systems, supply chain risk is the hardest problem in attempting to ensure secure computation. The proof that a given hardware chip design is secure only helps if hardware which the software is run on is actually the hardware that was designed. This assumption sounds trivial but it may be false, because it is possible that

[210] Yampolskiy and Spellchecker, "Artificial Intelligence Safety and Cybersecurity."

the hardware includes a manufactured-in trap door. Based on the revelations about the U.S. National Security Agency (NSA) serving national security letters to software companies forcing them to disclose user information, it is possible, indeed likely, that the NSA has already served national security letters to hardware companies including Intel and AMD requiring them to install trap doors into their hardware, which the NSA can later choose to trigger (Gustin, 2014)[211]. Fearing billions of USD in profit losses after the revelations in 2014, IBM's President Weber was quick to point out in an open letter to clients that the hardware giant would not comply with such letters (Weber, 2014)[212]. However, the severe penalties associated with disobedience or disclosure should cause us to be skeptical. None of today's proofs of software security can defend against such trap doors.

Open source processor design as possibility to overcome trustworthiness issues of hardware

In the near term, one can imagine a technology example that can be secure against those risks: a good open source processor design for which there is a proof of security comparable to the proof of security of the seL4 software. There are many open source processor designs that are sufficiently high performance that, when run on a field-programmable gate array (FPGA), can run fast enough to be practical for many applications. By combining these well-designed processors with a layout algorithm that randomizes layout decisions, the processor could be randomly laid out for each individual hardware instance. Given this randomized layout, there is no feasible corruption of the FPGA hardware that can escape notice under electron microscopes and that would also be able to successfully corrupt most instances of the processor.

Even if trustworthy processor is theoretically possible, it is likely too expensive

However, even if it was possible to build a secure processor, it would be hopelessly unadoptable. The current norm in secure software holds that if a software security mechanism costs a factor of 3% more than

[211] Gustin, "IBM Says It Hasn't Given the NSA Any Client Data."
[212] Weber, "A Letter to Our Clients About Government Access to Data."

insecurable mechanisms, widespread adoption becomes very unlikely. The trustworthy hardware in the form of a secure FPGA described above would result in a factor of at least an order of magnitude in performance cost over producing chips the standard way, which renders its adoption unrealistic.

III.9.4.2.7 THE BLOCKCHAIN ECOSYSTEM AN APPROACH FOR SAFETY

Ethereum and blockchain evolving in hostile ecosystem

A counterexample to the difficulties expressed above is Ethereum's current approach. Both Bitcoin and Ethereum are evolving in an ecosystem that is already under the very hostile attack pressures described earlier in this paper. When insecurity leads to losses, the players have no other recourse to compensate. Systems that are not bulletproof will be killed early and visibly, and therefore these ecosystems remain populated only by bulletproof systems. The bulletproof security of these systems is an essential part of their value proposition.

Ethereum as virtual machine that is trustworthy

Regarding the problem of trustable hardware mentioned earlier, if Ethereum is a virtual machine, it is a factor of at least ten thousand times costlier in performance than the FPGA approach mentioned earlier, that was already too expensive to be adopted. Ethereum is trustworthy in the same sense that Bitcoin is trustworthy; Bitcoin is a payment system and Etherium is a general purpose virtual machine (CPU, memory, limited IO). Both are synthesized by cryptographic protocols and massive redundancy among their players, based on a blockchain—an agreed order of messages. In order for either to take action in an untrustworthy fashion, a supermajority of participants would have to perform actions that were visibly illegitimate.

The Ethereum and Bitcoin systems per se are holding up very well. The publicized attacks on these systems do not reveal weaknesses in their foundation, but rather in the participants—at two different abstraction levels. Bitcoin exchanges were hacked, with losses of several hundred million dollars, due to insecurity of the platforms used by the exchanges,

not due to any flaw in the Bitcoin protocols. Ethereum, as a virtual machine, runs programs written by its users, such as the DAO (Decentralized Autonomous Organization) smart contract.

Example of DAO as bad software deployed on top of trustworthy Ethereum machine

The DAO was the first significant piece of software deployed by a commercial participant on Ethereum, and it was not bulletproof. While the software withstood initial code review, it should have been subject to (at least) more code review, or preferably a formal proof of correctness. As explained above, machine checked formal proofs of correctness can be and have been successfully performed on much larger and more complicated pieces of software such as seL4.

In the case of the DAO, once this insecure piece of software was deployed, hackers exploited a known bug, started diverting money, and successfully removed US$60 million worth of Ether. This provoked the Ethereum ecosystem to engage in a "hard fork," a deliberate change of software that created a new version of the Ethereum system in which the Ether was not stolen (Buterin, 2016)[213]. Resetting the system in this way was a serious compromise of the founding principles of these cryptographic smart contract systems, which is that they are permissionless, i.e., that "code is law." It established a terrible precedent that future actions within the systems may be overridden by retroactive fiat.

Blockchain ecosystem as hope for building something that is secure against cyber attacks

Despite this early misstep, we are optimistic that the universe of cryptographic smart contracts can be the beginning of an ecosystem in which projects can grow up under extraordinarily hostile conditions. Such projects are evolving with a degree of adversarial testing that can create the seeds for a system that can survive a magnitude of cyberattack that would destroy conventional software. If this type of secure system grows enough before the world is subject to such

[213] Buterin, "Hard Fork Completed."

cyberattacks, then a successful genetic takeover scenario might be achieved.

Safety of the proposed system still relies on counterfactual assumption

Bostrom states that even if one has a truly secure system, an AGI is likely to be able to break out of it, because "even a 'fettered superintelligence,' that was running on secure hardware on an isolated computer that can communicate only via text interface, might be able to break out of its confinement by persuading its handlers to release it" (Bostrom, 2003)[214]. While we cannot rule out this possibility, this degree of human gullibility does not seem plausible to us. Perhaps some pre-AGI experiments could help quantify this issue.

The proposed system's formal safety is independent of attacker intelligence, so would remain safe not only under cyberattack but also under AGI

While these other threat vectors are problematic, it is important to emphasize that, to the degree to which these systems are formally secure, that security is independent of the intelligence of the attacker. Thus, if humanity succeeds at building systems before AGI that are actually secure, which can in principle be done, then those systems should remain formally secure under AGI. The formal security of systems such as seL4, and the adversarial testing carried out on smart contracts, is likely to create an ecosystem of software systems which are secure against AGIs, because the threshold that needs to be crossed to guarantee security can be crossed well before AGI is reached. (In fact, this level could have been crossed before reaching the level of current machine intelligence.) There is no prerequisite of one on the other.

III.9.4.2.8 SUCCESSFUL CONSTITUTIONS AS ROLE MODELS FOR A SAFE SYSTEM

As a few examples of organizational arrangements that have had some long-term success at managing competing superintelligences, we can point to the Swiss Federal Constitution, the U.S. Constitution, and the (partly unwritten, but real) U.K. Constitution. Here we take the U.S.

[214] Bostrom, "Ethical Issues in Advanced Artificial Intelligence."

Constitution as an illustration, primarily due the authors' relative familiarity with it; later work should address a wider variety of successful arrangements.

Founding fathers were trying to create a Constitution that depessimizes

The originators of the U.S. Constitution, termed the Founding Fathers, were faced with a Bostrom-like nightmare of having to design a single institution that was going to be superintelligent, and that was composed of systems of people that individually want to take many actions that society would collectively not want any of them to do. However, these originators felt that they had no choice but to design this institution and attempt to create an architecture that was inherently constructed to maintain its integrity, not at being ideal but at avoiding being very seriously flawed. This strategy is generally known as depessimizing, rather than optimizing as advocated by most in the AI safety field. The worst-case scenarios of our future are extremely negative and numerous, so by simply avoiding the worst cases humanity would be doing extraordinarily well. Attempting to do better among the non-worst-case scenarios can be viewed as a very minor issue compared to safety against the worst cases. Gentzel advocates this depessimizing approach for AGI work: "the most sensible medium-term goal for human society is to guide the advance of technology in a rational way that has reasonable odds of getting past the current phase of development without causing global annihilation or other horrible catastrophes" (Goertzel, 2015)[215].

Success of Constitution as lending support to feasibility of building safe AGI

In the case of AGI, instead of attempting to build an optimal system, humanity should focus on not building a system that turns into a worst-case scenario. In the case of the U.S. Constitution, instead of attempting to design the Constitutions as an optimized utility-function that would serve everyone's interests, the originators' main objective was to avoid having it becoming a tyranny. It is extraordinary that this Constitution maintained most of its integrity of mechanism, as well as integrity of

[215] Goertzel, "Superintelligence: Fears, Promises and Potentials."

purpose, for its first 150 years and maintains much of this even today. It shows that this type of effort can succeed and is worth taking on.

AI safety is harder than Constitution because less familiar knowledge to build predictions on

AI safety is harder in the sense that when formulating the Constitution, the Founding Fathers could rely on their knowledge of human nature and the history of politics and human institutions. With regard to AI safety, there is less solid ground; however, the basic mechanism can be based on the above, in a more focused and less decentralized way. Just as future AGIs will dwarf current superintelligences with regard to intelligence, so are current superintelligences dwarfing the expectations of what the Founding Fathers imagined when designing the Constitution over two centuries ago. Nevertheless, the Constitution was only designed as a starting point, on which later, more intelligent agents could build, and it is still surprisingly relevant one industrial revolution later. Rather than inventing a safety approach from first principles, a useful approach could make use of the immense body of historic and cultural knowledge that can be relied on to ensure a more organic AGI world, similar to the one envisioned by Kurzweil: "Ultimately, the most important approach we can take to keep AI safe is to work on our human governance and social institutions. We are already a human--machine civilization. The best way to avoid destructive conflict in the future is to continue the advance of our social ideals, which has already greatly reduced violence" (Kurzweil, 2014)[216]

Elon Musk also appears to favor a decentralized approach: "The important thing is that if we do get some sort of runaway algorithm, then the human AI collective can stop the runaway algorithm. But if there's a large, centralized AI that decides, then there's no stopping it" (Dowd, 2017)[217]. Musk, Thiel, Reid Hoffman, Sam Altman, and others have founded and pledged a total of $1 billion to OpenAI, a foundation

[216] Kurzweil, "Ray Kurzweil: Don't Fear Artificial Intelligence."
[217] Dowd, "Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse."

with the purpose of developing and distributing AI widely as a safety strategy (Risley, 2015; OpenAI, 2017)[218].

Co-Chair Sam Altman explains, "Just like humans protect against Dr. Evil by the fact that most humans are good, and the collective force of humanity can contain the bad elements, we think it's far more likely that many, many AIs, will work to stop the occasional bad actors than the idea that there is a single AI a billion times more powerful than anything else" (Levy, 2015)[219]. The Open AI approach has attracted a $30 million grant from the Open Philanthropy Project (Open Philanthropy Project, 2017)[220]. For guidance on the AI transition, Altman has looked to James Madison's notes on the Constitutional Convention (Friend, 2016)[221].

Multilateralism and gridlock as important part of the system

Previously we mentioned civilization being very widely multilateral. In that sense, the evolved institutions of civilization are the result of this decentralized, ongoing negotiation among institutional frameworks having a very wide diversity of interests. Additionally, the Madison form of government was a perpetually explicitly renegotiated framework among these small number of divergent interests that were purposely put its opposition with each other, including division of power, checks and balances, and significant decentralization. Building the system to be in conflict with itself is a much more realistic strategy than to pursue building a unitary system that wants the right goals. While the checks and balances designed into such a system lead to a decrease in speed and efficiency, this is a positive tradeoff in exchange for a reduction in much more serious risks.

In addition to the UK, US, and Swiss constitutions, those attempting to design governance systems for AGI safety may find inspiration from (1) John Locke on institutional checks and balances, (2) John Adams on

[218] "About OpenAI"; Risley, "Elon Musk, Peter Thiel, Reid Hoffman and Others Commit $1B to Stop AI from Taking over the World."
[219] Levy, "How Elon Musk and Y Combinator Plan to Stop Computers From Taking Over."
[220] "OpenAI — General Support."
[221] Friend, "Sam Altman's Manifest Destiny."

federal-state balance, based on his study of the United Provinces of the Netherlands, Switzerland, the Holy Roman Empire, and the Peloponnesian League confederation in ancient Greece, and (3) the later cases of the Canadian, Australian, postwar German, and postwar Japanese constitutions (Bennett, 2017)[222].

III.9.5 SECURING HUMAN INTEREST IN AN AGI WORLD

How to achieve human interests once AGI is reached

Regardless of the exact timeframe, if our civilization and technology continue to progress, AGI will ultimately be reached. As Sam Harris points out, the only reason why AGI would not be reached eventually will be that an even worse event occurs, which destroys technology or civilization before it reaches that state (Harris, 2016)[223]. To ensure that civilization still serves human interests when AGI is reached, we argue that humans need a claim to capital to be able to participate in exchange, and that a promising way for humans to obtain this claim to capital is through a one-time distribution of unclaimed resources, referred to as "Inheritance Day" (Drexler, 1986)[224].

III.9.5.1 PROVIDING HUMANS WITH CAPITAL BARGAINING POWER

This section argues that to ensure civilization still serves human interests once AGI is reached, humans need capital to participate in exchange for two main reasons.

Civilization serves human interests as long as humans contribute either skill or capital in exchange

Earlier in the paper we argued that civilization tends to serve human interests, albeit imperfectly. However, the argument that the system serves human preferences depends on humans having something to offer in exchange, either proceeds from their capital or their skills. Historically, much of what humans had to offer in exchange was based

[222] Bennett, "Private Communications to Christine Peterson."
[223] Harris, "Surviving the Cosmos."
[224] Drexler, *Engines of Creation*.

on human skills. These skills were of two kinds: human mechanical skills (the ability of humans to perform actions physically) and mental skills. Currently, while it is still possible to earn income using their skills, many humans don't have capital.

Once AGI is reached human skills become irrelevant but capital has high returns

Machines have already displaced humans from being able to earn much income via direct contribution of human mechanical skill and will continue to do so (McKinsey, 2015)[225]. Human dexterity has a lot to contribute currently, but that is largely due to its coupling with human mental ability. Human mental ability in the abstract contributes a great deal today, but humanity should anticipate the day when a machine intelligence achieves general intelligence. Once AGI is widely deployed, it is anticipated that human skills will be outcompeted and have little or no economic value.

An economy that is so productive that human skill is irrelevant is also an economy that grows extraordinarily quickly, similar to Robin Hanson's description of an economy in which the GDP is doubling in weeks (Hanson, 2014) [226] . Any economy growing at this rate offers extraordinary returns to capital. Capital is defined here as the ownership of resources, where resources are both physical objects as well as ownership of created abstract rights (i.e., corporate stock) that are part of the fabric of the civilization. This capital itself can become investments which the economy would reward extraordinarily.

If humanity enters into the transition to AGI with insufficient preparation, much of humanity will have no capital and their skills would be irrelevant. One strategy to ensure that the dynamic of civilization still contributes to human well-being once human skills are irrelevant is to arrange that most human beings have some capital claim that they can continue to trade on, get capital returns on, and live well.

[225] Chui, Manyika, and Miremadi, "Four Fundamentals of Workplace Automation | McKinsey & Company."
[226] Hanson, "When the Economy Transcends Humanity."

III.9.5.2 INHERITANCE DAY AS STRATEGY TO PROVIDE CAPITAL CLAIM

Inheritance Day as strategy to provide capital claim

Since humans need a claim to capital to ensure that civilization still serves their interests, this section deals with a potentially useful way of granting capital claims to humans. We argue that one possible strategy to provide humans with a capital claim is via a strategy called "Inheritance Day."

One-time distribution can provide humans with a capital claim

Capital claims can be assigned to individuals either by redistributing existing capital claims or distributing new capital claims. Currently, redistribution is the most common strategy to assigning capital claims to individuals in need. However, redistribution leads to political opposition, because it involves giving new beneficiaries a claim to capital by taking it away from the previous owners. Continual redistribution also appears to reward high reproductive rates, leading to additional opposition. To avoid this political opposition, rather than redistributing capital that has already been claimed, society could distribute capital that has not been claimed yet. While the great majority of the Earth's land and much of its undersea area are already claimed, there is an entire universe of (according to present knowledge) unclaimed, unowned resources in space.

Inheritance Day is a promising proposal for distribution

Generally, in the past, unoccupied land has become owned via homesteading, in which the prospective owner occupies the land physically and develops it. However, in principle homesteading destroys economic value by giving rise to competition to become the entity performing the homesteading, which is a deadweight loss compared to the economic benefits of making use of the resources once they are claimed. The Inheritance Day proposal described by Drexler suggests that humanity select a day on which every human being alive that day is assigned an equal share of the as-yet-unclaimed resources of the universe. According to the Coase Theorem, given clear title to resources and multilateral ownership, ignoring transaction cost problems,

subsequent trade leads to a good utilization of those claimed resources and to a Pareto efficient outcome for all parties, regardless of the initial distribution of resources (Coase, 1960)[227].

Inheritance Day could provide individuals with the necessary capital to increase the likelihood that civilization will still serve human interests once AGI is reached and their skills are no longer economically sufficient.

III.9.5.3 IMPLEMENTING INHERITANCE DAY

Drexler quote on Inheritance Day implementation

When describing Inheritance Day, Drexler states that "this involves distributing ownership of the resources of space (genuine, permanent, transferable ownership) equally among all people—but doing so only once, then letting people provide for their progeny (or others') from their own vast share of the wealth of space. This will allow different groups to pursue different futures, and it will reward the frugal rather than the profligate. It can provide the foundation for a future of unlimited diversity for the indefinite future, if active shields are used to protect people from aggression and theft" (Drexler, 1986)[228].

Timing of release of Inheritance Day assets to individuals

One idealistic interpretation of the proposal, not recommended here, is that individuals receive full title to their entire share of newly-assigned resources with complete ability to trade immediately. The Coase Theorem seems to suggest that this would be the most economically efficient solution. However, the problem is that most individuals are not yet experienced at managing capital, much less ownership of space resources. Historically, when the wealthy plan to leave an inheritance to children who are still underage, they create a trust which gradually releases the resources to benefit the beneficiary until that beneficiary has crossed an age threshold such that the grantor is willing to entrust them with the rest of the wealth. With regard to Inheritance Day, understanding that at the moment that Inheritance Day is implemented none of the beneficiaries are as yet experienced at managing capital of

[227] Coase, "The Problem of Social Cost."
[228] Drexler, *Engines of Creation*.

this nature, it would be advisable to grant some of those resources immediately to individuals, including the ability to trade, but to hold most of the resources in trust and gradually release them over time. This would enable all individuals to continue to have a gradual stream of capital that can be invested and produce financial returns as the economy continues to grow. This would help ensure that individuals are protected from making terribly egregious, early foolish mistakes (Miller, 1980s-90s)[229].

Defining an equal share of space resources

Defining what constitutes an equal share of the resources of space is a hard problem that remains unsolved to date. What counts as an equal share relies on individuals' assessments of that share and these subjective values can differ greatly (Harms, 1989)[230]. However, a promising protocol for division is the "I cut, you choose" principle for envy-free distribution of resources with agents which have different preferences. One agent divides the resources, the other partner chooses first, and the divider receives the remaining share. While there are some new algorithms for cake-cutting for multiple agents with multiple preferences, the distribution of all resources in space remains a complex problem (Aziz, 2016)[231]. Another role model would be the approach to the privatization of resources in Poland via national wealth management funds when transitioning from a communist economy to a market economy. The government retained some of the shares of the newly privatized enterprises, gave some to the company's employees, and distributed the rest to competing National Wealth Management Funds, so that one investment group had primary responsibility for modernizing a given enterprise. From these funds, 27 million individuals received vouchers, equivalent to American-style mutual funds (Goldman, 2016)[232].

[229] Miller, "A Series of Private Communications to Christine Peterson in the 1980s and 1990s."
[230] Harms, "NanoCon: The Cosmic Pie, Harms."
[231] Aziz and Mackenzie, "A Discrete and Bounded Envy-Free Cake Cutting Protocol for Any Number of Agents."
[232] Goldman, *Revolution and Change in Central and Eastern Europe*.

When speaking on AI risk, Jaan Tallinn references a thought experiment involving negotiations between humanity and a powerful alien fleet which doesn't care about humanity. He says: "Even if we could secure just one galaxy out of the 100 billion as consolation prize for the losers, this would translate into 50 personal star systems for every human alive today. This illustrates two things: (1) Even if we mostly screw up, things might turn up to be pretty okay in the end and (2) the worst we can do is continue our current political zero-sum games, which cost us 50 galaxies per second" (Tallinn, 2017)[233]. Both points hold for Inheritance Day in a similar way: (1) Even if some might deem the details such as the choice of date or exact distribution of space resources as arbitrary, the sheer size of space could still eventually allow every person to be well-off, and (2) continuing to delay action and perpetuating the current system is just as much a decision as taking action to change, and is one that is potentially costly.

Inheritance Day is orthogonal to redistribution questions

The proposal for redistribution of resources, e.g., a basic income as supported by Elon Musk, Sam Altman, and other prominent figures (Agreelist, 2017)[234] and similar to the current experiment in Finland, is a separate issue. These two approaches to attempting to ensure human financial well-being—a one-time gradual distribution of unclaimed resources, and a continual redistribution of already-owned resources—are in principle unrelated. Either could be implemented on its own, or they could be combined. Whether either or both actually become implemented are political decisions for society to make.

III.9.6 CONCLUSIONS

Biotech risks can be seen as a subset of longer-term and more challenging Stage 4 nanotech risks; both derive from systems of molecular machines. Similarly, cyber risks can be seen as a subset of later AGI risks. Computer security is identified as important across many risk domains. Defensive, decentralized, bottom-up, open source approaches are suggested for addressing a variety of risk areas.

[233] Tallinn, *AI and Value Alignment*.
[234] Agreelist, "Universal Basic Income - Do You Agree?"

Inspiration can be provided by analogies with successful defense scenarios across domains, from the immune system in biology to the U.S. Constitution in politics. Timing estimates for these anticipated powerful technologies vary widely, therefore it is advisable to attempt to find strategies that are robust across these differing time estimates. A gradual, one-time distribution of unclaimed resources could help ameliorate the concern that human labor becomes much less valuable in a world with AGI. These concepts are presented as options for consideration and possible elaboration, rather than as complete policy recommendations.

III.9.7 ACKNOWLEDGEMENTS

We thank Mark S. Miller for substantial contributions and feedback on the ideas in this paper, and James C. Bennett, Steve Burgess, Ben Goertzel, Tanya Jones, Richard Mallah, Gayle Pergamit, Glenn Reynolds, Marcia Seidler, Leif Smith, Bas Steunebrink, Roman Yampolskiy, and Eliezer Yudkowsky for helpful comments; any remaining errors are our own. We thank Foresight Institute for financial support and UCLA's B. John Garrick Institute for the Risk Sciences for organizing the First Colloquium on Existential and Catastrophic Risk which stimulated this project.

IV ADDITIONAL LITERATURE ON CATASTROPHIC AND EXISTENTIAL RISK

IV.1 VALUE OF GCR INFORMATION: COST EFFECTIVENESS-BASED APPROACH FOR GLOBAL CATASTROPHIC RISK REDUCTION

AUTHOR AND LINK

Author: Anthony Michael Barrett

Link: https://pubsonline.informs.org/doi/abs/10.1287/deca.2017.0350

ABSTRACT

In this paper, we develop and illustrate a framework for determining the potential value of global catastrophic risk (GCR) research in reducing uncertainties in the assessment of GCR risk levels and the effectiveness of risk-reduction options. The framework uses the decision analysis concept of the expected value of perfect information (EVPI) in terms of the cost effectiveness of GCR reduction. We illustrate these concepts using available information on impact risks from two types of near earth objects (asteroids or extinct comets) as well as nuclear war, and consideration of two risk reduction measures. We also discuss key challenges in extending the calculations to all GCRs and risk-reduction options, as part of an agenda for comprehensive, integrated GCR research. While real-world research would not result in perfect information, even imperfect information could have significant value in informing GCR reduction decisions. Unlike most value of information approaches, our equation for calculating value of information is based on risk reduction cost effectiveness, to avoid implicitly equating lives and dollars e.g. using a value of statistical life (VSL), which may be inappropriate given the scale of GCRs. Our equation for value of information may be useful in other domains where VSLs would not be appropriate.

IV.2 A MODEL OF PATHWAYS TO ARTIFICIAL SUPERINTELLIGENCE CATASTROPHE FOR RISK AND DECISION ANALYSIS

AUTHOR AND LINK

Author: Anthony M. Barrett & Seth D. Baum

Link: http://dx.doi.org/10.1080/0952813X.2016.1186228

ABSTRACT

An artificial superintelligence (ASI) is an artificial intelligence that is significantly more intelligent than humans in all respects. Whilst ASI does not currently exist, some scholars propose that it could be created sometime in the future, and furthermore that its creation could cause a severe global catastrophe, possibly even resulting in human extinction. Given the high stakes, it is important to analyze ASI risk and factor the risk into decisions related to ASI research and development. This paper presents a graphical model of major pathways to ASI catastrophe, focusing on ASI created via recursive self-improvement. The model uses the established risk and decision analysis modelling paradigms of fault trees and influence diagrams in order to depict combinations of events and conditions that could lead to AI catastrophe, as well as intervention options that could decrease risks. The events and conditions include select aspects of the ASI itself as well as the human process of ASI research, development and management. Model structure is derived from published literature on ASI risk. The model offers a foundation for rigorous quantitative evaluation and decision-making on the long-term risk of ASI catastrophe.

IV.3 ANALYZING AND REDUCING THE RISKS OF INADVERTENT NUCLEAR WAR BETWEEN THE UNITED STATES AND RUSSIA

AUTHOR AND LINK

Author: Anthony M. Barrett, Seth D. Baum & Kelly Hostetler

Link: https://doi.org/10.1080/08929882.2013.798984

ABSTRACT

This article develops a mathematical modeling framework using fault trees and Poisson processes for analyzing the risks of inadvertent nuclear war from U.S. or Russian misinterpretation of false alarms in early warning systems, and for assessing the potential value of options to reduce the risks of inadvertent nuclear war. The model also uses publicly available information on early warning systems, near-miss incidents, and other factors to estimate probabilities of a U.S.–Russia crisis, the rates of false alarms, and the probabilities that leaders will launch missiles in response to a false alarm. The article discusses results, uncertainties, limitations, and policy implications.

Supplemental materials are available for this article. Go to the publisher's online edition of Science & Global Security to view the free online appendix with additional tables and figures.

IV.4 UNDERSTANDING AND MITIGATING THE IMPACTS OF MASSIVE RELOCATIONS DUE TO DISASTERS

AUTHOR AND LINK

Author: Vicki Bier

Link: https://link.springer.com/article/10.1007/s41885-017-0003-4

ABSTRACT

We have grown used to thinking of displaced persons as a developing-world problem. However, Hurricane Katrina and the Japanese tsunami/nuclear disaster made clear that even in the developed world people may need to leave their homes due to natural or man-made disasters. This can occur for reasons ranging from nuclear accidents, to natural disasters (e.g., hurricanes), to terrorism (e.g., a major anthrax attack), to climate change (e.g., coastal flooding). In addition to the social consequences of forced relocation, massive relocation can have significant economic costs, including not only property damage, but also business interruption, loss of housing services, and decline of property

values. Economic consequences can be expected to be highly nonlinear in both magnitude and duration of relocation. With regard to duration, a brief evacuation may be minimally disruptive, if people are able to return to their homes within a few days. A relocation of a few months or a year would be much more disruptive per day, but eventually, costs per day would diminish or approach zero. By contrast, costs can be expected to increase monotonically but non-linearly in the number of people needing to be relocated. Costs may also vary greatly depending on the nature of the assets that are interdicted. Unfortunately, disasters in populated areas can easily result in the need to relocate a million people or more. This argues for the need for research on interventions to encourage relocation before a disaster in areas under significant threat, and to increase resilience after massive relocations.

IV.5 THE EMERGENCE OF GLOBAL SYSTEMIC RISK

AUTHOR AND LINK

Author: Miguel A. Centeno, Manish Nag, Thayer S. Patterson, Andrew Shaver, and A. Jason Windawi

Link: http://www.annualreviews.org/doi/abs/10.1146/annurev-soc-073014-112317

ABSTRACT

In this article, we discuss the increasing interdependence of societies, focusing specifically on issues of systemic instability and fragility generated by the new and unprecedented level of connectedness and complexity resulting from globalization. We define the global system as a set of tightly coupled interactions that allow for the continued flow of information, capital, goods, services, and people. Using the general concepts of globality, complexity, networks, and the nature of risk, we analyze case studies of trade, finance, infrastructure, climate change, and public health to develop empirical support for the concept of global systemic risk. We seek to identify and describe the sources and nature of such risks and methods of thinking about risks that may inform future academic research and policy-making decisions.

IV.6 SYSTEMIC RISK IN GLOBAL AGRICULTURE

AUTHOR AND LINK

Author: Conference (Princeton-Columbia Joint Conference)

Link: https://risk.princeton.edu/img/Princeton-Columbia_Agriculture_Conf_Report_2014-10-24_(v2016-09-27).pdf

ABSTRACT (EXECUTIVE SUMMARY)

The Green Revolution is estimated to have led to our collective ability to feed over one billion additional people largely through scientific advances in crop and soil science. Modern information technology, communication, and transportation have woven interdependent networks that provide greater efficiency of both production and delivery of food, which has led to a systems-driven revolution in agriculture. However, the large scale and advanced technical nature of these complex systems comes at the cost of greater fragility. The critical nature of the agricultural system as the source and sustenance of life elevates the study and remediation of this fragility to a global priority.

The emerging research fields of systemic risk and systems thinking provide insight into understanding and mitigating the current risks and challenges in our global agriculture network. This network is a system-of-systems that begins beneath the ground with our aquifers and soil. Subsequently, it extends through the crops with bidirectional effects between environment and climate, trade and finance, and human health and livelihood. Finally, with its effect on political stability, the network extends into the realm of policy and governance.

This conference and the summary of the proceedings first explore the current challenges to modern agriculture. Next, we seek to contribute to the research field by applying systems thinking in order to explain these critical challenges. Finally, we attempt to understand the implications for prescriptive analysis and governance in pursuit of the goals of greater productivity, mitigating risk, and increasing resilience

IV.7 TAKING ACTIONS TO PREPARE SOCIETY FOR CATASTROPHIC RISKS

AUTHOR AND LINK

Author: B. John Garrick and Roger L. McCarthy

Link:
https://static1.squarespace.com/static/54628adae4b0f587f5d3e03f/t/57d83e48e58c62769f265609/1473789514378

INTRODUCTION

Will human beings make it through the next century?

It's not a frivolous question. Barring a cataclysm that would make the question irrelevant, the answer is "yes." The real question is how much unnecessary suffering are we going to endure?

Catastrophic risks are not something that most people seriously think about. The human race has all too often addressed disasters only after they happen. There are always more immediate concerns, and the public is becoming increasingly inured to predictions of disaster caused by everything ranging from Y2K to pandemics that never materialize. Unfortunately, disaster fear is fueled by news media hyperbole and entertainment industry fantasies, so potential catastrophes are sensationalized to the detriment of their rational consideration. Even the three simultaneous core meltdowns at Fukushima during the Great East Japan Earthquake of 2011 produced nothing close to the China syndrome or any other apocalyptic casualty scenario popularly associated with nuclear energy.

The stark reality is the human race is at catastrophic risk, more than ever before. Besides threats that have always been with us – and, sadly, always will be – such as plagues, mass warfare, or natural disasters, we are living in ever denser and thus more fragile urban concentrations, as the 2004 and 2011 tsunamis and the 2005 Hurricane Katrina so painfully illustrated. We also have much greater interconnection, as SARS (severe acute respiratory syndrome) also painfully illustrated. Denser urban

population centers provide greatly increased leverage for these long known natural catastrophes to create mass casualties.

In addition, burgeoning technologies have the potential for creating catastrophic events. Advances in artificial intelligence, nanotechnology, biological engineering, and even particle physics are occurring at an exponentially growing rate, with implications that are breathtaking, and certainly with some unintended and unknown consequences. Fueled by Moore's law3 in computer power and by the expansion of Internet access to virtually the entire world, these and other technologies present challenges that are serious now and could become catastrophic in the not-too-distant future. Unfortunately, popular culture has too often caused this field to be occupied with holocaust fantasies of everything from plagues (too many zombie movies to count), to earthquakes (innumerable, but most recently "San Andreas") and takeovers by computers/robots (n+1 "Terminator" movies), etc. Asteroids? Genetically modified organisms? Aliens? Let's not go there. Making catastrophe the subject of so much fiction with little technical accuracy measurably impacts the public perception of catastrophic risk (Satpahi and Smith, Undated) and thus makes more challenging the rational quantification and serious scientific consideration of such risks.

IV.8 BOOK REVIEW

BOOK INFO

Title: *Quantifying and Controlling Catastrophic Risks*

Author and Date: B. John Garrick (2008)

Publisher: Academic Press xxii + 351 pages

REVIEWER INFO

Reviewer: D. Warner North

Link: http://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.2010.01508.x/full

BOOK INFO

Author: Marc Gerstein (with Michael Ellsberg)

Title: *Flirting with Disaster: Why Accidents Are Rarely Accidental*

Publication: New York: Union Square Press, 2008

Author: Nancy G. Leveson

Title: *Engineering a Safer World: Systems Thinking Applied to Safety*

Publication: Cambridge, MA: The MIT Press, 2011

(Available for free download: http://mitpress.mit.edu/books/engineering-safer-world)

Author: Nate Silver

Title: *The Signal and the Noise: Why So Many Predictions Fail, but Some Don't*

Publication: New York: The Penguin Press, 2012

Author: Nassim Nicholas Taleb

Title: *Antifragile: Things That Gain from Disorder*

Publication: New York: Random House, 2012

Author: James Owen Weatherall

Title: *The Physics of Wall Street: A Brief History of Predicting the Unpredictable*

Publication: Boston: Houghton Mifflin Harcourt, 2013

INTRODUCTION

Sally Kane, Book Review Editor for Risk Analysis, invited me to follow my review of three books in 2012 with another "mega-review" of books I thought important for the Risk Analysis community. These five books are my selection, and I recommend all of them, though for different purposes, as discussed next. One of them, Nate Silver's the Signal and the Noise, is also reviewed in more detail in this issue by Songjong Roh, a student of Area Editor Katherine McComas. Two are on a subject that has become a major interest for me, safety culture and how better safety can be achieved

IV.10 THE ECONOMIC IMPACT OF SPACE WEATHER - WHERE DO WE STAND

AUTHOR AND LINK

Author: J.P. Eastwood, E. Biffis, M.A. Hapgood, L. Green, M.M. Bisi, R.D. Bentley, R. Wicks, L.-A. Mickinnell, M. Gibbs, and C. Burnett

Link: http://onlinelibrary.wiley.com/doi/10.1111/risa.12765/full

ABSTRACT

Space weather describes the way in which the Sun, and conditions in space more generally, impact human activity and technology both in space and on the ground. It is now well understood that space weather represents a significant threat to infrastructure resilience, and is a source of risk that is wide-ranging in its impact and the pathways by which this impact may occur. Although space weather is growing rapidly as a field, work rigorously assessing the overall economic cost of space weather appears to be in its infancy. Here, we provide an initial literature review to gather and assess the quality of any published assessments of space weather impacts and socioeconomic studies. Generally speaking, there is a good volume of scientific peer-reviewed literature detailing the likelihood and statistics of different types of space weather phenomena. These phenomena all typically exhibit "power-law" behavior in their severity. The literature on documented impacts is not as extensive, with many case studies, but few statistical

studies. The literature on the economic impacts of space weather is rather sparse and not as well developed when compared to the other sections, most probably due to the somewhat limited data that are available from end-users. The major risk is attached to power distribution systems and there is disagreement as to the severity of the technological footprint. This strongly controls the economic impact. Consequently, urgent work is required to better quantify the risk of future space weather events.

IV.11 CAN SISYPHUS SUCCEED? GETTING U.S. HIGH-LEVEL NUCLEAR WASTE INTO A GEOLOGICAL RESPOSITORY

AUTHOR AND LINK

Author: D. Warner North

Link: https://www.ncbi.nlm.nih.gov/pubmed/23311528

ABSTRACT

The U.S. government has the obligation of managing the high-level radioactive waste from its defense activities and also, under existing law, from civilian nuclear power generation. This obligation is not being met. The January 2012 Final Report from the Blue Ribbon Commission on America's Nuclear Future provides commendable guidance but little that is new. The author, who served on the federal Nuclear Waste Technical Review Board from 1989 to 1994 and subsequently on the Board on Radioactive Waste Management of the National Research Council from 1994 to 1999, provides a perspective both on the Commission's recommendations and a potential path toward progress in meeting the federal obligation. By analogy to Sisyphus of Greek mythology, our nation needs to find a way to roll the rock to the top of the hill and have it stay there, rather than continuing to roll back down again.

IV.12 SPACE WEATHER: INTRODUCING A SURVEY PAPER – AND A RECENT EXECUTIVE ORDER

AUTHOR AND LINK

Author: D. Warner North

Link: http://onlinelibrary.wiley.com/doi/10.1111/risa.12778/full

INTRODUCTION

As an Area Editor for Risk Analysis, it is my pleasure and privilege to manage the review of submitted manuscripts. It is gratifying to find an excellent survey paper on a risk that is relatively unfamiliar to many of our readers, and to bring it though our editorial process to publication. This is the case with our lead paper for this February issue, "The Economic Impact of Space Weather – Where Do We Stand?"

IV.13 DEPARTMENT OF HOMELAND SECURITY BIOTERRORISM RISK ASSESSMENT: A CALL FOR CHANGE

CONTRIBUTOR AND LINK

Contributors: National Research Council; Division on Engineering and Physical Sciences; Division on Earth and Life Studies; Board on Mathematical Sciences and Their Applications; Board on Life Sciences; Committee on Methodological Improvements to the Department of Homeland Security's Biological Agent Risk Analysis

Link: https://www.nap.edu/catalog/12206/department-of-homeland-security-bioterrorism-risk-assessment-a-call-for

DESCRIPTION

The mission of *Department of Homeland Security Bioterrorism Risk Assessment: A Call for Change*, the book published in December 2008, is to independently and scientifically review the methodology that led to the 2006 Department of Homeland Security report, *Bioterrorism Risk Assessment* (BTRA) and provide a foundation for future updates.

This book identifies a number of fundamental concerns with the BTRA of 2006, ranging from mathematical and statistical mistakes that have corrupted results, to unnecessarily complicated probability models and models with fidelity far exceeding existing data, to more basic questions about how terrorist behavior should be modeled.

Rather than merely criticizing what was done in the BTRA of 2006, this new NRC book consults outside experts and collects a number of proposed alternatives that could improve DHS's ability to assess potential terrorist behavior as a key element of risk-informed decision making, and it explains these alternatives in the specific context of the BTRA and the bioterrorism threat.

IV.14 REVIEW OF THE DEPARTMENT OF HOMELAND SECURITY'S APPROACH TO RISK ANALYSIS

CONTRIBUTOR AND LINK

Contributor: National Research Council; Committee to Review the Department of Homeland Security's Approach to Risk Analysis

Link: https://www.nap.edu/catalog/12972/review-of-the-department-of-homeland-securitys-approach-to-risk-analysis

DESCRIPTION

The events of September 11, 2001 changed perceptions, rearranged national priorities, and produced significant new government entities, including the U.S. Department of Homeland Security (DHS) created in 2003. While the principal mission of DHS is to lead efforts to secure the nation against those forces that wish to do harm, the department also has responsibilities in regard to preparation for and response to other hazards and disasters, such as floods, earthquakes, and other "natural" disasters. Whether in the context of preparedness, response or recovery from terrorism, illegal entry to the country, or natural disasters, DHS is committed to processes and methods that feature risk assessment as a critical component for making better-informed decisions.

Review of the Department of Homeland Security's Approach to Risk Analysis explores how DHS is building its capabilities in risk analysis to inform decision making. The department uses risk analysis to inform decisions ranging from high-level policy choices to fine-scale protocols that guide the minute-by-minute actions of DHS employees. Although DHS is responsible for mitigating a range of threats, natural disasters, and pandemics, its risk analysis efforts are weighted heavily toward terrorism. In addition to assessing the capability of DHS risk analysis methods to support decision-making, the book evaluates the quality of the current approach to estimating risk and discusses how to improve current risk analysis procedures.

Review of the Department of Homeland Security's Approach to Risk Analysis recommends that DHS continue to build its integrated risk management framework. It also suggests that the department improve the way models are developed and used and follow time-tested scientific practices, among other recommendations.

IV.15 UNDERSTANDING AND MANAGING RISK IN SECURITY SYSTEMS FOR THE DOE NUCLEAR WEAPONS COMPLEX

CONTRIBUTORS AND LINK

Contributors: National Research Council; Division on Earth and Life Studies; Nuclear and Radiation Studies Board; Committee on Risk-Based Approaches for Securing the DOE Nuclear Weapons Complex

Link: https://www.nap.edu/catalog/13108/understanding-and-managing-risk-in-security-systems-for-the-doe-nuclear-weapons-complex

DESCRIPTION

A nuclear weapon or a significant quantity of special nuclear material (SNM) would be of great value to a terrorist or other adversary. It might have particular value if acquired from a U.S. facility--in addition to acquiring a highly destructive tool, the adversary would demonstrate an inability of the United States to protect its nuclear assets. The United States expends considerable resources toward maintaining effective

security at facilities that house its nuclear assets. However, particularly in a budget-constrained environment, it is essential that these assets are also secured efficiently, meaning at reasonable cost and imposing minimal burdens on the primary missions of the organizations that operate U.S. nuclear facilities.

It is in this context that the U.S. Congress directed the National Nuclear Security Administration (NNSA)--a semi-autonomous agency in the U.S. Department of Energy (DOE) responsible for securing nuclear weapons and significant quantities of SNM--asked the National Academies for advice on augmenting its security approach, particularly on the applicability of quantitative and other risk-based approaches for securing its facilities. In carrying out its charge, the committee has focused on what actions NNSA could take to make its security approach more effective and efficient.

The committee concluded that the solution to balancing cost, security, and operations at facilities in the nuclear weapons complex is not to assess security risks more quantitatively or more precisely. This is primarily because there is no comprehensive analytical basis for defining the attack strategies that a malicious, creative, and deliberate adversary might employ or the probabilities associated with them. However, using structured thinking processes and techniques to characterize security risk could improve NNSA's understanding of security vulnerabilities and guide more effective resource allocation.

IV.16 LESSONS LEARNED FROM THE FUKUSHIMA NUCLEAR ACCIDENT FOR IMPROVING SAFETY AND SECURITY OF U.S. NUCLEAR PLANTS: PHASE 2

CONTRIBUTORS AND LINK

Contributors: National Academies of Sciences, Engineering, and Medicine; Division on Earth and Life Studies; Nuclear and Radiation Studies Board; Committee on Lessons Learned from the Fukushima Nuclear Accident for Improving Safety and Security of U.S. Nuclear Plants

Link: https://www.nap.edu/catalog/21874/lessons-learned-from-the-fukushima-nuclear-accident-for-improving-safety-and-security-of-us-nuclear-plants

DESCRIPTION

The U.S. Congress asked the National Academy of Sciences to conduct a technical study on lessons learned from the Fukushima Daiichi nuclear accident for improving safety and security of commercial nuclear power plants in the United States. This study was carried out in two phases: Phase 1, issued in 2014, focused on the causes of the Fukushima Daiichi accident and safety-related lessons learned for improving nuclear plant systems, operations, and regulations exclusive of spent fuel storage. This Phase 2 report focuses on three issues: (1) lessons learned from the accident for nuclear plant security, (2) lessons learned for spent fuel storage, and (3) reevaluation of conclusions from previous Academies studies on spent fuel storage.

IV.17 LESSONS LEARNED FROM THE FUKUSHIMA NUCLEAR ACCIDENT FOR IMPROVING SAFETY OF U.S. NUCLEAR PLANTS

CONTRIBUTORS AND LINK

Contributors: National Research Council; Division on Earth and Life Studies; Nuclear and Radiation Studies Board; Committee on Lessons Learned from the Fukushima Nuclear Accident for Improving Safety and Security of U.S. Nuclear Plants

Link: https://www.nap.edu/catalog/18294/lessons-learned-from-the-fukushima-nuclear-accident-for-improving-safety-of-us-nuclear-plants

DESCRIPTION

The March 11, 2011, Great East Japan Earthquake and tsunami sparked a humanitarian disaster in northeastern Japan. They were responsible for more than 15,900 deaths and 2,600 missing persons as well as physical infrastructure damages exceeding $200 billion. The earthquake and tsunami also initiated a severe nuclear accident at the Fukushima

Daiichi Nuclear Power Station. Three of the six reactors at the plant sustained severe core damage and released hydrogen and radioactive materials. Explosion of the released hydrogen damaged three reactor buildings and impeded onsite emergency response efforts. The accident prompted widespread evacuations of local populations, large economic losses, and the eventual shutdown of all nuclear power plants in Japan.

Lessons Learned from the Fukushima Nuclear Accident for Improving Safety and Security of U.S. Nuclear Plants is a study of the Fukushima Daiichi accident. This report examines the causes of the crisis, the performance of safety systems at the plant, and the responses of its operators following the earthquake and tsunami. The report then considers the lessons that can be learned and their implications for U.S. safety and storage of spent nuclear fuel and high-level waste, commercial nuclear reactor safety and security regulations, and design improvements. *Lessons Learned* makes recommendations to improve plant systems, resources, and operator training to enable effective ad hoc responses to severe accidents. This report's recommendations to incorporate modern risk concepts into safety regulations and improve the nuclear safety culture will help the industry prepare for events that could challenge the design of plant structures and lead to a loss of critical safety functions.

In providing a broad-scope, high-level examination of the accident, *Lessons Learned* is meant to complement earlier evaluations by industry and regulators. This in-depth review will be an essential resource for the nuclear power industry, policy makers, and anyone interested in the state of U.S. preparedness and response in the face of crisis situations.

IV.18 IMPROVING THE ASSESSMENT OF THE PROLIFERATION RISK OF NUCLEAR FUEL CYCLES

CONTRIBUTORS AND LINK

Contributors: National Research Council; Division on Earth and Life Studies; Nuclear and Radiation Studies Board; Committee on Improving the Assessment of the Proliferation Risk of Nuclear Fuel Cycles

Link: https://www.nap.edu/catalog/18335/improving-the-assessment-of-the-proliferation-risk-of-nuclear-fuel-cyclesdescription

DESCRIPTION

The material that sustains the nuclear reactions that produce energy can also be used to make nuclear weapons—and therefore, the development of nuclear energy is one of multiple pathways to proliferation for a non-nuclear weapon state. There is a tension between the development of future nuclear fuel cycles and managing the risk of proliferation as the number of existing and future nuclear energy systems expands throughout the world. As the Department of Energy (DOE) and other parts of the government make decisions about future nuclear fuel cycles, DOE would like to improve proliferation assessments to better inform those decisions.

Improving the Assessment of the Proliferation Risk of Nuclear Fuel Cycles considers how the current methods of quantification of proliferation risk are being used and implemented, how other approaches to risk assessment can contribute to improving the utility of assessments for policy and decision makers. The study also seeks to understand the extent to which technical analysis of proliferation risk could be improved for policy makers through research and development.

V LIST OF ATTENDEES

Name	Role	Title	Affiliation
Aguirre, Anthony	Participant	Professor	UC Santa Cruz
Apostolakis, George	Moderator	Chairman	Japan Nuclear Risk Research Center
Barrett, Anthony	Participant	Co-Founder and Director of Research	Global Catastrophic Risk Institute
Baum, Seth	Lecturer	Executive Director	Global Catastrophic Risk Institute
Bier, Vicki	Participant	Professor	University of Wisconsin
Bley, Dennis	Participant	President	Buttonwood Consulting
Bourgon, Malo	Participant	COO at MIRI	Machine Intelligence Research Institute
Carnesale, Albert	Lecturer	Professor; Chancellor Emeritus	UC Los Angeles
Centeno, Miguel A.	Participant	Professor	Princeton
Cornwall, John	Participant	Professor Emeritus	UC Los Angeles
Crowley, Kevin	Participant	Division Director	US National Academy of Engineering
Diaconeasa, Mihai A.	Participant	Postdoctoral Scholar	UC Los Angeles
Dhir, Vijay	Participant	Professor	UC Los Angeles
Duettmann, Allison	Participant	Research Staff	Foresight Institute
Frankel, Oliver	Participant	Retired	Goldman Sachs
Garrick, B. John	Participant	Adjunct Professor	UC Los Angeles
Hamrick, Barbara	Participant	Radiation Safety Officer	UC Irvine
Hartman, Steve A.	Participant	President	S A Hartman & Associates
Jackson, Christopher	Participant	Research Staff	The B. John Garrick Institute of the Risk Sciences

Johnson, David	Participant	Consultant; GIRS Affiliate	The B. John Garrick Institute of the Risk Sciences
Katona, Peter	Lecturer	Professor	UC Los Angeles Geffen School of Medicine
Kosson, David	Participant	Professor; Director of CRESP	Vanderbilt University
Lauta, Kristian	Participant	Professor	University of Copenhagen
Liu, Hin-Yan	Participant	Professor	University of Copenhagen
Matheny, Jason	Participant	Director	Intelligence Advanced Research Projects Activity
McCarthy, Roger	Moderator	President	McCarthy Engineering
Mennen, Alex	Participant	Graduate Student	UC Los Angeles
Miller, Mark	Participant	Senior Fellow	Foresight Institute
Mosleh, Ali	Participant	Professor	UC Los Angeles
North, Warner	Participant	President	Northworks
Ó hÉigeartaigh, Seán	Lecturer	Executive Director	Center for the Study of Existential Risk
Parson, Edward	Participant	Professor; Faculty Co-Director	UC Los Angeles Law
Peterson, Christine	Lecturer	Writer and Advisor	Self Employed
Petroski, Henry	Moderator	Professor	Duke University
Ramberg, Bennett	Participant	Writer	Self Employed
Rhodes, Catherine	Lecturer	Academic Project Manager	Center for the Study of Existential Risk
Sandberg, Anders	Lecturer	Senior Research Fellow	University of Oxford
Shepherd, Joseph E.	Participant	Professor; VP of Student Affairs	California Institute of Technology
Suphamongkhon, Kantathi	Participant	Senior Fellow	UC Los Angeles

Unwin, Stephen	Participant	Division Director	Pacific North National Laboratory
Von Winterfeldt, Detlof	Lecturer	Professor; Director of CREATE	University of Southern California
Whipple, Chris	Participant	Writer	Self Employed
Wiener, Jonathan	Lecturer	Professor	Duke University
Woo, Gordon	Participant	Adjunct Professor	Nanyang Technological University, Singapore

REFERENCES

"1984 Rajneeshee Bioterror Attack." *Wikipedia*, October 20, 2017. https://en.wikipedia.org/w/index.php?title=1984_Rajneeshee_bioterror_attack&oldid=806201668.

"About OpenAI." OpenAI. Accessed November 8, 2017. https://openai.com/about/.

Agreelist. "Universal Basic Income - Do You Agree?" Agreelist.com, 2017. http://www.agreelist.org/a/basic-income.

Aguirre, Anthony. "Metaculus," 2017. https://www.metaculus.com/questions/.

Al, R. Acuna-Soto et. "Megadrought and Megadeath in 16th Century Mexico - Volume 8, Number 4—April 2002 - Emerging Infectious Disease Journal - CDC." Accessed November 3, 2017. https://doi.org/10.3201/eid0804.010175.

Alibek, Ken, and Stephan Handelman. *Biohazard*. Accessed October 31, 2017. https://www.penguinrandomhouse.com/books/2070/biohazard-by-ken-alibek-with-stephen-handelman/9780385334969.

Ambrose, Stanley H. "Late Pleistocene Human Population Bottlenecks, Volcanic Winter, and Differentiation of Modern Humans." *Journal of Human Evolution* 34, no. 6 (June 1, 1998): 623–51. https://doi.org/10.1006/jhev.1998.0219.

Armstrong Stuart. "Smarter Than Us." Machine Intelligence Research Institute. Accessed November 13, 2017. https://intelligence.org/smarter-than-us/.

Aziz, Haris, and Simon Mackenzie. "A Discrete and Bounded Envy-Free Cake Cutting Protocol for Any Number of Agents." *ArXiv:1604.03655 [Cs]*, April 13, 2016. http://arxiv.org/abs/1604.03655.

Barrett, Anthony M., and Seth D. Baum. "A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis." *Journal of Experimental & Theoretical Artificial Intelligence* 29, no. 2 (March 4, 2017): 397–414. https://doi.org/10.1080/0952813X.2016.1186228.

Barrett, Anthony Michael. "Value of Global Catastrophic Risk (GCR) Information: Cost-Effectiveness-Based Approach for GCR Reduction." *Decision Analysis* 14, no. 3 (August 24, 2017): 187–203. https://doi.org/10.1287/deca.2017.0350.

Bartholomew, Robert E., and Benjamin Radford. *The Martians Have Landed!: A History of Media-Driven Panics and Hoaxes*. McFarland, 2011.

Baum, Seth D. "The Far Future Argument for Confronting Catastrophic Threats to Humanity: Practical Significance and Alternatives." *Futures*, Confronting Future Catastrophic Threats To Humanity, 72, no. Supplement C (September 1, 2015): 86–96. https://doi.org/10.1016/j.futures.2015.03.001.

———. "The Great Downside Dilemma for Risky Emerging Technologies." *Physica Scripta* 89, no. 12 (2014): 128004. https://doi.org/10.1088/0031-8949/89/12/128004.

Baum, Seth D., David C. Denkenberger, Joshua M. Pearce, Alan Robock, and Richelle Winkler. "Resilience to Global Food Supply Catastrophes." *Environment Systems and Decisions* 35, no. 2 (June 1, 2015): 301–13. https://doi.org/10.1007/s10669-015-9549-2.

Baum, Seth D., and Itsuki C. Handoh. "Integrating the Planetary Boundaries and Global Catastrophic Risk Paradigms." *Ecological Economics* 107, no. Supplement C (November 1, 2014): 13–21. https://doi.org/10.1016/j.ecolecon.2014.07.024.

Baumgaertner, Emily. "Trump's Proposed Budget Cuts Trouble Bioterrorism Experts." *The New York Times*, May 28, 2017, sec. Politics. https://www.nytimes.com/2017/05/28/us/politics/biosecurity-trump-budget-defense.html.

Beckstead, Nicholas. "On the Overwhelming Importance of Shaping the Far Future." Ph.D., Rutgers The State University of New Jersey - New Brunswick, 2013. https://search.proquest.com/docview/1442191960/abstract/5180B9289DC64F48PQ/1.

Benatar, David. "Why It Is Better Never to Come into Existence." *American Philosophical Quarterly* 34, no. 3 (1997): 345–55.

Bennett, James C. "Private Communications to Christine Peterson," 2017.

"Biological Weapons, Bioterrorism, and Vaccines | History of Vaccines." Accessed October 31, 2017. https://www.historyofvaccines.org/content/articles/biological-weapons-bioterrorism-and-vaccines.

"Biological Weapons Convention." *Wikipedia*, September 25, 2017. https://en.wikipedia.org/w/index.php?title=Biological_Weapons_Convention&oldid=802284264.

Bostrom, Nick. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15, no. 3 (November 2003): 308–14. https://doi.org/10.1017/S0953820800004076.

———. "Ethical Issues in Advanced Artificial Intelligence," January 1, 2003.

———. "Existential Risk Prevention as Global Priority." *Global Policy* 4, no. 1 (February 1, 2013): 15–31. https://doi.org/10.1111/1758-5899.12002.

———. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards" 9 (March 2003). https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c.

———. "Pascal's Mugging." *Analysis* 69, no. 3 (2009): 443–45.

———. *Superintelligence: Paths, Dangers, Strategies*. OUP Oxford, 2014.

Bostrom, Nick, and Milan M. Cirkovic. *Global Catastrophic Risks*. OUP Oxford, 2011.

Bovsun, M. "Guru of Poison: Bioterrorrists Spread Salmonella in Oregon." NY Daily News, May 13, 2013. http://www.nydailynews.com/news/justice-story/guru-poison-bioterrorrists-spread-salmonella-oregon-article-1.1373864.

Bradley, Ben. "The Value of Endangered Species." *The Journal of Value Inquiry* 35, no. 1 (March 1, 2001): 43–58. https://doi.org/10.1023/A:1010383322591.

Brin, David. "The Transparent Society." *Harvard Journal of Law & Technology* 12 (1999 1998): 513–32.

Brooks, Teresa Nicole. "Survey of Automated Vulnerability Detection and Exploit Generation Techniques in Cyber Reasoning Systems." *ArXiv:1702.06162 [Cs]*, February 20, 2017. http://arxiv.org/abs/1702.06162.

Broome, John. "Discounting the Future." *Philosophy & Public Affairs* 23, no. 2 (April 1, 1994): 128–56. https://doi.org/10.1111/j.1088-4963.1994.tb00008.x.

Buterin, Vitalik. "Hard Fork Completed." *Ethereum Blog* (blog), July 20, 2016. https://blog.ethereum.org/2016/07/20/hard-fork-completed/.

Cairns-Smith, A. G. *Genetic Takeover: And the Mineral Origins of Life*. 1 edition. Cambridge: Cambridge University Press, 1987.

Centeno, Miguel A., Manish Nag, Thayer S. Patterson, Andrew Shaver, and A. Jason Windawi. "The Emergence of Global Systemic Risk." *Annual Review of Sociology* 41, no. 1 (2015): 65–85. https://doi.org/10.1146/annurev-soc-073014-112317.

Chernov, Dmitry, and Didier Sornette. *Man-Made Catastrophes and Risk Information Concealment: Case Studies of Major Disasters and Human Fallibility*. Springer, 2015.

Chui, Michael, James Manyika, and Mehdi Miremadi. "Four Fundamentals of Workplace Automation | McKinsey & Company." Accessed November 14, 2017. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/four-fundamentals-of-workplace-automation.

Ćirković, Milan M. "Small Theories and Large Risks—Is Risk Analysis Relevant for Epistemology?" *Risk Analysis* 32, no. 11 (November 1, 2012): 1994–2004. https://doi.org/10.1111/j.1539-6924.2012.01914.x.

Clancey, Tom. "Tom Clancy Quotes." Accessed October 31, 2017. http://thinkexist.com/quotation/the_difference_between_fiction_and_reality/225193.html.

Clouceff, M, and J Greenhalgh. "The Next Pandemic Could Be Dripping On Your Head." NPR.org. Accessed October 31, 2017. http://www.npr.org/sections/goatsandsoda/2017/02/21/508060742/the-next-pandemic-could-be-dripping-on-your-head.

Coase, R. H. "The Problem of Social Cost." *The Journal of Law & Economics* 3 (1960): 1–44.

Cotton-Barrat, Owen, Sebastian Farquahar, John Halstead, Stefan Schubert, and Andrew Snyder-Beattie. "Global Catastrophic Risks 2016." Global Challenges Foundation, 2016. https://api.globalchallenges.org/static/reports/Global-Catastrophic-Risk-Annual-Report-2016.pdf.

Cotton-Barrat, Owen, and T Ord. “Existential Risk and Existential Hope.” Future of Humanity Institute -- Technical Report #2015-1, 2015. https://www.fhi.ox.ac.uk/Existential-risk-and-existential-hope.pdf.

Cotton-Barrat, Owen, and A Sandberg. “Classifying Risks of Human Extinction.” Forthcoming, 2017.

Council, National Research, Division on Earth and Life Studies, Board on Life Sciences, Division on Engineering and Physical Sciences, Board on Mathematical Sciences and Their Applications, and Committee on Methodological Improvements to the Department of Homeland Security’s Biological Agent Risk Analysis. *Department of Homeland Security Bioterrorism Risk Assessment: A Call for Change*. National Academies Press, 2008.

Cropper, M. L. “Regulating Activities with Catastrophic Environmental Effects.” *Journal of Environmental Economics and Management* 3, no. 1 (June 1, 1976): 1–15. https://doi.org/10.1016/0095-0696(76)90009-7.

Crutzen, Paul J. “Albedo Enhancement by Stratospheric Sulfur Injections: A Contribution to Resolve a Policy Dilemma?” *Climatic Change* 77, no. 3–4 (August 1, 2006): 211. https://doi.org/10.1007/s10584-006-9101-y.

Denkenberger, David, and Joshua M. Pearce. *Feeding Everyone No Matter What: Managing Food Security After Global Catastrophe*. Academic Press, 2014.

Derbes, Vincent J. “De Mussis and the Great Plague of 1348.” *JAMA* 196, no. 1 (April 4, 1966): 59–62. https://doi.org/10.1001/jama.1966.03100140113030.

Dezecache, Guillaume. “Human Collective Reactions to Threat.” *Wiley Interdisciplinary Reviews: Cognitive Science* 6, no. 3 (May 1, 2015): 209–19. https://doi.org/10.1002/wcs.1344.

Diamond, Jared. *Collapse: How Societies Choose to Fail or Succeed*. Penguin, 2005.

Dillon, Robin L., and Catherine H. Tinsley. “How Near-Misses Influence Decision Making Under Risk: A Missed Opportunity for Learning.” *Management Science* 54, no. 8 (June 10, 2008): 1425–40. https://doi.org/10.1287/mnsc.1080.0869.

Dillon, Robin L., Catherine H. Tinsley, and Matthew Cronin. “Why Near-Miss Events Can Decrease an Individual’s Protective Response to Hurricanes.” *Risk Analysis* 31, no. 3 (March 1, 2011): 440–49. https://doi.org/10.1111/j.1539-6924.2010.01506.x.

Doudna, Jennifer A., and Samuel H. Sternberg. *A Crack in Creation: Gene Editing and the Unthinkable Power to Control Evolution*. Houghton Mifflin Harcourt, 2017.

Dowd, Maureen. “Elon Musk’s Billion-Dollar Crusade to Stop the A.I. Apocalypse,” 2017. https://cacm.acm.org/careers/215162-elon-musks-billion-dollar-crusade-to-stop-the-a-i-apocalypse/fulltext.

Dr. Patricia Lewis, Sasan Aghlani, and Benoît Pelopidas Heather Williams. “Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy.” Chatham House, April 28, 2014. https://www.chathamhouse.org//node/13981.

"Drake Equation." *Wikipedia*, November 1, 2017. https://en.wikipedia.org/w/index.php?title=Drake_equation&oldid=808232889.

Drexler, Eric. *Engines of Creation: The Coming Era of Nanotechnology*. Reprint edition. New York: Anchor, 1987.

———. *MDL Intelligence Distillation: Exploring Strategies for Safe Access to Superintelligent Problem-Solving Capabilities*, 2015.

Drexler, K. Eric. "Nanotechnology and Exploratory Engineering." presented at the Stanford Univesity course taught spring quarter 1988, Standford University, 1988.

———. "Reframing AI Prospects (Index Page)." Google Docs, 2017. https://docs.google.com/document/d/145yJBoNTYHOJ_FMOO2hO-x2KnJQT45hxhtd0I84HVLE/edit?usp=embed_facebook.

Drexler, K. Eric, and Dennis Pamlin. "Nano-Solutions for the 21st Century." Low Carbon Leaders, 2013. https://www.oxfordmartin.ox.ac.uk/downloads/academic/201310Nano_Solutions.pdf.

Drury, John, David Novelli, and Clifford Stott. "Psychological Disaster Myths in the Perception and Management of Mass Emergencies." *Journal of Applied Social Psychology* 43, no. 11 (November 1, 2013): 2259–70. https://doi.org/10.1111/jasp.12176.

Duettman, Allison. "The Reflective Equilibrium as Ethical Standard for AI." London School of Economics, 2014. Master's Thesis in the Department of Philosophy, Logic and Scientific Method.

Edwards, Richard. "Poison-Tip Umbrella Assassination of Georgi Markov Reinvestigated," June 19, 2008, sec. News. http://www.telegraph.co.uk/news/2158765/Poison-tip-umbrella-assassination-of-Georgi-Markov-reinvestigated.html.

Eliezer, Yudkowsky. "Rationality: From AI to Zombies." Machine Intelligence Research Institute, March 11, 2015. https://intelligence.org/rationality-ai-zombies/.

Executive Office of the President. "National Electric Grid Security and Resilience Action Plan." *Https://Www.Whitehouse.Gov*, December 1, 2016. https://www.hsdl.org/?abstract&did=797486.

Ezell, Barry Charles, Steven P. Bennett, Detlof Von Winterfeldt, John Sokolowski, and Andrew J. Collins. "Probabilistic Risk Analysis and Terrorism Risk." *Risk Analysis* 30, no. 4 (April 1, 2010): 575–89. https://doi.org/10.1111/j.1539-6924.2010.01401.x.

Fahmida Y. Rashid. "Cyber Attack on Power Grid Could Top $1 Trillion in Damage: Report | SecurityWeek.Com," June 16, 2015. http://www.securityweek.com/cyber-attack-power-grid-could-top-1-trillion-damage-report?v=_H-uxRq2w-c.

Farquhar, Sebastian, John Halstead, Owen Cotton-Barrat, Haydn Belfield, Stefan Schubert, and Andrew Snyder-Beattie. "Existential Risk -- Diplomacy and Governance." Global Priorities Project, 2017. https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf.

Filardo, N. W. “Research Report: Mitigating LangSec Problems with Capabilities.” In *2016 IEEE Security and Privacy Workshops (SPW)*, 189–97, 2016. https://doi.org/10.1109/SPW.2016.57.

Fiona O’Cleirigh. “Bill Binney, the ‘Original’ NSA Whistleblower, on Snowden, 9/11 and Illegal Surveillance.” ComputerWeekly.com, 2015. http://www.computerweekly.com/feature/Interview-the-original-NSA-whistleblower.

Fisher, Kathleen. “Using Formal Methods to Enable More Secure Vehicles: DARPA’s HACMS Program.” In *Proceedings of the 19th ACM SIGPLAN International Conference on Functional Programming*, 1–1. ICFP ’14. New York, NY, USA: ACM, 2014. https://doi.org/10.1145/2628136.2628165.

Freudenberg, Drew, and Jean Tirole. “Game Theory.” MIT Press, 1991. https://mitpress.mit.edu/books/game-theory.

Friend, Tad. “Sam Altman’s Manifest Destiny.” *The New Yorker*, October 3, 2016. https://www.newyorker.com/magazine/2016/10/10/sam-altmans-manifest-destiny.

Frye, P. “Life On Mars Bad? Star Trek Dream Could Be Killed By NASA’s 2020 Mars Rover.” The Inquisitr, September 30, 2014. https://www.inquisitr.com/1508522/life-on-mars-bad-star-trek-dream-could-be-killed-by-nasas-2020-mars-rover/.

Future of Life Institute. *Interactions between the AI Control Problem and the Governance Problem | Nick Bostrom*. Accessed December 1, 2017. https://www.youtube.com/watch?v=_H-uxRq2w-c.

Garcia, Ryan J. B., and Detlof von Winterfeldt. “Defender–Attacker Decision Tree Analysis to Combat Terrorism.” *Risk Analysis* 36, no. 12 (December 1, 2016): 2258–71. https://doi.org/10.1111/risa.12574.

Garrick, B. John. *Quantifying and Controlling Catastrophic Risks*. Academic Press, 2008.

Gates, Bill. “A New Kind of Terrorism Could Wipe out 30 Million People in Less than a Year — and We Are Not Prepared.” Business Insider. Accessed November 22, 2017. http://www.businessinsider.com/bill-gates-op-ed-bio-terrorism-epidemic-world-threat-2017-2.

“GCR Concept Project | Global Catastrophic Risk Institute.” Accessed November 20, 2017. http://gcrinstitute.org/concept/.

Gernot Wagner, and Martin L. Weitzman. “Climate Shock.” Princeton University Press. Accessed October 10, 2017. https://press.princeton.edu/titles/10414.html.

Gjerde, Jon, Sverre Grepperud, and Snorre Kverndokk. “Optimal Climate Policy under the Possibility of a Catastrophe.” *Resource and Energy Economics* 21, no. 3 (August 1, 1999): 289–317. https://doi.org/10.1016/S0928-7655(99)00006-8.

“Global Catastrophic Risk.” *Wikipedia*, December 10, 2017. https://en.wikipedia.org/w/index.php?title=Global_catastrophic_risk&oldid=814764514.

Goertzel, Ben. “Artificial General Intelligence: Concept, State of the Art, and Future Prospects.” *Journal of Artificial General Intelligence* 5, no. 1 (2014): 1–48. https://doi.org/10.2478/jagi-2014-0001.

———. “Superintelligence: Fears, Promises and Potentials,” November 2015. http://jetpress.org/v25.2/goertzel.htm.

Goldman, Minton. *Revolution and Change in Central and Eastern Europe: Political, Economic and Social Challenges*. 1 edition. Armonk, NY: Routledge, 2016.

Good, Irving John. “Speculations Concerning the First Ultraintelligent Machine**Based on Talks given in a Conference on the Conceptual Aspects of Biocommunications, Neuropsychiatric Institute, University of California, Los Angeles, October 1962; and in the Artificial Intelligence Sessions of the Winter General Meetings of the IEEE, January 1963 [1, 46].The First Draft of This Monograph Was Completed in April 1963, and the Present Slightly Amended Version in May 1964.I Am Much Indebted to Mrs. Euthie Anthony of IDA for the Arduous Task of Typing.” In *Advances in Computers*, edited by Franz L. Alt and Morris Rubinoff, 6:31–88. Elsevier, 1966. https://doi.org/10.1016/S0065-2458(08)60418-0.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. “When Will AI Exceed Human Performance? Evidence from AI Experts.” *ArXiv:1705.08807 [Cs]*, May 24, 2017. http://arxiv.org/abs/1705.08807.

Graham, David A. “Rumsfeld’s Knowns and Unknowns: The Intellectual History of a Quip.” *The Atlantic*, March 27, 2014. https://www.theatlantic.com/politics/archive/2014/03/rumsfelds-knowns-and-unknowns-the-intellectual-history-of-a-quip/359719/.

Gustin, Sam. “IBM Says It Hasn’t Given the NSA Any Client Data.” Time Magazine, 2014. http://time.com/25410/ibm-nsa-letter/.

Hadhazy, Adam. “How We Could Actually Build a Space Colony.” Popular Mechanics, October 2, 2014. http://www.popularmechanics.com/science/space/deep/how-we-could-actually-build-a-space-colony-17268252.

Häggström, Olle. *Here Be Dragons: Science, Technology and the Future of Humanity*. Oxford University Press, 2016.

Hanson, Robin. “Combinatorial Information Market Design.” *Information Systems Frontiers* 5, no. 1 (January 1, 2003): 107–19. https://doi.org/10.1023/A:1022058209073.

———. “When the Economy Transcends Humanity.” *The Futurist* 48 (January 1, 2014): 27–30.

Harms, Tracy. “NanoCon: The Cosmic Pie, Harms.” *Nanocon Proceedings*, 1989, Appendices.

Harris, Sam. “Surviving the Cosmos,” 2016. https://www.samharris.org/podcast/item/surviving-the-cosmos/.

Herfst, Sander, Eefje J. A. Schrauwen, Martin Linster, Salin Chutinimitkul, Emmie de Wit, Vincent J. Munster, Erin M. Sorrell, et al. “Airborne Transmission of Influenza A/H5N1 Virus

Between Ferrets." *Science* 336, no. 6088 (June 22, 2012): 1534–41. https://doi.org/10.1126/science.1213362.

Hertsgaard, Mark. "Mikhail Gorbachev Explains What's Rotten in Russia." Accessed October 18, 2017. https://www.salon.com/2000/09/07/gorbachev/.

Homer-Dixon, Thomas, Brian Walker, Reinette Biggs, Anne-Sophie Crépin, Carl Folke, Eric Lambin, Garry Peterson, et al. "Synchronous Failure: The Emerging Causal Architecture of Global Crisis." *Ecology and Society* 20, no. 3 (July 14, 2015). https://doi.org/10.5751/ES-07681-200306.

Hook, Sidney. In *A Free Man's Choice*, 10–12. 1958a. The New Leader, 26 May.

———. In *Bertrand Russell Retreats*, 25–28. 1958b. The New Leader, 7-14 July.

"Interview: K. Eric Drexler." *Bulletin of the Atomic Scientists* 63, no. 1 (January 1, 2007): 55–58. https://doi.org/10.2968/063001018.

John Garrick, B., James E. Hall, Max Kilger, John C. McDonald, Tara O'Toole, Peter S. Probst, Elizabeth Rindskopf Parker, et al. "Confronting the Risks of Terrorism: Making the Right Decisions." *Reliability Engineering & System Safety*, Confronting the risks of terrorism: making the right decisions, 86, no. 2 (November 1, 2004): 129–76. https://doi.org/10.1016/j.ress.2004.04.003.

John Hopkins Bloomberg School of Public Health. "Centers for Disease Control and Prevention (CDC) Classification of Bioterrorism Microorganisms," May 23, 2017. http://ocw.jhsph.edu/courses/BiologicalAgentsOfWaterAndFoodborneBioterrorism/PDFs/WaterFoodTerror3.pdf.

Jun. 20, and 2017. "Bioterrorism Rule Blocks Some U.S. Researchers from Studying Bird Flu." Science | AAAS, October 4, 2017. http://www.sciencemag.org/news/2017/06/bioterrorism-rule-blocks-some-us-researchers-studying-bird-flu.

Kaplan, Stanley, and B. John Garrick. "On The Quantitative Definition of Risk." *Risk Analysis* 1, no. 1 (March 1, 1981): 11–27. https://doi.org/10.1111/j.1539-6924.1981.tb01350.x.

Kennaway, Robert N. M. Watson, Jonathan Anderson, Ben Laurie, Kris. "A Taste of Capsicum: Practical Capabilities for UNIX." Accessed November 15, 2017. https://cacm.acm.org/magazines/2012/3/146252-a-taste-of-capsicum/fulltext.

Konopinski, Emil. "Ignition of the Atmosphere with Nuclear Bombs," August 14, 1946. http://library.sciencemadness.org/lanl1_a/lib-www/la-pubs/00329010.html.

Kornwitz, Jason. "The Cybersecurity Risk of Self-Driving Cars," 2017. https://phys.org/news/2017-02-cybersecurity-self-driving-cars.html.

Kunreuther, Howard, and Geoffrey Heal. "Managing Catastrophic Risk." Working Paper. National Bureau of Economic Research, June 2012. https://doi.org/10.3386/w18136.

Kunreuther, Howard, Nathan Novemsky, and Daniel Kahneman. “Making Low Probabilities Useful.” *Journal of Risk and Uncertainty* 23, no. 2 (September 1, 2001): 103–20. https://doi.org/10.1023/A:1011111601406.

Kurzweil, Ray. “Ray Kurzweil: Don’t Fear Artificial Intelligence.” Time Magazine, 2014. http://time.com/3641921/dont-fear-artificial-intelligence/.

Kurzweil, Ray, and Christopher Lane. *How to Create a Mind: The Secret of Human Thought Revealed*. MP3 Una edition. Brilliance Audio, 2012.

Kuznick, Peter J. “Prophets of Doom or Voices of Sanity? The Evolving Discourse of Annihilation in the First Decade and a Half of the Nuclear Age.” *Journal of Genocide Research* 9, no. 3 (September 1, 2007): 411–41. https://doi.org/10.1080/14623520701528940.

Lengel, Allan. “Little Progress In FBI Probe of Anthrax Attacks,” September 16, 2005. http://www.washingtonpost.com/wp-dyn/content/article/2005/09/15/AR2005091502456.html.

Leslie, John. *The End of the World: The Science and Ethics of Human Extinction*. Psychology Press, 1998.

Levy, Steven. “How Elon Musk and Y Combinator Plan to Stop Computers From Taking Over.” WIRED, 2015. https://www.wired.com/2015/12/how-elon-musk-and-y-combinator-plan-to-stop-computers-from-taking-over/.

Lipsitch, Marc, and Thomas V. Inglesby. “Moratorium on Research Intended To Create Novel Potential Pandemic Pathogens.” *MBio* 5, no. 6 (December 31, 2014): e02366-14. https://doi.org/10.1128/mBio.02366-14.

Loria, Kevin. “Over and over Again, the Military Has Conducted Dangerous Biowarfare Experiments on Americans.” Business Insider. Accessed November 2, 2017. http://www.businessinsider.com/over-and-over-again-americans-have-been-subjected-to-secret-military-experiments-2015-9.

Luke Muehlhauser. “When Will AI Be Created?” Machine Intelligence Research Institute, May 15, 2013. https://intelligence.org/2013/05/15/when-will-ai-be-created/.

Maher, Timothy M., and Seth D. Baum. “Adaptation to and Recovery from Global Catastrophe.” *Sustainability* 5, no. 4 (March 28, 2013): 1461–79. https://doi.org/10.3390/su5041461.

Mallah, Richard. “The Landscape of AI Safety and Beneficence Research: Presentation on the Set of Technical Research Topics for Robust, Interpretable, and Safe AI.” e27. Accessed November 15, 2017. https://e27.co/event/the-landscape-of-ai-safety-and-beneficence-research-presentation-on-the-set-of-technical-research-topics-for-robust-interpretable-and-safe-ai/.

Martin, Ian W. R., and Robert S. Pindyck. “Averting Catastrophes: The Strange Economics of Scylla and Charybdis.” *American Economic Review* 105, no. 10 (October 2015): 2947–85. https://doi.org/10.1257/aer.20140806.

Matheny, Jason G. “Reducing the Risk of Human Extinction.” *Risk Analysis* 27, no. 5 (October 1, 2007): 1335–44. https://doi.org/10.1111/j.1539-6924.2007.00960.x.

McLarty, Thomas F, and Thomas L. Ridge. “SECURING THE US ELECTRICAL GRID,” 2014.

Mellers, Barbara A, Alan Schwartz, Katty Ho, and Ilana Ritov. “Decision Affect Theory: Emotional Reactions to the Outcomes of Risky Options.” *Psychological Science* 8, no. 6 (November 1, 1997): 423–29. https://doi.org/10.1111/j.1467-9280.1997.tb00455.x.

Miller, James D., and D. Felton. “The Fermi Paradox, Bayes’ Rule, and Existential Risk Management.” *Futures* 86, no. Supplement C (February 1, 2017): 44–57. https://doi.org/10.1016/j.futures.2016.06.008.

Miller, Mark S. “A Series of Private Communications to Christine Peterson in the 1980s and 1990s,” 1980s-90s.

Miller, Mark S., Tom Van Cutsem, and Bill Tulloh. “Distributed Electronic Rights in JavaScript,” 2013. https://research.google.com/pubs/pub40673.html.

Miller, Mark S., and K. Eric Drexler. “Comparative Ecology: A Computational Perspective.” In *The Ecology of Computation*. Elsevier Science Publishers, 1988. https://e-drexler.com/d/09/00/AgoricsPapers/agoricpapers/ce/ce0.html.

Miller, Mark S., and Bill Tulloh. “Decision Alignment (Extended Abstract),” 2016. https://research.google.com/pubs/pub45571.html.

———. “The Elements of Decision Alignment.” In *30th European Conference on Object-Oriented Programming (ECOOP 2016)*, edited by Shriram Krishnamurthi and Benjamin S. Lerner, 56:17:1–17:5. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2016. https://doi.org/10.4230/LIPIcs.ECOOP.2016.17.

Miller, Mark S., Ka-Ping Yee, and Jonathan Shapiro. “Capability Myths Demolished.” Technical Report SRL2003-02, Johns Hopkins University Systems Research Laboratory, 2003. http://www. erights. org/elib/capability/duals, 2003.

Millett, Piers, and Andrew Snyder-Beattie. “Human Agency and Global Catastrophic Biorisks.” *Health Security* 15, no. 4 (July 26, 2017): 335–36. https://doi.org/10.1089/hs.2017.0044.

Minsky, Marvin. *The Society of Mind*. Pages Bent edition. New York: Simon & Schuster, 1988.

Morris, M. “Forgotten Horrors: The Human Experiments of Unit 731.” *KnowledgeNuts* (blog). Accessed November 3, 2017. http://knowledgenuts.com/2013/07/23/forgotten-horrors-the-human-experiments-of-unit-731/.

Mueller, John. “Is There Still a Terrorist Threat?: The Myth of the Omnipresent Enemy.” *Foreign Affairs*, September 1, 2006. https://www.foreignaffairs.com/articles/2006-09-01/there-still-terrorist-threat-myth-omnipresent-enemy.

Munthe, Christian. *The Black Hole Challenge: Precaution, Existential Risks and the Problem of Knowledge Gaps*. Vol. Accepted, 2017.

Nanotechnology for Chemical and Biological Defense | Margaret Kosal | Springer. Accessed November 14, 2017. //www.springer.com/us/book/9781441900616.

Nebehay, Stephanie. “Avoid Caves in Uganda after Marburg Death: WHO.” *Reuters*, July 11, 2008. https://www.reuters.com/article/us-uganda-marburg/avoid-caves-in-uganda-after-marburg-case-who-advice-idUSL1166433320080711.

Ng, Andrew. “Andrew Ng: Why ‘Deep Learning’ Is a Mandate for Humans, Not Just Machines.” WIRED, 2015. https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines/.

Ng, Y.-K. “Should We Be Very Cautious or Extremely Cautious on Measures That May Involve Our Destruction?” *Social Choice and Welfare* 8, no. 1 (February 1, 1991): 79–88. https://doi.org/10.1007/BF00182449.

Nichols, Tom. *The Death of Expertise: The Campaign Against Established Knowledge and Why It Matters*. Oxford University Press, 2017.

Nordholz, Jan Christoph, and Jean-Pierre Seifert. “Efficient Virtualization on Hardware with Limited Virtualization Support,” 2011.

O’Malley, Pat. “Catastrophe: Risk and Response, by Richard A. PosnerOxford University Press, 2004; Vii + 322 Pp.” *Law, Probability and Risk* 4, no. 3 (September 1, 2005): 187–89. https://doi.org/10.1093/lpr/mgi011.

Omohundro, Steve. “Autonomous Technology and the Greater Human Good.” *Journal of Experimental & Theoretical Artificial Intelligence* 26, no. 3 (July 3, 2014): 303–15. https://doi.org/10.1080/0952813X.2014.895111.

O’Neill, Gerard K, David Gump, Space Studies Institute, and Space Frontier Foundation. *The High Frontier: Human Colonies in Space*. Burlington, Ont.: Apogee Books, 1976. http://catalog.hathitrust.org/api/volumes/oclc/45644248.html.

O’Neill, John. “The Varieties of Intrinsic Value.” *The Monist* 75, no. 2 (May 1, 1992): 119–37. https://doi.org/10.5840/monist19927527.

“OpenAI — General Support.” Open Philanthropy Project, February 1, 2017. https://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence/openai-general-support.

Ouagrham, Ben. “Biological Weapons Threats from the Former Soviet Union.” *Llechtenstein Institute of Self-Determination at Princeton University*, Working Paper Series on Russia and the Former Soviet States, August 2003.

Palmer, Megan J., Bruce C. Tiu, Amy S. Weissenbach, and David A. Relman. “On Defining Global Catastrophic Biological Risks.” *Health Security* 15, no. 4 (August 1, 2017): 347–48. https://doi.org/10.1089/hs.2017.0057.

Parfit, Derek. *Reasons and Persons*. OUP Oxford, 1984.

Patel, Prachi. “Computer Vision Leader Fei-Fei Li on Why AI Needs Diversity.” IEEE Spectrum: Technology, Engineering, and Science News, October 19, 2016. https://spectrum.ieee.org/tech-talk/at-work/tech-careers/computer-vision-leader-feifei-li-on-why-ai-needs-diversity.

Pinker, Steven. *The Better Angels of Our Nature: Why Violence Has Declined*. Reprint edition. New York Toronto London: Penguin Books, 2012.

Popper, Karl R., Alan Ryan, and E. H. Gombrich. *The Open Society and Its Enemies*. New One-Volume edition with a New introduction by Alan Ryan and an essay by E. H. Gombrich edition. Princeton: Princeton University Press, 1945.

"Precautionary Principle." *Wikipedia*, December 5, 2017. https://en.wikipedia.org/w/index.php?title=Precautionary_principle&oldid=813753683.

Rampino, Michael R., and Stanley H. Ambrose. "Volcanic Winter in the Garden of Eden: The Toba Supereruption and the Late Pleistocene Human Population Crash." *Special Paper of the Geological Society of America* 345 (2000): 71–82. https://doi.org/10.1130/0-8137-2345-0.71.

Randle, Melanie, and Richard Eckersley. "Public Perceptions of Future Threats to Humanity and Different Societal Responses: A Cross-National Study." *Futures*, Confronting Future Catastrophic Threats To Humanity, 72, no. Supplement C (September 1, 2015): 4–16. https://doi.org/10.1016/j.futures.2015.06.004.

Raup, David M. "Cohort Analysis of Generic Survivorship." *Paleobiology* 4, no. 1 (January 1978): 1–15. https://doi.org/10.1017/S0094837300005649.

Redlener, Irwin. *Americans at Risk: Why We Are Not Prepared for Megadisasters and What We Can Do*. Knopf Doubleday Publishing Group, 2006.

Rees, Martin. *Our Final Century: Will Civilisation Survive the Twenty-First Century?: Will the Human Race Survive the Twenty-First Century?* New Ed edition. London: Arrow, 2004.

Reinhardt, Jason C., Xi Chen, Wenhao Liu, Petar Manchev, and M. Elisabeth Paté-Cornell. "Asteroid Risk Assessment: A Probabilistic Approach." *Risk Analysis* 36, no. 2 (February 1, 2016): 244–61. https://doi.org/10.1111/risa.12453.

REUTERS, China Photo /. "Civet Cat Becomes SARS Scapegoat." msnbc.com, January 8, 2004. http://www.nbcnews.com/id/3908790/ns/health-infectious_diseases/t/civet-cat-becomes-sars-scapegoat/.

Risley, James. "Elon Musk, Peter Thiel, Reid Hoffman and Others Commit $1B to Stop AI from Taking over the World." GeekWire, December 11, 2015. https://www.geekwire.com/2015/non-profit-ai-research-facility-backed-by-elon-musk-peter-thiel-created-to-protect-against-corperate-ai/.

Rosoff, H., and D. Von Winterfeldt. "A Risk and Economic Analysis of Dirty Bomb Attacks on the Ports of Los Angeles and Long Beach." *Risk Analysis* 27, no. 3 (June 1, 2007): 533–46. https://doi.org/10.1111/j.1539-6924.2007.00908.x.

Russell, Allan Dafoe and Stuart. "Yes, the Experts Are Worried about the Existential Risk of Artificial Intelligence." MIT Technology Review. Accessed October 10, 2017. https://www.technologyreview.com/s/602776/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/.

Russell, Bertrand. “Freedom to Survive,” 7–4. The New Leader, 1958. http://ucelinks.cdlib.org:8888/sfx_local?sid=google&auinit=B&aulast=Russell&atitle=Freedom+to+Survive&title=The+New+Leader&date=1958&spage=7&issn=0028-6044.

———. In *World Communism and Nuclear War*, 9–10. 1958a. New Leader, 26 May. http://ucelinks.cdlib.org:8888/sfx_local?sid=google&auinit=B&aulast=Russell&atitle=World+Communism+and+Nuclear+War&title=The+New+Leader&date=1958&spage=9&issn=0028-6044.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3 edition. Upper Saddle River: Pearson, 2009.

Sagan, Carl. “Nuclear War and Climatic Catastrophe: Some Policy Implications.” *Foreign Affairs* 62, no. 2 (1983): 257–92. https://doi.org/10.2307/20041818.

Schlesinger, Arthur Meier. *A Thousand Days: John F. Kennedy in the White House*. Houghton Mifflin Harcourt, 2002.

Schlosser, Eric. *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety*. Penguin, 2013.

Schulte, Peter, Laia Alegret, Ignacio Arenillas, José A. Arz, Penny J. Barton, Paul R. Bown, Timothy J. Bralower, et al. “The Chicxulub Asteroid Impact and Mass Extinction at the Cretaceous-Paleogene Boundary.” *Science* 327, no. 5970 (March 5, 2010): 1214–18. https://doi.org/10.1126/science.1177265.

Selk, Avi. “Bill Gates: Bioterrorism Could Kill More than Nuclear War — but No One Is Ready to Deal with It.” *Washington Post*, February 18, 2017, sec. WorldViews. https://www.washingtonpost.com/news/worldviews/wp/2017/02/18/bill-gates-bioterrorism-could-kill-more-than-nuclear-war-but-no-one-is-ready-to-deal-with-it/.

Sell, Tara Kirk. “How Trump’s Budget Makes Us All Vulnerable to Bioterrorism.” Text. TheHill, May 31, 2017. http://thehill.com/blogs/pundits-blog/homeland-security/335813-how-the-trumps-budget-makes-us-all-vulnerable-to.

Simpson, Corbin. *Monte: A Dynamic Language Inspired by Python and E*. Emacs Lisp. 2014. Reprint, monte-language, 2017. https://github.com/monte-language/monte.

“Smallpox.” *Wikipedia*, October 25, 2017. https://en.wikipedia.org/w/index.php?title=Smallpox&oldid=807066337.

Soares, Nate. “The Value Learning Problem.” *Machine Intelligence Research Institute, Berkeley, CA, USA*, 2015. https://intelligence.org/files/ValueLearningProblem.pdf.

Steffen, Will, Katherine Richardson, Johan Rockström, Sarah E. Cornell, Ingo Fetzer, Elena M. Bennett, Reinette Biggs, et al. “Planetary Boundaries: Guiding Human Development on a Changing Planet.” *Science* 347, no. 6223 (February 13, 2015): 1259855. https://doi.org/10.1126/science.1259855.

Steunebrink, Bas R., Kristinn R. Thórisson, and Jürgen Schmidhuber. “Growing Recursive Self-Improvers.” In *International Conference on Artificial General Intelligence*, 129–139. Springer, 2016.

Szigeti, Balázs, Padraig Gleeson, Michael Vella, Sergey Khayrulin, Andrey Palyanov, Jim Hokanson, Michael Currie, Matteo Cantarelli, Giovanni Idili, and Stephen Larson. "OpenWorm: An Open-Science Approach to Modeling Caenorhabditis Elegans." *Frontiers in Computational Neuroscience* 8 (2014): 137. https://doi.org/10.3389/fncom.2014.00137.

Tallinn, Jaan. *AI and Value Alignment*. 2017 Beneficial AI Conference, 2017. https://www.youtube.com/watch?v=d6pIk-JxfGw.

———. "CSaP Distinguished Lecture: The Intelligence Stairway - Networks of Evidence and Expertise for Public Policy," 2012. http://www.csap.cam.ac.uk/events/csap-distinguished-lecture-intelligence-stairway/.

The White House. "Homeland Security Presidential Directive / HSPD-10: Biodefense for the 21st Century," April 28, 2004. https://fas.org/irp/offdocs/nspd/hspd-10.html.

Tonn, Bruce. "Transcending Oblivion." *Futures* 31 (April 1, 1999): 351–59. https://doi.org/10.1016/S0016-3287(98)00137-2.

Tonn, Bruce, and Dorian Stiefel. "Evaluating Methods for Estimating Existential Risks." *Risk Analysis* 33, no. 10 (October 1, 2013): 1772–87. https://doi.org/10.1111/risa.12039.

Toynbee, Arnold Joseph, and D. C. Somervell. *A Study of History. Abridgement of V. I-Vi by D.C. Somervell*. Oxford University Press, 1947.

Villar, Rodrigo G., Sean P. Elliott, and Karen M. Davenport. "Botulism: The Many Faces of Botulinum Toxin and Its Potential for Bioterrorism." *Infectious Disease Clinics of North America* 20, no. 2 (June 2006): 313–327, ix. https://doi.org/10.1016/j.idc.2006.02.003.

Walsh, Toby. "The Singularity May Never Be Near." *ArXiv:1602.06462 [Cs]*, February 20, 2016. http://arxiv.org/abs/1602.06462.

Wampler, Robert A, and Thomas S Blanton. "Anthrax at Sverdlovsk, 1979." The National Security Archive, November 15, 2001. https://nsarchive2.gwu.edu/NSAEBB/NSAEBB61/.

Watson, Robert NM, Peter G. Neumann, Jonathan Woodruff, Jonathan Anderson, Ross Anderson, Nirav Dave, Ben Laurie, Simon W. Moore, Steven J. Murdoch, and Philip Paeps. "CHERI: A Research Platform Deconflating Hardware Virtualization and Protection." In *Workshop Paper, Runtime Environments, Systems, Layering and Virtualized Environments (RESoLVE 2012)*, 2012.

Weber, Robert C. "A Letter to Our Clients About Government Access to Data." THINK Blog, March 14, 2014. https://www.ibm.com/blogs/think/2014/03/open-letter-data/.

Wiener, Jonathan B. "The Tragedy of the Uncommons: On the Politics of Apocalypse." *Global Policy* 7 (May 1, 2016): 67–80. https://doi.org/10.1111/1758-5899.12319.

"WikiLeaks - The Hackingteam Archives." Accessed November 15, 2017. https://wikileaks.org/hackingteam/emails/emailid/96008.

Wilson, Grant. "Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law Note." *Virginia Environmental Law Journal* 31 (2013): 307–64.

Woo, Gordon. "Counterfactual Disaster Risk Analysis." Variance Journal, 2016.

Yampolskiy, Roman V. “From Seed AI to Technological Singularity via Recursively Self-Improving Software.” *ArXiv:1502.06512 [Cs]*, February 23, 2015. http://arxiv.org/abs/1502.06512.

Yampolskiy, Roman V., and M. S. Spellchecker. “Artificial Intelligence Safety and Cybersecurity: A Timeline of AI Failures.” *ArXiv:1610.07997 [Cs]*, October 25, 2016. http://arxiv.org/abs/1610.07997.

Yang, Wesley. “Is the ‘Anthropocene’ Epoch a Condemnation of Human Interference — or a Call for More?” *The New York Times*, February 14, 2017, sec. Magazine. https://www.nytimes.com/2017/02/14/magazine/is-the-anthropocene-era-a-condemnation-of-human-interference-or-a-call-for-more.html.

Yudkowsky, Eliezer. “Coherent Extrapolated Volition.” *Singularity Institute for Artificial Intelligence*, 2004.

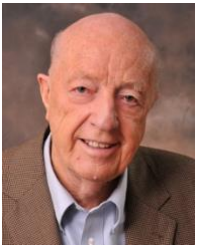
———. “Corporations vs. Superintelligences.” Accessed November 15, 2017. https://arbital.com/p/corps_vs_si/.

———. “Task-Directed AGI,” 2016. https://arbital.com/p/task_agi/.

———. “The Value Loading Problem,” 2015. https://www.edge.org/response-detail/26198.

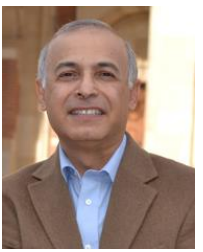
THE B. JOHN GARRICK INSTITUTE FOR THE RISK SCIENCES

B. JOHN GARRICK, FOUNDER



B. John Garrick, Ph.D. and M.S., Engineering and Applied Science, UCLA; B.S., Physics, BYU; graduate, Oak Ridge School of Reactor Technology, is a pioneer in the risk sciences relating to the assessment and management of the risk of complex systems, both natural and man-made. He has authored texts on the risk sciences and served a White House appointment for two terms as Chairman of the U.S. Nuclear Waste Technical Review Board. He retired as CEO of PLG, Inc., an international engineering and management consulting firm following the start of his career as a physicist for the U.S. Atomic Energy Commission. He is a Distinguished Adjunct Professor of Engineering and Applied Science, UCLA, and a fellow of three professional societies, the American Nuclear Society, the Society for Risk Analysis, and the Institute for the Advancement of Engineering. He is a past President of the international Society for Risk Analysis, receiving that society's highest award, the Distinguished Achievement Award. Dr. Garrick was elected to the National Academy of Engineering in 1993 "for making quantitative risk assessment an applied science and a fundamental part of engineering design." He is founder and senior advisor of the UCLA B. John Garrick Institute for the Risk Sciences and received the 2014 Alumnus of the Year award from the UCLA Henry Samueli School of Engineering and Applied Science.

ALI MOSLEH, DIRECTOR



Dr. Ali Mosleh is a Distinguished University Professor and Evelyn Knight Chair in Engineering at UCLA where he is also the director of the UCLA Garrick Institute for the Risk Sciences. Previously he was the Nicole J. Kim Eminent Professor of Engineering and Director of the Center for Risk and Reliability at the University of Maryland. He was elected to the US National Academy of Engineering in 2010, and is a Fellow of the Society for Risk Analysis, and the American Nuclear Society, recipient of several scientific achievement awards, and technical advisor to numerous organizations, including appointment by President George W. Bush to the U.S. Nuclear Waste Technical Review Board. He conducts research on methods for probabilistic risk analysis and reliability of complex systems and has made contributions in diverse fields of theory and application. He holds several patents, and has edited and authored over 500 publications.

Published by:

The B. John Garrick Institute for the Risk Sciences
University of California, Los Angeles

Engineering VI
404 Westwood Plaza
Los Angeles, CA 90095

+1 (310) 794-5141
info@risksciences.ucla.edu