QUANTIFYING GLOBAL CATASTROPHIC RISKS
One Day Workshop

PROCEEDINGS

Editor: Dr. B. John Garrick

Sponsored by:
The B. John Garrick Institute for the Risk Sciences, University of California, Los Angeles
The Centre for the Study of Existential Risk, University of Cambridge, United Kingdom
Global Catastrophic Risk Institute, USA

Held at:
Luskin Convention Center, UCLA
September 16, 2018

https://risksciences.ucla.edu
QUANTIFYING GLOBAL CATASTROPHE RISKS
One Day Workshop

PROCEEDINGS

Editor: Dr. B. John Garrick
This workshop on Quantifying Global Catastrophic Risks is a spinoff of the First International Colloquium on Catastrophic and Existential Risk held at the same location March 27-29, 2017. Many of the same experts and support staff were involved in both conferences. Seth Baum represented the Global Catastrophic Risk Institute and Julius Weitzdörfer represented The Center for the Study of Existential Risk, University of Cambridge, United Kingdom. The third sponsor was the host, The Garrick Institute for the Risk Sciences, University of California, Los Angeles. The Editor acknowledges the assistance of Ali Mosleh and Mihai Diaconeasa of the Garrick Institute and Barbara Hamrick of the University of California, Irvine, not only for their participation in the workshop but their expert and professional input to the proceedings. We wish to thank David Johnson, Chairman of the PSAM 14 conference, for allowing us to piggyback this workshop on the PSAM 14 international conference. We wish also to acknowledge the technical support for audio and visual aids provided by Jeonghyun Lee, Arjun, and Yuan-Shang Chang of the Garrick Institute.
# TABLE OF CONTENTS

Acknowledgments ................................................................. iii

Table of Contents ................................................................... iv

I  Introduction ........................................................................... 1

II  Summary and Evaluation ...................................................... 3
  II.1  Workshop Overview .......................................................... 3
  II.2  Presentation Summaries ..................................................... 4
  II.3  Takeaway Messages .......................................................... 16

III  Presenter Papers ................................................................. 19
  III.1  Reflections on the Risk Analysis of Nuclear War – Seth Baum ... 19
  III.4  Reproduced: Classifying Global Catastrophic Risks – Julius Weitzdörfer .... 94
  III.5  Summary: The Varieties of Uncertainty Quantification – Stephen Unwin .... 103
  III.6  Environmental and Natural Disasters: “Surprise” versus “Degradation” – Holly Buck, Peter Kareiva, Claire Hirashiki ................................................................. 106

IV  Presenters and Biographical Sketches ................................... 125

V  List of Attendees ..................................................................... 128
I  INTRODUCTION

Increasing our understanding of global catastrophic risks is key to their prevention, mitigation, and, for selected risks, our ability to delay their occurrence. Global catastrophic risks, referred to simply as “global risks,” are risks of events that could significantly harm or even destroy human civilization at the global scale (Hempsell, 2004; Baum and Barrett, 2017). In fact, some global risks are paths to extinction.

Catastrophic and Existential Risk was the subject of a colloquium held at UCLA on March 27-29, 2017 (Garrick [editor], 2017). A frequent observation during the colloquium was how essential it is to better quantify and prioritize such risks. This workshop has chosen to focus on “quantifying global catastrophic risks.”

Recognizing the need for better understanding global risks has been researched for decades by such scholars as Bostrom and Čirković (Bostrom and Čirković [editors], 2008). Selected quotes below from the “conclusions and future directions” of this reference strongly support the need for a more disciplined and analytical approach to characterizing global risks. Key points specifically relevant to the workshop are underlined.

“In the study of individual risks, focus more on producing actionable information, such as early-warning signs, metrics for measuring progress towards risk reduction, and quantitative models for risk assessment.”

“Develop and implement better methodologies and institutions for information aggregation and probabilistic forecasting...”

“Foster a critical discourse aimed at addressing questions of prioritization in a more reflective and analytical manner than is currently done; and consider global catastrophic risks and their mitigation within a broader context of challenges and opportunities for safeguarding and improving the human condition.”

The workshop was executed as part of the ongoing international discourse to better characterize global risks and seek methods for their quantification. Sponsors of the workshop were The B. John
Garrick Institute for the Risk Sciences, University of California, Los Angeles; The Centre for the Study of Existential Risk, University of Cambridge, United Kingdom; and the Global Catastrophic Risk Institute, USA. The approach taken was to invite some 30 scholars and practitioners from academia, research institutions, and the private sector to participate. The invitees were from several different nations. The format was six presentations taking about 55% of the time leaving about 45% of the time for discussion. Participants are listed in the appendices, including short Curricula Vitae on the presenters. The presentations, papers upon which they were based, or (in one case) a detailed synopsis of the presentation are included as a part of these proceedings.
II  SUMMARY AND EVALUATION

II.1  WORKSHOP OVERVIEW

The first three presentations were coordinated to provide a reference frame for discussing the quantitative assessment of global risks. This reference frame was staged as a high-level view of the principles and practices of current methods for quantifying facility and natural risks. The purpose was to consider if known methods could be extended to global risks; in particular, to those risks for which there was little direct evidence of their prior occurrence. Such risks might include events that may have previously occurred, but for which no human records of the events exist as they occurred before humans existed or before they had evolved to the point of making written records of such events.

The chosen reference frame included the identification of a global risk example to which current methods could be applied to add clarity to the methodological changes and research that may be required to in fact quantify global risks. The example chosen was the risk of nuclear war—a most challenging example to be sure, but a very real global risk and one of concern to all of society. The rarity of the use of nuclear weapons, as well as the very brief history of the nuclear age adds to the complexity of the quantification process.

To further establish the reference frame, it was considered essential to identify and describe all current unclassified investigations into modeling the risk of nuclear war. The first presentation took on this task. The second presentation provided the basic concepts and methods supporting current quantitative assessments of both natural and anthropogenic risks, including complex facilities handling dangerous and hazardous materials. The third presentation applied these methods to a particular initiating event scenario of an unintentional nuclear war. The underlying questions guiding all three presentations were as follows:

1. Are existing methods of quantifying facility and natural hazards risk extendable to the assessment of global risks?
2. To what extent do decision making processes impact global risks?
3. How can the risk of an event be quantified having never (or rarely) been observed by humans?

II.2 PRESENTATION SUMMARIES

Six presentations were given across a broad range of global catastrophic risk topics. A summary is provided below. The summaries were based on full papers and material provided in Section III.

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

- Reflections on the Risk Analysis of Nuclear War – Seth Baum
- A Quantitative Risk Assessment Study of the Initiation of an Inadvertent Nuclear War – Mihai A. Diaconeasa, Theresa Stewart
- Classifying Global Catastrophic Risks – Julius Weitzdörfer
- The Varieties of Uncertainty Quantification – Stephen Unwin
- Environmental and Natural Disasters: “Surprise” versus “Degradation” – Holly J. Buck, Claire Hirashiki

**The first presentation was made by Seth Baum**, the Executive Director of the Global Catastrophic Risk Institute and a distinguished scholar in the field. Baum summarized prior work in the nuclear war risk field, provided an overview of modeling nuclear war risk, and shared elements of the extensive unclassified data base that exists on near-miss scenarios for initiating a nuclear war. He also provided some insights on the policy implications of risk assessment work on such issues as nuclear disarmament and deterrence.
The prior work consists of several studies on the sequence of events necessary to initiate a nuclear war. The primary logic models employed were event trees, event sequence diagrams (scenarios) and fault trees, all described in some detail within the papers included in these proceedings. Different categories of initiators of nuclear war have been investigated in each of two general categories, namely intentional and unintentional. The investigations have resulted in the development of a robust amount of evidence of different scenarios that without intervention could have resulted in a nuclear conflict. Many of the scenarios have been classified as near-misses.

Of course, the Cuban Missile Crises of 1962 is considered by most scholars as the event that came closest to resulting in an exchange of nuclear weapons. Another example of a near miss was the January 1995 Norwegian weather rocket incident. The unidentified rocket launch was detected by both the USA and Russia. Russia, thinking that they may be under attack, followed their protocol for launching a counter attack. Fortunately, they ultimately concluded it was not a nuclear threat and did not launch the counter attack. Other near-misses have involved faulty computer chips, misinterpretation of training exercises, and mistaken atmospheric effects on satellite monitoring activities. In fact, there are 100s of events that have been identified as supporting evidence for structuring scenarios for assessing nuclear war risk.

The research and investigation performed by Baum and others have been critical to support taking the next step in the rigor of the risk assessments. While their work has definitely produced some insights on just what the nuclear war risk is, it has not advanced to the stage of being termed quantitative. That is the next step and was the theme for this workshop. The transition from qualitative to quantitative risk assessment is the topic of the second presentation.

The second presentation was made by B. John Garrick, Founder of UCLA’s B. John Garrick Institute for the Risk Sciences. Garrick presented the basic concepts that have evolved over five decades for quantifying the risk of complex
systems and natural hazards with the nuclear power industry clearly leading the way.

Garrick described the challenges of quantifying the risk of rare events, events that have never occurred, or events that occurred prior to humans being able to record them. He also noted scholars have been outspoken about the folly of trying to predict such events. He countered this objection with the suggestion that by changing the goal from pure prediction to an evidence-based assessment that approaches pure prediction, one can use quantitative risk assessment as an effective means to reach an end similar to that of an actual prediction.

A rigorous risk assessment can provide valuable insights to our understanding and expectations of rare events. Such an approach requires taking full advantage of all evidence related to the event in question, including the evidence of non-occurrence. Embracing the uncertainty sciences and utilizing Bayes’ theorem of inferential reasoning aid the maximization of the information content of the available evidence.

Central to quantifying the risk of any event is the definition of risk and risk assessment. The chosen definition by many practitioners and regulators is the so-called “triplet definition of risk.” The triplet definition is based on the premise that when one asks “what is the risk” of something, they are asking three fundamental questions, “What can happen?” “How likely is it that it will happen?” “If it does happen, what are the consequences?”

Quantitative risk assessment is defined as a process of probabilistic evidential and inferential analysis of the response of events, systems, or activities to different threats and challenges based on the fundamental rules of logic and plausible reasoning. The key is “the optimal processing of incomplete information.” To claim that the assessment is quantitative requires that the uncertainty be quantified to make transparent the answer to the question, “what is the risk?” Other terms critical to quantifying risk are “evidence,” “probability,” and “risk measure.”

Evidence is any information that can impact the likelihood or probability of events. Evidence includes, but is not restricted to,
data. While data is clearly evidence, the word connotes failure rates or event rates, whereas evidence encompasses all relevant information whether or not it is countable.

The chosen definition of probability that has served the risk community well has its roots in the work of E.T. Jaynes (Jaynes, 2003). Probability is defined as the credibility of a hypothesis based on the totality of the supporting evidence. It obeys Bayes’ theorem and the fundamental postulates of probability and plausible reasoning.

Risk can be measured as a “frequency,” a “probability,” or a “probability of frequency.” The latter has proved to be an effective risk measure for extremely rare events as it is generally easier to assess the frequency of an event than its actual probability. However, for the measure to be classified as “quantitative” the uncertainties in the frequencies must be quantified. This is achieved by applying Bayes theorem and the supporting evidence to construct a probability distribution about the frequency.

These concepts, definitions, and interpretations together with such other logic models as Bayesian networks enjoy tremendous success in assessing the risk of complex systems having extremely rare recurrence intervals, including the case of no human record of the event. Contemporary investigators (Wang, 2007) have labeled the integration of the algorithms of event trees, fault trees and Bayesian networks as the Hybrid Casual Logic (HCL) methodology. Appropriate integration of the three logic models represented by HCL can yield major insights into the risk of such global risks as super volcanoes, earth impacting events, and nuclear war.

The form of the results is discussed in the later sections of the proceedings in the actual paper presented. The results include risk curves, contributors to the risk ranked in terms of their importance, and specific scenarios dominating the risk.

The third presentation was made by Mihai A. Diaconeasa and Theresa Stewart. Diaconeasa is a Postdoctoral Research Scholar and Stewart is a UCLA engineering graduate student both affiliated with The B. John Garrick Institute for the Risk
The presentation and associated paper examined the risk of inadvertently initiating a nuclear war. The following caveat on this presentation and the example is important:

The example is not intended to be a comprehensive and rigorous assessment of the chosen scenario. That would require a much greater effort as well as the need to collaborate with National Laboratory subject matter experts and to have access to classified information. The primary goal was to introduce quantitative methods of analysis to a specific nuclear war scenario to build confidence in the ability to be more rigorous and quantitative in assessing the different scenarios by which a nuclear war could be initiated. In particular, the real results of the example are the methods and models employed, not the numbers. However, the numbers are not totally irrelevant as they are based on a variety of existing data sources and used to illustrate how the actual evidence is processed.

The goals of analyzing this example were to (1) identify the events that can stimulate the initiation of a nuclear war, (2) capture potential failure mechanisms, and (3) quantify the likelihood of those events, including their uncertainties.

The assessment of this event used quantitative, that is, probabilistic risk assessment (PRA) methods to examine the likelihood of a nuclear war caused by a false belief of a nuclear attack, specifically caused by a false alarm from a satellite detection. As indicated in the caveat above, this assessment is not intended to be a rigorous analysis of the risk of nuclear war. The emphasis is on methods, not a rigorous quantitative result. For this demonstration the authors relied on accessible information in the open literature. Therefore, the systems characterized are dated and the results reflect these limitations.

---

1 Quantitative risk assessment (QRA) and probabilistic risk assessment (PRA) are considered to be the same.
The key assumptions are (1) a launch under attack strategy, (2) intercontinental ballistic missile technology, (3) USA/Russian conditions of the 1980s, and (4) the use of only unclassified information.

The HCL methodology was developed for risk scenario analysis in PRAs of technological systems that consider not only the risks associated with “hard” causes such as hardware systems, but also the risks generated by “soft” causes such as human activities, physical environment, and socio-economic environment. This methodology offers a multi-layered modeling approach so that each individual domain of the system is modeled with the most appropriate technique. Embedded in the methodology are numerous algorithms, including the Phoenix Method (a model-based human reliability analysis methodology) of quantifying human performance, Bayesian inference analysis, and probability arithmetic methods such as Monte Carlo to propagate the uncertainties through the model.

The event sequence was modeled by starting with an initial detection of the false alarm by a satellite, following which the alarm must pass through five stages either by a deliberate decision at that stage, or through a time-out. If the false nature of the alarm is caught at any stage, there is still the possibility of a launch by a lower level official.

Failure to collect correct information, make the appropriate decision given the available information, or take an action is modeled at each stage using fault trees to identify the cause of the failure. These causes are linked to Bayesian networks which allow consideration of human and environmental factors. The initial false alarm detection via satellite and potential second detection of a false alarm via radar are both modeled using fault trees and consider causes from objects passing the detectors threshold and faults in the system itself.

Supporting evidence for quantifying the different phases of the decision-making process exists from multiple data sources including the extensive collections of Ellsberg, Hellman, and the...
Global Catastrophic Risk Institute. Detailed references are in the full paper included in these proceedings. Data exists on false alarms, including those caused by satellite and radar detection faults. Reliability databases are robust sources for failures of systems, subsystems, structures, and components. Expert knowledge is also a valuable source, especially in areas of “intent” and human behavior. Research, development and testing information is a major source for aerospace related systems.

In areas where evidence is sparse, such as respecting the matter of “intent” to initiate an attack or a counter attack, embracing the uncertainty and behavioral sciences as well as expert knowledge and other sources of relevant evidence ameliorates the challenge. Developing evidence for rare events or events not known to have ever occurred requires more extensive research and investigation than is generally practiced in most risk and reliability applications.

The Garrick paper included in these proceedings discusses the various ways in which risk results can be represented. The form of the results often used to show the risk of a particular event is a probability density function. For this exercise, the authors used a “probability of frequency format” as shown in the full paper. The example assessed illustrates with high confidence that the analytical apparatus is available to quantify rare global risks.

**The fourth presentation was made by Julius Weitzdörfer** of the Centre for the Study of Existential Risk, University of Cambridge, United Kingdom. Weitzdörfer, a lawyer, performs research on the law and governance of extreme technological risks. His presentation was on, “Classifying Global Catastrophic Risks,” the basis of which is the paper of the same title included in these proceedings and openly accessible.

The motivation for the research highlighted in the paper is the need for organizing global risks into a more structured form to facilitate their better understanding and the establishment of priorities for taking preventive and mitigating actions. The point is made that while there is reasonable understanding of the global risk scenarios of greatest concern, “there is currently no known methodology for compiling a comprehensive, interdisciplinary view of global catastrophic risks.” The classification system
provides a roadmap for the “convergence” between risk scenarios and insights on research needs as well as policy considerations. It also increases the transparency of the factors that affect global risks.

An important feature of the classification system is the breaking down of the analysis of global risk scenarios into three key components: (1) a critical system(s)\(^2\) whose safety boundaries are breached by a potential threat, (2) the mechanisms by which this threat might spread globally and affect the majority of the human population, and (3) the manner in which we might fail to prevent or mitigate both (1) and (2). The classification system facilitates the performance of global risk assessment by decomposing it into separate manageable parts such as separating the analysis of global spreading mechanisms from the analysis of critical systems. The decomposition process also makes clear such separations as natural global risks and anthropogenic risks including specific critical systems.

The authors advocate three ways that the classification system could be potentially used: (1) prioritization of risk reduction efforts; (2) creation of a live reference list of expertise for different risk scenarios; and, (3) highlighting of uncertain or neglected corners of the global risk possibility space.

While it is very clear that such a classification system would resolve many issues having to do with synthesizing the myriad of elements attendant to global risk in terms of the combinations and permutations involved, it is also clear that it is heavily dependent on the overall knowledge base of the risks. Two major factors in implementing the system as described are knowing at least something about the likelihood and consequences of the risk and the associated uncertainties. This brings us to the theme of the workshop, namely the need to better quantify global risks. Clearly, a classification system such as proposed provides a framework for systematically quantifying such risks. On the other hand, there would be extensive iteration between the quantification exercises and the structure of the classification

\(^2\) Critical system is defined as any system or process that, if disturbed beyond a certain limit or scale, could trigger a significant reduction in humanity’s ability to survive in its current form.
system. In other words, classifying the risks as proposed is only possible in combination with activities that make the risks more transparent in terms of their likelihood and consequence. Such analyses are the very foundation for establishing priorities for taking action to prevent or mitigate the consequences.

The fifth presentation was made by Stephen D. Unwin, Head of Risk Programs Development of the Pacific Northwest National Laboratory. As alluded to earlier, quantification of risks implies quantification of the uncertainties, which has to be part of the answer to the question, “what is the risk?” The oft taken approach of doing conservative analyses or worst case analysis to represent the uncertainties does not provide the necessary transparency to make the best decisions in managing risk—especially the risk of extremely rare events. The uncertainties must be front and center in the analysis to have real meaning for effective decision making.

Unwin noted that diverse approaches have been adapted to uncertainty quantification using methods based on fundamentally disparate concepts, uncertainty metrics, and means of analysis. A taxonomy of uncertainty models includes the following:

- Qualitative or bounding methods
- Classical statistical methods
- Bayesian probabilistic methods
- Non-Bayesian quantitative methods

An example of qualitative uncertainty analysis is a simple bounding approach in which model input parameters are varied over their ranges, and the corresponding ranges of the output parameters are determined. A major shortcoming of such an approach results from the fact that no weighting metric is placed on the input parameter ranges, such that while the limitations on, and uncertainties of, the input parameters may be reasonably well-characterized, they do not (necessarily) correspond to the limitations and uncertainties of the output, because probabilistic mechanics play no role in propagation. The upside of the method is there is no need to define and defend input probability distributions, but the price of poor output interpretability is high.
The use of statistical models to produce classical confidence intervals is a practical method only for an extremely narrow problem set; specifically, when a likelihood function is available to interpret the data, and where the model logic is sufficiently simple to allow propagation of the confidence intervals. Few real-world problems have such features.

Unwin discussed Bayesian methods, beginning with the conceptual revolution in which probability becomes interpreted as a measure of partial belief, thus allowing probabilities to be attached to general propositions and not only to statistical data. This put probabilities at the center of a theory of evidence, substantially broadening the potential domains of application, and examples of Bayesian analysis were described along with the decision/policy/regulatory questions to which they have been applied.

Common criticisms of Bayesianism as a theory of evidence were discussed, including the opinion that probability theory does not intuitively represent the way evidence influences belief, and two alternatives were described:

- Dempster-Shafer (D-S) theory, which is a generalization of probability theory.
- Possibility theory, which is based in the algebra of fuzzy logic.

In each of these approaches, belief is represented by two distinctive measures: Support versus Plausibility in the case of D-S theory, and Necessity versus Possibility in the case of Possibility theory. The former measure in each case reflects the degree to which evidence points towards the truth of a proposition, while the latter measure reflects the degree to which evidence fails to point away from the proposition. In this way, both theories propose remedies to the argument against the probabilistic model.

Finally, the presentation assessed the advantages and disadvantages of each theory of evidence, with a focus on comparison of Bayesian methods to D-S and Possibility theory, with Unwin concluding that Bayesian probability theory is the most
robust and defensible approach to uncertainty characterization in real-world applications.

There are the usual language problems in connection with the use of the word “belief” in this context. This editor is a disciple of Jaynes’ (Jaynes, 2003) concept of Bayesian probability assessment. Jaynes puts the emphasis on probabilities being “knowledge based” – that is, based on the supporting evidence and plausible reasoning – not “belief based.” Experts assign probabilities on the basis of “expert knowledge,” not “expert opinion” or “expert judgment.” Belief implies the concept of faith, and probabilities should not be based on faith, but on the supporting evidence. To be sure the evidence can be challenged, but that is where the debate should be—including on the assignment of a probability distribution that best represents the supporting evidence. Telling the truth should be the driving force, not whether it can be believed. The truth often involves uncertainty, which must be quantified—when done with rigor and transparency it is the component of a probability that separates “belief” from the debate.

**The sixth presentation was made by Holly Buck** who is a Nature-Set Science Fellow at UCLA’s Institute of the Environment and Sustainability. The title of her presentation was Environmental and Natural Disasters: “Surprise” versus “Degradation.”

The presentation provided a very interesting change of pace. As the title implies, the focus was on the environment and natural disasters. The presentation centered on three analogues that are believed informative about the dynamics of these risks: disasters, extinctions, and ecosystem collapses. Known disasters and ecosystem collapses, while not necessarily global or existential in terms of consequences, are nevertheless useful analogs for more consequential events. The worst disasters can be an analog to an existential risk. Ecosystem collapses are also more like an existential risk, because they are systemic failures.

The authors examined factors that led to the events investigated. These factors were categorized as unsolvables (the event was not a surprise, but no one knew how to prevent it); failures (there
were ways to prevent it, but no one took action); and, surprises (the event was not anticipated). The following two key research questions were posed during the presentation. Were the most devastating disasters or collapses surprises, unsolvables, or failures? Are there properties of systems that generically reduce existential risk by enhancing resilience?

The top ten disasters examined were primarily “failures” with only one that could be categorized as a “surprise.” That event was the double whammy of the earthquake and cholera epidemic in Haiti in 2010. The failures with respect to the top ten disasters were largely due to inadequate emergency preparedness. The randomly drawn disasters were mainly “unsolvables,” events which were anticipated and for which appropriate measures had been taken to the extent possible.

Next Buck and her colleagues considered extinctions, with illustrative examples chosen from the twentieth century and beyond, the time when reliable information about the causes of extinctions started to be available. The question raised was “when a species ends, is it a surprise?” Again, the categories were (1) unsolvables: the extinction is inevitable, but a practical solution is not; and, (2) failures: there is a practical solution and it may not be expensive, but extinction seems inevitable due to a failure of leadership, governance, and society. Looking at examples of the heath hen, African elephant, vaquita, and passenger pigeon, it was found that extinctions were generally not due to just one thing, that criminal behavior is hard to overcome, and that a highly unlikely run of bad luck can undermine what might seem to be a successful effort. In modern times, rarely is the failure to prevent extinction a reflection solely of governance or social failure.

For the final analysis, Buck and her colleagues used a dataset called the Red List of Ecosystems. This is a dataset maintained by the International Union for Conservation of Nature (IUCN), and their hypothesis is that ecosystem risk is a function of the species that compose them, their interactions, and the ecological processes they depend on. The research team considered ecosystems categorized as Critically Endangered, Endangered, and Vulnerable, and found that for most of these ecosystems, multiple factors contributed to their decline. Their evaluation of
the Aral Sea in more detail indicated that it had at least eight factors in its collapse.

The conclusion of their research was that these three analogies stress the importance of multiple whammies. A multi-dimensional examination of risk factors, and their synergistic interaction, must be fully integrated into risk analysis. This interaction of risks was also addressed by Weitzdörfer in building the case for the need to have a formal classification system for global and existential risks.

The authors also suggested that scholars should look deeper into non-failures, which may hold clues to the properties of resilient systems. The point is made that resilience is conferred by attributes such as heterogeneity, modular structures, redundancy, introducing negative feedback loops to counteract positive feedback loops, and anticipating surprises. Their opinion is that resilience hasn’t yet been operationalized, or translated from an ecological principle into social action. Their research has identified three important challenges: operationalizing resilience, anticipating surprises, and overcoming organized irresponsibility.

II.3 TAKEAWAY MESSAGES

- Catastrophic events initiated, exacerbated or otherwise modified by human action (or inaction), as well as rare natural events still pose a challenge to the quantification of the risk(s) of these events, but there are existing and developing methods that provide a platform for quantification. In general, much more technology is available for quantifying risk than is actually applied.

- While the ability to make quantitative “predictions” of extremely rare events may be challenged on the basis of too much uncertainty, there are surrogates such as quantitative risk assessment that can add considerable transparency to the likelihood of such events and therefore some insight on their timing.

- The absence of historical evidence for human-caused and rare natural events is overstated as a reason for not
being able to rigorously quantify global risks. If quantification is taken to mean an evidence-based probabilistic representation of the likelihood of a rare event, then seldom has the full panoply of evidence been used that is actually available for assessing such risks.

- Developing evidence to support quantification of human-caused and natural rare event risk requires not only statistical analysis, but the forensic sciences as well. It is an interdisciplinary exercise extensively involving both the physical and social sciences. Other disciplines such as law and economics may also be part of the process.

- Modeling uncertainties is a critical element of quantitative risk analyses, and there are multiple methods that can be applied to a given problem to illuminate the boundaries of our knowledge. Contemporary practices indicate increasing confidence in Bayesian methods for quantifying contributors to risk, including the associated uncertainties.

- The value of research in this area for policy- and decision-makers extends well beyond specific quantitative results, as it lays out tentative boundaries as to what can happen, and what can be done to prevent, delay, or mitigate those events. The research explices what is known and unknown, and opens windows onto “unknown unknowns.”

- An overarchingly point of discussion was not just the need to quantify the risk of the global events themselves, but equally important to provide a well-defined and risk-based roadmap of the most beneficial recovery actions. This same idea equally applies to regional disasters such as occurred in the USA, namely the 2017 hurricanes Harvey and Maria, the 2018 hurricane Michael, and the 2018 California wild fires. The main benefit of such an approach is making the contributing factors to the disasters much more transparent to facilitate the most effective recovery options.
• Categorization and prioritization of risks, and even defining the measure of consequences require additional discussion and elucidation. Artificial dichotomies, such as natural vs. manmade events, can foreclose the examination of the risk of synergistic or cascading events; prioritization based solely on consequences may not result in practical recommendations if an event is not preventable or cannot be mitigated; and, consequences measured in actual deaths, for example, ignore the consequences of lost future generations. A comprehensive inventory of consequences to be avoided can serve as the basis for an international, multi-cultural prioritization of risks.

• Additional value is found in applying quantitative risk analysis to assess changes from a baseline risk to a modified risk, including the risk of the transition itself; i.e., even if the risk of an event cannot be narrowly cabined, the change in risk between the baseline state to the modified state may be quantifiable, as well as the risks attendant to the transition period.

• A special case of nuclear deployment (a potential retaliatory deployment based on a false alarm of a first strike) was presented as a case study in quantification of these complex risks, with a demonstration that quantification can be achieved.
III PRESENTER PAPERS

Six presentations were made as a basis for extensive discussion in a workshop format. All of the presentations are summarized and evaluated in the Summary and Evaluation section presented earlier. Five of the presentations were based on papers included in these proceedings. Unwin’s presentation did not involve a paper, but it is summarized in this section.

As noted earlier, the first three presentations were coordinated to provide a framework for discussion on how existing methods of quantitative risk assessment would fare in quantifying global catastrophic risks. Full papers are provided for these three presentations.

III.1 REFLECTIONS ON THE RISK ANALYSIS OF NUCLEAR WAR – SETH BAUM

ABSTRACT

This paper presents reflections on the use of risk analysis for understanding and informing policy decisions about nuclear war. A quantitative evaluation of risk arguably should be central to many important nuclear weapons decisions, such as disarmament and launch alert status, because these decisions involve tradeoffs between different risks. However, nuclear war is a difficult risk to analyze, little effort has been made to analyze it, and nuclear war policy decisions have made little use of risk analysis. The paper demonstrates this via a detailed review of the nuclear war risk literature, a summary of a new model for nuclear war risk produced by the Global Catastrophic Risk Institute, and a discussion of the use of risk analysis in nuclear war decision-making. Despite the challenges, there are significant opportunities for progress on both the analysis and the decision-making. The paper finds that, at this time, the limiting factor is mainly the use of risk analysis for decision-making, such that people working on nuclear war risk should emphasize outreach to decision-makers. The paper’s discussion is of relevance for guiding efforts to understand and reduce nuclear war risk, and is likewise applicable to many other risks, especially other global catastrophic risks.
Consider this question: Would the world be safer with or without nuclear weapons? This is one of the most vital questions for the international community in its ongoing debate over nuclear disarmament. This is also a question for which a quantitative risk perspective is highly relevant. Unfortunately, the international community has not made much use of risk analysis in its discussion of this question, or of the numerous other important nuclear weapons policy questions for which risk analysis could also be highly relevant.

The aim of this paper is to reflect on the risk analysis of nuclear war, both in terms of its intellectual substance and its role in policy discussions and decisions. The paper draws on my own experience analyzing nuclear war risk and policy, my experience in and observations of international policy discussions about nuclear war, as well as the broader literature and other discussions of the topic.

A few central points can be made. First, nuclear war is a difficult risk to analyze, because a nuclear war would be complex and largely unprecedented. Second, there has been little effort to analyze nuclear war risk, though the efforts to date have made some meaningful progress. Third, nuclear war policy discussions have generally not sought input from risk analysis. These three points are interrelated, and they combine to paint a picture of a nuclear war policy debate that is not as well informed by risk analysis as it could be.

To illustrate this, let us revisit the question of whether the world would be safer with or without nuclear weapons. This is a key factor in the policy debate about nuclear disarmament. A large segment of the international community is concerned that the nuclear-armed states are not disarming rapidly enough. Their concern has translated into the new Treaty on the Prohibition of Nuclear Weapons. Meanwhile, other portions of the international

---

3 See in particular Barrett et al. (2013), Baum (2015a; 2015b), Baum and Barrett (2018), and Baum et al. (2015; 2018).
community, primarily the nuclear-armed countries and their close allies, have argued against rapid nuclear disarmament.

Both sides of the debate are talking about risk, but they are talking about different aspects of risk. The rapid disarmament side argues that nuclear weapons increase the severity of conflict by emphasizing the large humanitarian consequences of nuclear weapons use (e.g., Fihn 2013). The non-rapid disarmament side argues that nuclear weapons decrease the frequency of major war by emphasizing nuclear deterrence (e.g., Mies 2012). Both sides’ arguments could be correct, but that would not resolve which side has the better disarmament policy.

A full risk analysis would consider the effect of nuclear weapons on both the frequency and the severity of war (or, more generally, on violent conflict). Figure 1 sketches what such a risk analysis might look like, assuming that nuclear weapons do indeed increase the severity and decrease the frequency of war. An important question is whether the decrease in frequency is enough to justify the increase in severity. If it is, then a case could be made for gradual disarmament or the permanent retention of nuclear weapons. Essentially, this would be to say that, in risk terms, the world would be safer with nuclear weapons than without.

4 The non-rapid disarmament side includes people arguing for gradual disarmament, no disarmament, and increased nuclear weapons proliferation, all of whom often base their arguments in the efficacy of nuclear deterrence.

5 I briefly pursued a somewhat more detailed analysis of the risk of nuclear vs. non-nuclear war in Baum (2015c), concluding in favor of rapid disarmament, i.e. that the risk of non-nuclear war is smaller. My argument was based on the outsized severity of nuclear winter. However, at this time, I am not convinced that this analysis is correct. For example, it does not account for various prospects for the long-term trajectories of human civilization, a matter explored further in Baum et al. (2018b).
Figure 1. A possible sketch of the risk of major war with nuclear weapons (nuclear war) vs. without nuclear weapons (conventional war).

Risk need not be the only factor in nuclear disarmament policy. There can be other important factors, such as the cost of retaining or disarming nuclear weapons and the political preferences of various countries and other stakeholders. But risk would clearly seem to be an important factor. Risk analysis provides the means of answering the basic question of whether the world is safer with or without nuclear weapons. So we should expect to see significant demand from policy communities for risk analysis of both nuclear and conventional war. However, this has not been the case. Instead, policy communities have mostly just talked about aspects of the risk without any serious effort to analyze it.

It is important to note that the tradeoff shown in Figure 1 requires a quantitative risk analysis. The underlying question is whether nuclear weapons decrease the frequency of war more than they increase the severity. This is not a question that can be answered via qualitative characterization of the risk. This is important because, as this paper will show, quantifying this risk is not an easy task. Perhaps it will prove insurmountable, but there has not yet been enough effort to reach this conclusion.

There are other important policy questions that also demand a quantitative analysis of nuclear war risk. One concerns nuclear weapons launch alert status. The case for having nuclear weapons on high alert is based on the notion that doing so strengthens deterrence, because high-alert weapons can be launched more quickly, making it harder for the other side’s first-
strike attack to succeed. The case for having nuclear weapons on low alert is based on the possibility that high-alert weapons could more readily be launched accidentally or by rogue or unauthorized actors. Thus, launch alert policy faces a tradeoff between the risk of deterrence failure and the risk of accidental/unauthorized launch. Quantitative risk analysis is needed to evaluate this tradeoff.

There are also some more general policy questions that apply to a wide range of risks, including nuclear war. First, how much of a priority should nuclear war be? There are many issues that compete for attention, funding, and other scarce resources. Arguably, the larger the risk of nuclear war is, the more of a claim it has for these resources. Second, what are the best or most effective ways to reduce nuclear war risk? There are many actions that could potentially reduce the risk, from improving relations between nuclear-armed countries to developing resources to aid post-war survivors (Baum 2015a). Arguably, efforts to reduce nuclear war risk should focus on those actions that would cause the largest decrease in the risk.

It follows that a quantitative understanding of nuclear war risk is, or at least arguably should be, important for nuclear weapons policy. Therefore, the remainder of this paper focuses on the current state of quantitative analysis of nuclear war risk, the prospect for future progress, and what it all means for public policy and risk analysis research. Section 2 reviews prior literature. Section 3 presents an overview of the most detailed model currently available, that of Baum et al. (2018a) and Baum and Barrett (2018). Section 4 outlines an agenda for future research on nuclear war risk. Section 5 discusses the research and policy implications and concludes.

It should be noted that many of the issues raised here are not unique to nuclear war. To the contrary, many risks are difficult to analyze and/or have received little risk analysis attention by analysts or policy communities. This holds in particular for other global catastrophic risks. All global catastrophic risks face the same basic data challenge: no global catastrophe has ever destroyed modern global civilization. Many of the risks also face similar complexities. For example, Baum (2018) shows that
asteroid risk—one of the few global catastrophic risks that has been analyzed at length—has highly uncertain human consequences. A major point of uncertainty concerns how well survivors would fare in a post-catastrophe world. The same uncertainty surrounds all global catastrophes that leave some survivors.

### III.1.2 PRIOR LITERATURE

This section reviews literature that analyzes the probability and severity of nuclear war, with emphasis on attempts to quantify probability and severity. This section does not attempt to cover the literature that presents more qualitative discussion of nuclear weapons issues and frames the discussion in terms of risk, such as Lewis et al. (2014), Borrie et al. (2017), and Acton (2018). This literature is often insightful and of relevance for the study of nuclear war risk, but it is beyond the scope of this paper.

#### III.1.2.1 THE PROBABILITY OF NUCLEAR WAR

The earliest dedicated analyses of the probability of nuclear war available in the public record appear to be by Bernard Bereanu of the Centre of Mathematical Statistics in Bucharest (Bereanu 1981; 1982; 1983), and by Michael Intriligator and Dagobert Brito, economists at University of California, Los Angeles and Tulane University (Intriligator and Brito 1981). Prior to these publications, i.e. over the first several decades after the invention of nuclear weapons, the literature contains relatively limited discussion of the probability of nuclear war.

The Bereanu papers model the probability of nuclear war due to false alarms from early warning system technical glitches. The false alarms are assumed to occur at random intervals following a Poisson process. The model compares the time it takes to assess whether the alarm is true or false to the time until adversary missiles reach their targets. If the alarm cannot be resolved before targets would be hit, it is assumed that the side experiencing the alarm will launch its nuclear weapons. The papers further assume that the time for missiles to reach their targets will steadily decrease, due to progress in missile technology. Under these assumptions, nuclear war will eventually occur “with probability 1” (Bereanu 1983, p.49, paper abstract).
The papers propose negotiations on delivery systems to increase the time available to resolve alarms, and they propose nuclear disarmament as the policy needed to avoid nuclear war.

The various assumptions of the Bereanu model are made with little or no empirical justification and can readily be questioned. First, it is plausible that the rate of false alarms would tend to decrease over time due to improvements in warning system technology. Second, it is not a certainty that weapons would be launched if the false alarm is not resolved within the time for adversary weapons to reach their targets. Indeed, there are documented instances of military personnel and political leadership opting to not launch due to their own yet-unconfirmed suspicion of the alarm being false, such as the 1983 Stanslav Petrov incident and the 1995 Norwegian rocket incident. There are even some nuclear-armed states, such as China, that are believed to typically keep their nuclear weapons in a low alert state, such that the weapons cannot be fired under such short notice. Third, the time for adversary weapons to reach their targets would not necessarily continue to decrease. Indeed, missile speed has not changed substantially over the years.

Intriligator and Brito (1981) model the effect of nuclear proliferation on the probability of nuclear war. The model is a theoretical framework for evaluating the effect on the probability of war of additional states acquiring nuclear weapons. The paper does not attempt to quantify the effect, but instead presents qualitative discussion of likely trends. It considers the possibility of new nuclear-armed states initiating nuclear war as well as the possibility of them restraining nuclear war via deterrence and via potentially dominating the postwar order in the event that they are not party to the war. The discussion is largely theoretical and does not bring in significant empirical evidence to inform the analysis. A later publication (Brito and Intriligator 1996) proceeds along similar lines.

Marsh (1985) models the probability of the US launching nuclear weapons due to a false alarm of incoming attack. The study

---

6 These and other false alarm incidents are described in Baum et al. (2018a).
includes rich empirical detail on such matters as flight times of Soviet nuclear weapon delivery vehicles, US warning systems and decision structures, and the survivability of US nuclear forces. The study introduces a dataset of US false alarms during 1977-1983 that was subsequently used by Wallace et al. (1986) and Barrett et al. (2013). The data suggest one false alarm per year. For certain launch postures (in particular “launch under attack”), alarms from both radar and satellite are needed (“dual phenomenology”). Both alarms would need to occur during a narrow window corresponding to the flight time of the Soviet missile, which the study calculates to be 27 minutes for ICBMs and 9 minutes for SLBMs. That corresponds to a 0.000051 annual probability of two concurrent false alarms for ICBMs and a 0.000017 annual probability for SLBMs, or approximately once per 20,000 years for ICBMs or 60,000 years for SLBMs (p.68).

Paté-Cornell and Neu (1985) model the effect of US command, control, communications, and intelligence (C3I) systems on the probability of nuclear war between the US and the Soviet Union. C3I systems are supposed to provide national leadership with accurate information on whether the nation is under attack, and to faithfully relay any launch orders that the leadership decides to make. The study models an array of scenarios based on whether the Soviet Union has launched an attack, whether US C3I correctly assesses the presence or absence of an attack, whether or not the US national leadership decides to launch an attack based on the information that US C3I has provided it, and whether or not US C3I accurately relays launch orders to weapon operators. The model additionally explores different C3I configurations that can affect these various probabilities, such as launch on warning vs. launch on impact. Model parameters are quantified using illustrative point estimates based on the author’s judgments that are “deliberately arbitrary” and “bear no known relation to any realistic probabilities” (p.132). The study finds that the probability of inadvertent US nuclear launch due to C3I false alarm “dominates” the probability of US non-launch due to C3I failing to report an incoming attack (p.135), though given the allegedly arbitrary nature of the parameter estimates, this finding may have no real-world meaning.
Wallace et al. (1986) model the probability of the US launching nuclear weapons due to a false alarm of incoming attack. This study combines the 1977-1983 US false alarm data from Marsh (1985) with information on the time available to resolve false alarms. The underlying idea is that if an alarm cannot be resolved, then it may be interpreted as a real attack, thereby prompting launch in what is presumed to be a counterattack. The study calculates probabilities of false alarms occurring during major crises (such as the Cuban missile crisis), given the frequency of false alarms and the duration of a given crisis. However, the study falls short of calculating the probability of nuclear war because it does not consider the frequency of crises. Additionally, it does not rigorously quantify the probability of a false alarm prompting a nuclear weapon launch: it focuses on the time available to resolve alarms and not the human decision process of ordering launches.

Avenhaus et al. (1989) model the total probability of nuclear war. The paper explores how changes in the annual probability of nuclear war affect the long-term probability. The analysis is entirely theoretical and does not appear to have any basis in actual estimates of the probability of nuclear war. Instead, it is essentially just an inquiry into the mathematics of sequences of probabilities that happens to be framed in the context of nuclear war but could just as easily be for any ongoing probability.

Paté-Cornell and Fischbeck (1995) apply Bayesian probability theory to the interpretation of C3I signals by the US President. The study is not an analysis of the probability of nuclear war per se, but instead is an analysis of the probability of incoming attack that the President is likely to assign in the event that US C3I indicates an incoming attack. The study assumes that the President’s reasoning conforms to Bayesian probability theory. The study further assumes certain specifics about the President’s beliefs, for example that the President’s prior probability of attack follows a beta distribution. The study postulates that, with the Cold War having ended, the probability of incoming attack may be low relative to the probability of C3I false alarm, and that the President may overestimate the probability of incoming attack in the event of a C3I alarm unless the President’s thinking conforms with Bayesian probability theory.
Hellman (2008) models the probability of Russia-U.S. nuclear war that is caused by a crisis similar to the Cuban missile crisis. The model includes the frequency of events that could escalate into such crises and the conditional probabilities of escalation from initial event to crisis, from crisis to nuclear weapons launch, and from nuclear weapons launch to all-out nuclear war. The study uses historical data for initial event frequency and escalation to crisis. There is no historical data for escalation to nuclear weapons launch and all-out nuclear war, so the study uses ranges of probabilities. The study calculates the probability of this type of nuclear war as being in the range of $2 \times 10^{-4}$ to $5 \times 10^{-3}$ per year. Arguably, ranges of probabilities should have also been used for the first two parameters, which are also uncertain. Furthermore, this study uses a sparse information set to quantify its parameters, suggesting a significant amount of ongoing uncertainty.

Lundgren (2013) models the probability that nuclear war could have occurred during the Cold War. The model includes crises, false alarms, and conventional war escalating to nuclear war. The model uses a 21.3% probability for nuclear war via the Cuban missile crisis based on personal estimates of President Kennedy and his national security advisor, McGeorge Bundy. Many of the other probability estimates are the author’s personal estimates and are not easily assessed. The study also does not consider uncertainty in any of its parameter estimates. The study calculates a 61% chance that nuclear war could have occurred during the Cold War. Of course, it is now known with certainty that nuclear war did not occur during the Cold War, and present and future circumstances are different from the Cold War.

Barrett et al. (2013) model the probability of inadvertent Russia-U.S. nuclear war. As defined in this paper, inadvertent nuclear war occurs when one country misinterprets a false alarm as a nuclear attack by another country and launches nuclear weapons in what it mistakenly believes is a retaliation but is in fact the first strike. The study models the process of a false alarm making its way through the launch decision process. The study uses the Marsh (1985) false alarm data, assuming that the ongoing false alarm rate is consistent with this older dataset. The study also uses probability distributions for uncertain parameters. The study
calculates probability distributions for the rate of inadvertent Russia-U.S. nuclear war under two sets of assumptions. The distributions are quite wide, with the 5% to 95% ranges spanning from 0.0002 to 0.07 and from 0.00001 to 0.05 depending on the assumptions.

III.1.2.2 THE SEVERITY OF NUCLEAR WAR

The literature on the severity of nuclear war is more diffuse and more difficult to summarize. Thus, this section will only cover a few select highlights from this literature.

Perhaps the first analysis of the severity of nuclear war is Konopinski et al. (1946), a study completed as part of the Manhattan Project. This study examined the possibility of nuclear detonations igniting the atmosphere, resulting in global catastrophe. The study concluded that ignition was unlikely but did not definitively rule it out. The study was conducted prior to the Trinity test, which was the first-ever nuclear detonation. Had the study found ignition to be more likely, the Trinity test may not have proceeded.

An especially detailed study of the severity of nuclear war is Glasstone and Dolan (1977), published jointly by the US Departments of Defense and Energy. This study documents several physical effects of nuclear war: air blast, ground shock, thermal radiation, ionizing radiation, and electromagnetic pulse. The study presents the physics of these effects in considerable detail. It also covers secondary effects on built infrastructure and human bodies (i.e., medical effects). Other secondary effects, such as economic and political effects, are not covered. The study contains some quantitative analysis but does not seek to tabulate the net severity of nuclear detonations, nor does it consider the aggregate severity of nuclear war.

OTA (1979), a study by the US Office of Technology Assessment and commissioned by the US Senate Committee on Foreign Relations, assesses a range of consequences of nuclear war. It emphasizes that “the effects of a nuclear war that cannot be calculated are at least as important as those for which calculations are attempted” (p.3). This line specifically refers to the calculations of military planners, which focus on more
predictable consequences of nuclear war and exclude less predictable consequences such as social and economic disruption. The study analyzes four nuclear war scenarios: attacks on individual cities (using Detroit and Leningrad as examples), 10-missile attacks on oil refineries, counterforce attacks on ICBM silos, and all-out counterforce and countervalue attacks. It also considers the relative advantages and disadvantages of the US and Soviet political-economic systems for managing the aftermath of nuclear war.

Turco et al. (1983) presents the first scientific study of the global environmental consequences of nuclear war known as nuclear winter. This study focuses exclusively on environmental effects and does not seek to characterize impacts in human terms. Several studies have since examined the environmental consequences. For example, Robock et al. (2007) use more advanced climate models, finding less intense but more durable effects than Turco et al. (1983), while Reisner et al. (2018) use more advanced fire models, finding less smoke entering the stratosphere relative to previous studies.

Ehrlich et al. (1983) studies ecological consequences of nuclear war from ionizing radiation and various effects related to nuclear winter. The study anticipates massive ecological harms, including “the extinction of a major fraction of the plant and animal species on Earth”, and finds that human extinction “seems unlikely” but “cannot be ruled out” (p.1299). Loss of civilization in the Northern Hemisphere and possibly its entire population are seen as more likely possibilities. However, these possible human effects are not based on any careful analysis. Instead, the analysis focuses on general ecological effects, primarily to nonhuman species.

Cantor et al. (1989) is a rare extended inquiry into the economic consequences of nuclear war, produced by Oak Ridge National Laboratory on behalf of the U.S. Federal Emergency Management Agency. The study is framed generically in terms of social cataclysms and uses nuclear war as a central example. The study explores the possible forms of economic exchange in the aftermath of nuclear war. It considers possibilities such as the loss of property rights, the use of barter instead of money, the loss of trust in fiduciary authorities, and disruptions to transportation
and labor. The study draws on literatures from anthropology, economics, and sociology, and it reflects on the challenge of scientifically analyzing such an unprecedented event. It does not attempt to quantify the economic consequences of nuclear war, but nonetheless highlights why economic consequences can be an important factor to the overall severity of nuclear war.

Toon et al. (2007) assess a range of human and environmental consequences of nuclear war and nuclear terrorism. The study models human casualty and fatality rates as a function of distance from ground zero based on Hiroshima and Nagasaki data. Using this model, it presents calculations of casualties and fatalities from a similar (15KT) detonation in several nuclear terrorism and nuclear war scenarios. Calculations are presented as point estimates and do not account for uncertainty in the underlying model or the use of nuclear weapons with other yields. Human harms from ionizing radiation, including medical effects and territory abandonment, are described but not quantified. (There is quantification of some physical processes involving ionizing radiation.) Analysis of global environmental effects focuses on the amounts of soot produced and the corresponding effects on atmospheric chemistry; human harms are not considered.

EMP Commission (2008) presents a detailed analysis of the effects of electromagnetic pulse. The Commission was created by the 2001 US National Defense Authorization Act. The EMP Commission (2008) study presents an especially detailed analysis of the effects of nuclear war on civil infrastructure. The study focuses on the considerable effects of electromagnetic pulse on infrastructure, but much of it also applies to effects from other aspects of nuclear war. For example, disruptions to energy systems, telecommunications, food and water provision, and government functioning, all covered in EMP Commission (2008), can also occur from the direct damages from low-altitude nuclear detonations (blast, fire, etc.). The study describes the effects in detail but does not present aggregate quantifications of severity.

Robock (2010) surveys the scientific literature on nuclear winter and discusses some potential human harms. The study finds that “most of the world’s people are threatened with starvation following a full-scale [Russia-US] nuclear war” (p.424). The study
further states that “Although extinction of our species was not ruled out initial studies by biologists [such as Ehrlich et al. (1983)], it now seems that this would not take place. Especially in Australia and New Zealand, humans would have a better chance to survive.” It should be noted that the phrasing “a better chance to survive” continues to not rule out human extinction. Furthermore, the study focuses on environmental effects, not human consequences, and thus is arguably not well-positioned to evaluate prospects for human survival.

Helfand (2013) quantifies the human harms from nuclear winter in an India-Pakistan nuclear war scenario. The study finds that two billion people could be at risk of starvation. The study only claims that this number of people would be at risk of starvation, not that they in fact suffer or die from starvation. The analysis is based on crop modeling under nuclear winter climatic conditions and data on global food insecurity. The underlying idea is that nuclear winter would reduce food availability, which is especially worrisome for the present-day population that already faces food scarcity. While this underlying idea is likely to be robust, the two billion estimate is more suspect. Throughout the study, point estimates are used for highly uncertain parameters, and the uncertainty is seldom acknowledged. For example, the study states that “Even if agricultural markets continued to function normally, 215 million people would be added to the rolls of the malnourished over the course of a decade” (p.2, emphasis added), suggesting that it is known that this exact number of people would become malnourished under the described scenario. The study also identifies, but does not attempt to quantify, two additional effects that could factor significantly in the total severity: the possibility of food scarcity to cause or worsen disease outbreaks and violent conflicts. These possibilities speak to the considerable uncertainty pervading attempts to quantify the severity of nuclear winter and other impacts of nuclear war.

Fihn (2013) exemplifies the studies of the consequences of nuclear war undertaken by the portion of the international community that seeks more rapid nuclear disarmament. This publication presents discussions of a range of consequences: medical, environmental, agricultural, economic, and political. It also includes case studies of several actual and potential nuclear
detonations. It includes technical and quantitative analyses as well as attention to the human side, for example noting “The vast majority of injured people would die alone without so much as a human hand or voice to comfort them and without any relief for their agonising pain” (p.23). The publication concludes that the international community should “declare both the use and the possession of nuclear weapons as unacceptable, as there is no legitimate situation in which the impact of the use of a nuclear weapon can be justified” (p.100), though it considers neither the probability of nuclear weapon use nor the risk of violence in a world without nuclear weapons.

Frankel et al. (2013) evaluate the state of knowledge about the physical consequences of nuclear detonations. This study reviews data from test detonations and extrapolations based on underlying physical mechanisms. It covers a range of effects, with emphasis on surprises such as electromagnetic pulse, ozone depletion, and nuclear winter. A central theme is that the physical consequences remain uncertain, especially for large-scale nuclear war. Echoing OTA (1979), this study expresses concern about a tendency to “underestimate consequences by concentrating on selected physical phenomena that cause calculable damage to targets of interest to military planners” (p.31).

III.1.2.3 OVERARCHING THEMES IN THE PRIOR LITERATURE

In consideration of the prior literature for both the probability and severity of nuclear war, several themes emerge. First, the literature has many gaps. Probability studies tend to focus on specific nuclear war scenarios, while the one study attempting to cover the total probability (Avenhaus et al. 1989) has significant methodological limitations. Impacts studies tend to focus on the more readily quantifiable harms, especially physical and environmental effects and medical harms from direct exposure to nuclear detonations, while studies that cover other harms tend not to quantify the aggregate severity from these harms (e.g., Cantor et al. 1989; EMP Commission 2008).

Second, the risk of nuclear war is not easy to quantify. Attempts to quantify the probability are generally incomplete and easy to
poke holes in. Studies of the impacts often do not even attempt quantification, and when they do, it is only for a few relatively simple portions of the impacts. For both probability and impacts, quantification is challenged by pervasive complexities and a lack of empirical data.

Third, it can be argued that the literature is not particularly extensive. Section 2.1 has a more-or-less comprehensive survey of the probability literature; this is really not much for a topic that has been studied for almost 40 years. Section 2.2 does not have a comprehensive survey of the impacts literature, but the vast majority of this literature, including most of the studies in Section 2.2, is oriented toward general descriptions of impacts and not toward risk analysis. Analysts seeking to answer basic questions—such as what the total risk of nuclear war is and how the risk could be affected by various policies—have little in the way of resources to draw on. While the risk is difficult to quantify, quite a lot more could be done to that end.

III.1.3 THE GCRI MODEL

In recognition of the shortcomings of the prior literature, the GCRI model covers the entire risk. The model is summarized in Figure 2.

![Figure 2](image)

**Figure 2.** Summary of the GCRI model for the risk of nuclear war.

Figure 2 resembles a “bowtie” model commonly used in risk analysis.\(^7\) The element on the left depicts the set of causes that

---

\(^7\) Bowtie models typically put threats in the left bow, consequences in the right bow, and the hazard in the middle. For discussion, see e.g. [https://www.cgerisk.com/knowledgebase/The_bowtie_method](https://www.cgerisk.com/knowledgebase/The_bowtie_method).
can form nuclear war scenarios; this is used to model the probability of nuclear war. The element in the center depicts the details of the nuclear war itself, such as which countries participate, how many nuclear weapons are detonated, and which targets are struck by the nuclear detonations. The element on the right depicts the impacts of nuclear war, with each detonation branching out into a range of impacts. Work to date has focused on developing the model structure for probability and impacts. The model structure for the details of the nuclear war itself remains to be completed, as does quantifying parameters in each of these models. What follows is a summary of the probability and impacts models. Full detail can be found in Baum et al. (2018a) and Baum and Barrett (2018).

III.1.3.1 THE PROBABILITY MODEL

The probability model has two main branches, covering two major types of nuclear war scenarios: intentional first strike and first strike intended to be retaliation. Intentional first strike occurs when a country correctly believes that it is not under nuclear attack and makes the decision to initiate nuclear war. First strike intended to be retaliation occurs when a country incorrectly believes that it is under nuclear attack when in fact it is not under nuclear attack, and then makes the decision to launch nuclear weapons in what it believes is retaliation but is in fact the first strike. The two model branches are shown in Figure 3.

**Figure 3.** Two main branches of the GCRI nuclear war probability model.

The distinction between the two main branches is central to at least one important policy decision, concerning nuclear weapon launch alert status. When weapons are at high alert, they can be
launched more rapidly in response to warnings of incoming attacks. This is believed to strengthen nuclear deterrence by making it harder for one side to destroy the other side’s nuclear weapons in a surprise first-strike attack. However, high alert is also believed to increase the potential for accidental or unauthorized launch of nuclear weapons because weapons on high alert are easier to launch. This is essentially a tradeoff between the probability of intentional first strike and the probability of first strike intended to be retaliation: higher alert status could decrease the probability of intentional first strike (by strengthening deterrence) and increase the probability of first strike intended to be retaliation (by making it easier to launch nuclear weapons based on faulty information).

The model in Figure 3 can be translated into a probability equation in order to assess the probability of nuclear war and inform decisions such as launch alert status. The equation can be written is as follows:

\[ \lambda_{NW} = \lambda_T \cdot P_T + \lambda_F \cdot P_F \]  

(1)

In Equation 1, \( \lambda_{NW} \) is the annualized rate of nuclear war; \( \lambda_T \) is the annualized rate of events that could provoke intentional first strike; \( P_T \) is the conditional probability that one such event prompts an intentional first strike; \( \lambda_F \) is the annualized rate of a false belief of being under nuclear attack; and \( P_F \) is the conditional probability that the false threat prompts a first strike that is intended as retaliation. Annualized rates correspond to the probability of nuclear war occurring in a one-year interval; for more on the relation between rates and probabilities, see Baum et al. (2018a).

In terms of Equation 4, the launch alert policy decision can be conceptualized as an optimization problem in which the launch alert level is selected to minimize \( \lambda_{NW} \). Higher launch alert levels may decrease \( P_T \) by persuading countries that their intentional first strikes will be met with devastating retaliation. Higher launch alert levels may also decrease \( \lambda_T \) by motivating countries to avoid the sorts of conflicts and crises that could provoke intentional first strike. On the other hand, higher launch alert levels may increase \( P_F \) by making it easier for first strike attacks to be initiated. Higher
launch alert levels may also increase $\lambda_F$ by lowering the threshold for a country to believe it is under nuclear attack. Insofar as launch alert level does indeed pose a tradeoff between $(\lambda_T \times P_T)$ and $(\lambda_F \times P_F)$, it should be of considerable policy interest to assess which launch alert level minimizes $\lambda_{NW}$. This is one potential application of the probability model.

The rest of the probability model consists of details for each of the two main model branches. Figure 4 presents detail of the top branch. It includes two main pathways to intentional nuclear war: intentional escalation and inadvertent escalation. Intentional escalation occurs when one country’s leadership makes the decision to act in a way that prompts another country’s leadership to respond with a nuclear attack. Inadvertent escalation occurs when one the escalatory actions occur without any decisions from national leadership, such as in the “fog of war” (Posen 1991). Finally, both intentional and inadvertent escalation can occur during a conventional direct war (such as World War II), a conventional proxy war (such as the Vietnam War), or a non-war crisis (such as the Cuban missile crisis).

![Figure 4. Top branch of the GCRI nuclear war probability model.](image)

Figure 4 presents detail of the bottom branch. It includes two main pathways to nuclear war from a first strike intended as retaliation, based on two types of events that could cause a country to mistakenly believe it is under nuclear attack: non-war nuclear detonations and false alarms. Non-war nuclear detonations are any detonation that is not intended as an attack by the authorized leadership of another country. Non-war nuclear detonations include unauthorized detonations (such as by rogue actors within a nuclear-armed country or by terrorists who commandeer a country’s nuclear weapons), detonations of nonstate nuclear
weapons (if terrorists or other nonstate actors can build their own nuclear weapons), and accidental detonation of either domestic (a country’s own) nuclear weapons or foreign (another country’s) nuclear weapons. False alarms include events that look like nuclear attacks, including military exercises and nonmilitary events (such as scientific rocket launches), and monitoring system mistakes, including mistakes due to human error and glitches in monitoring system technology.

Figure 5. Bottom branch of the GCRI nuclear war probability model.

The probability model details can offer a wealth of valuable information, especially if model parameters can be quantified. For example, the probability of inadvertent escalation could inform decisions about how aggressive to be in conventional wars and crises. Some nuclear threats are often treated as “the threat that leaves something to chance” (Schelling 1960); which threats are made (and how) could be informed by (among other factors) exactly what that chance is. As another example, the various probabilities of different types of false alarms could inform decisions for how to allocate resources to reduce the probability of false alarm.
As a starting point for quantifying probability model parameters, Baum et al. (2018a) compile a dataset of 60 historical incidents that may have threatened to become a nuclear war. The dataset includes the one actual nuclear war (World War II) and 59 events that went partway towards nuclear war (such as the Cuban missile crisis). Further work is needed to assess how close each incident came to nuclear war and the ongoing probabilities of similar incidents, both of which are important steps toward quantifying the probability of various types of nuclear war.

Quantifying how close each incident came to nuclear war would be a subtly challenging endeavor. The incidents are all prone to historical interpretation. Indeed, analysts disagree on how close they were to nuclear war—for example, Lewis et al. (2014) consider them to have come pretty close, while Tertrais (2017) disagrees. One valuable contribution would be to analyze this debate in consideration of the broader study of the interpretation of near-miss events in risk analysis (e.g., Dillon et al. 2014).

III.1.3.2 THE IMPACTS MODEL

The probability model has five main branches, covering five major types of impacts of nuclear detonations: thermal radiation, blast, ionizing radiation, electromagnetic pulse, and human perceptions. The first four branches are widely documented, such as in the literature surveyed in Section 2.2. The human perceptions branch covers the social, political, and cultural reactions that humans can have to nuclear detonations. For example, the detonations in Hiroshima and Nagasaki may have shifted attitudes toward warfare and military conduct in ways that heavily structured the Cold War and other adversarial relationships. The five model branches are shown in Figure 6.
Figure 6. Five main branches of the GCRI nuclear war impacts model.

The rest of the impacts model is substantially more complex than the probability model, and likewise cannot be shown in full in this paper. (It is contained in full in Baum and Barrett 2018). Instead, this section will present select portions of the model in order to provide a general impression of it and to discuss some implications for research and policy.

A central feature of the impacts model is the use of modules that repeat in multiple model branches and interconnect with each other. The modules are essentially equivalent to modules or objects in object-oriented computer programming. Quantitative forms of the impacts model could likewise be implemented with such programming. As with other applications of object-oriented programming, the model’s use of modules is done because some types of impacts recur in different parts of the model. The model contains 15 modules: fire, blocked sunlight, damage to infrastructure, water supply disruption, agriculture disruption, food insecurity, healthcare disruption, infectious disease, transportation disruption, transportation systems disruption, energy supply disruption, satellite disruption, telecommunications disruption, shifted norms, and general malfunction of society.

To illustrate the modular nature of the module, the section will now present some model detail related to nuclear winter, which is an especially important potential impact of nuclear war. Nuclear winter comes largely from thermal radiation causing fire, which blocks sunlight. Figure 7 shows the model's module for blocked
sunlight. There are two main sets of impacts: agriculture disruption and changes to solar, wind, and hydro energy. As shown in Figure 8, agriculture disruption can cause changes to food consumption and a reduction in greenhouse gas emissions, the latter due to the large amount of greenhouse gas emissions currently produced by agriculture. As shown in Figure 9, the shifts in energy from blocked sunlight can disrupt energy supplies and can increase greenhouse gas emissions, as declines in renewable energy prompt increases in fossil fuel energy. Finally, Figure 10 shows the food insecurity module, whose impacts include the general malfunction of society due to the loss of labor and the outbreak of infectious diseases.

**Figure 7.** The blocked sunlight module.

**Figure 8.** The agriculture disruption module.

**Figure 9.** Model detail for shifts in energy from blocked sunlight.
Figures 7-10 show that the impacts of nuclear war could in turn affect at least two other global catastrophic risks: climate change and infectious diseases. The connection to infectious diseases has briefly been identified in previous literature (e.g., Helfand 2013) but has not been explored in significant detail, even though this is potentially one of the largest impacts of nuclear war. For comparison, the 1918 “Spanish” flu outbreak was arguably the largest impact of World War I. (The outbreak killed many more people than the war itself and may have been substantially milder if the war had not occurred.) The connection to climate change has not been discussed to any significant extent in prior literature and could also be an important factor.

The importance of including these indirect impacts of nuclear war is consistent with a theme in some of the literature surveyed in Section 2.2, especially OTA (1979), Cantor et al. (1989), and Frankel et al. (2013). Instead of starting by modeling the portions of the impacts that are easiest to model, this model starts by identifying the full range of potential impacts. Climate change and infectious disease are examples of impacts that are not included in most studies of nuclear war impacts and are not easy to model but could be major factors in the total severity. Future research is needed to quantify the severity of each of the various impacts.

**III.1.4 NUCLEAR WAR DECISION-MAKING**

While the analysis of nuclear war risk slowly chugs along, a variety of important decisions related to nuclear war are continually being made. Many of these are informed at least in part by aspects of nuclear war risk, such as disarmament advocates’ concern for the impacts of nuclear war and deterrence...
advocates’ concern for the probability of nuclear vs. non-nuclear war. However, this use of risk thinking has been somewhat superficial and largely disconnected from the various attempts to more carefully analyze the risk.

The disconnect is especially vivid in a passage from Paté-Cornell and Fischbeck (1995, p.31), which describes a hypothetical decision in which two Presidential advisors provide the President with advice based on two distinct nuclear war scenarios:

…they propose two different probabilities (0.2 and 0.3) that an attack is underway. Assume also that initially, having heard both arguments, the President gives each hypothesis a probability of 0.5 and that he has adopted as a measure of his own degree of belief the two estimates of the probability of attack conditional on the two hypotheses (0.2 and 0.3)… The mean prior probability that an attack is underway is therefore 0.25.

The analysis in this passage is logically coherent, but it has no apparent connection to actual US Presidential decision-making. While it is not the place of the current paper to speculate on the nature of US Presidential decision-making, it may nonetheless be plausible that, for at least some US Presidents, the passage does not describe how they think. Indeed, the US voters (or, more precisely, the US electoral college) have yet to elect a President with the sort of formal risk analysis background one may need to think in the terms described in the passage.

This passage is indicative of what I believe to be a wider tendency of risk analysts assuming, or at least hoping, that decision-makers share their analytical and ethical perspective, such that the risk analysts’ analysis could strongly inform important policy decisions. This phenomenon is hardly unique to nuclear war, but nuclear war offers an especially compelling case, given that many of the essential decisions must be made by the US President, the heads of other nuclear-armed states, and other top government officials. To be sure, there are a wide range of decisions to be made by a wide range of people that can affect the risk of nuclear
war, including decisions by ordinary citizens, technical experts, and other people from outside the top echelons of government (Baum 2015a). But these other populations are typically not seeking input from risk analysis either.

This circumstance may merit a more pragmatic approach by risk analysts. As elaborated in Baum (2015d) and Baum and Barrett (2017), one approach involves analysts and other like-minded colleagues reaching out to decision-makers in order to learn the decision-makers’ perspectives and opportunities and formulate policy ideas that make sense to them. Such an approach can be successful, but it places a considerable burden on analysts and their colleagues. Furthermore, any success could be transient, requiring analysts and colleagues to do customized outreach and policy formulation for each decision, with decision-makers otherwise poised to proceed in their own particular ways.

An alternative approach may be to advocate for a risk perspective among decision-makers. The aim would be to create durable interest in treating risk as an important factor in decision-making and likewise in the use of careful risk analysis. The case for risk in nuclear war decision-making arguably should be compelling, for reasons described throughout this paper, in particular for the central role that risk can play in policy decisions like nuclear disarmament and launch alert status. In my own brief experience engaging with nuclear weapons policy communities, I have seen enough interest in the risk perspective to believe there should be more effort to promote the risk perspective within these communities.

One important and very immediate decision is on whether analysts and their colleagues focus their effort on additional analysis or on outreach to policy communities. At this time, I believe the main limiting factor is interest from policy communities, and that the focus should therefore be on outreach. There has been significant progress in the research literature, including both the literature surveyed in Section 2 and the GCRI model outlined in Section 3. Yes, this research has major limitations, such that it cannot yet provide detailed guidance to policy decisions. Likewise, additional research progress can make risk more useful and compelling to decision-makers. But the
policy debate is not up to speed on even the limited existing literature, or on a more basic risk perspective. Until policy communities have a greater interest in nuclear war risk analysis, and greater sophistication in their use of it, there is little reason to continue producing nuclear war risk analyses. Furthermore, if they do become more interested in nuclear war risk analysis, then they may also create more institutional support for the analysis, which could accelerate research progress. One reason for further nuclear war risk analysis is to develop risk-reduction solutions that appeal to policy communities regardless of their interest in risk, as described above and in Baum (2015d) and Baum and Barrett (2017), but even this approach may be limited mainly by interest from policy communities. Thus, the case for more outreach would appear to be robust.

III.1.5 CONCLUSION

Nuclear war is an important risk. Analysis of the risk, including quantifying its probability and severity, could and arguably should yield valuable information for many important decisions that can affect the risk. Nuclear disarmament and nuclear weapon launch alerts status are two examples of decisions that pose tradeoffs between different risks and therefore benefit from the capacity to quantitatively evaluate the tradeoffs.

However, seven decades of scholarship on nuclear war risk, from the Manhattan Project study of Konopinski et al. (1946) to the GCRI model of Baum et al. (2018a) and Baum and Barrett (2018), has largely failed to inform these decisions. There are several reasons for this. Nuclear war is a difficult risk to quantify. The body of research remains fairly small. And decision-makers have generally not sought out risk analysis to inform their decisions.

There is no escaping the fact that nuclear war is a difficult risk to quantify, but progress can be made on both the research and the decision-making. At this time, it appears that the limiting factor is in the use of risk analysis in decision-making, such that analysts and their colleagues should focus on outreach to decision-makers instead of further research. However, further research can still be helpful, including to demonstrate the usefulness of risk analysis to decision-makers.
Nuclear war is not the only risk that faces these challenges. Indeed, all of the global catastrophic risks do to varying extents, as do many other risks. Nuclear war is thus both an important risk in its own right and also a valuable case for analyzing and managing risks, especially global catastrophic risks. Given the very high stakes, it is important that we continue to try to get the details right.

ACKNOWLEDGMENTS

Tony Barrett and Robert de Neufville provided helpful comments on an earlier draft of this paper. The paper also benefited from discussions with John Garrick and from the audience at the September 2018 workshop Quantifying Global Catastrophic Risks, hosted by the UCLA Garrick Institute for the Risk Sciences. All remaining errors are the author’s alone. The views of the paper are the author’s and not necessarily those of the Global Catastrophic Risk Institute.

REFERENCES


III.2 PRINCIPLES AND PRACTICES FOR QUANTIFYING GLOBAL CATASTROPHIC RISKS

– B. JOHN GARRICK

### III.2.1 INTRODUCTION

The purpose of this presentation is to present at a high-level principles and concepts of quantitative risk assessment applicable to quantifying global catastrophic risks. The Part I
presentation by Seth Baum set the stage for current efforts on quantifying the global risk of nuclear war. Part III, presented by Mihai Diaconeasa and associates, extend the model of Part I by applying the Part II principles and practices to a component of the nuclear war risk, namely the extremely feared global risk of an unintentional initiation of a nuclear war.

Global catastrophic risks, referred to simply as “global risks,” are risks of events that could significantly harm or even destroy human civilization at the global scale (Hempsell, 2004; Baum and Barrett, 2017). The major difference from much of our previous experience of quantifying risk is with respect to the form of the evidence supporting the risk assessments. For many global risks, the challenge is much greater than for facility and selected natural event risks because of the absence of direct experience. Thus, considerable emphasis is on the search for evidence of extremely rare events and especially those events for which there has been no direct human experience. We are being guided in our evaluation by the following questions:

1. Are existing methods of quantifying facility and natural risks extendable to the assessment of global risks?
2. To what extent do decision making processes impact global risks?\(^8\)
3. How can something be quantified that has never been directly experienced by human beings?

The task of quantifying global risks is considered by many eminent physical and social science scholars as essentially impossible. Consider the JASON\(^9\) report prepared in 2008 (MITRE/JASON, 2009). Here is what they reported concerning the assessment of events that have never occurred. “… it is

\(^8\) While many believe that human “intent” cannot be quantified, the decision making process can be analyzed for weaknesses and vulnerabilities and have an impact on human actions. This is particularly relevant to the global risk of an unintentional nuclear war. Mistakes and system failures can and have occurred. Question 2 is primarily addressed in the follow-on paper on the global risk of an unintentional nuclear war.

\(^9\) JASON is an independent group of distinguished scientists which advises the United States GOVERNMENT.
simply not possible to validate (evaluate) predictive models of rare events that have not occurred…” They were referring explicitly to a weapons of mass destruction terrorism event. Their finding certainly doesn’t provide much encouragement, for example of being able to quantify the risk of a nuclear war, the example chosen for this paper. The most complicating factor pointed out by the JASON study is “that rare event assessment is largely a question of human behavior, in the domain of the social sciences, and predictive social sciences models pose even greater challenges than predictive models in the physical sciences.” The global risk expert Milan M. Ćirković was more emphatic, “Very destructive events completely destroy predictability!”

So, where does this leave us? It leaves us with the challenge to maximize the knowledge about a global risk based on the evidence and traces of the events that do in fact exist or can be developed. A reasonable surrogate is to quantify the likelihood of the risk in the form of the frequency of the event. Frequencies are generally easier to develop, but of course there is uncertainty about the frequency and this is where probability plays a role. Processing the evidence according to Bayes theorem enables developing a probability distribution of the frequency of the event. This is not a “prediction” in the sense of a direct probability of the event, but it might be considered the next best thing to it and can result in a major increase in knowledge about the event. It does require some adjustments of algorithms and technical terms. Thus, our approach will be as follows:

- Define what we mean by “quantitative risk assessment” (QRA)
- Present an overarching framework for assessing and quantifying risk
- Define and interpret what we mean by such terms as “quantification,” “probability,” “evidence,” and “risk measure”
- Highlight the QRA methodology and form of the results
- Propose the example of quantifying the global risk of an unintentional nuclear war
III.2.2 QUANTITATIVE RISK ASSESSMENT

Quantitative risk assessment is a process of probabilistic evidential and inferential analysis of the response of events, systems, or activities to different threats and challenges based on the fundamental rules of logic and plausible reasoning. It is a thought process for answering the three basic risk questions about events, systems, or activities of what can go wrong, what is the likelihood if it does go wrong, and what are the consequences. The methodology has been evolving since the 1970s driven mostly by probabilistic risk assessments (PRA) of nuclear power plants. Most nuclear power plants throughout the world (~450 plants) have some form of a PRA, some much more rigorous than others. The technology has spread to other fields resulting in a robust experience base for doing quantitative risk assessments of both natural and anthropogenic systems and activities.

The task of quantifying risks requires interdisciplinary teams made up of the appropriate subject matter experts. Experience indicates among the disciplines that may be involved are the physical and social sciences, environmental science and engineering, legal, economics, medicine, and others. Of course, quantitative or probabilistic risk assessment has evolved as a discipline by itself and its experts usually are the principal investigators of such assessments.

Assessing the risk of global events, especially those that have not been physically observed or recorded by human beings represents a major challenge. This challenge is made even more significant in that global risks represent paths to existential risk. There are both natural and anthropogenic examples of global risks. Natural global events that we know have happened, but not always with the benefit of direct observation by human beings, are super volcanoes and large asteroid impacts. Examples of anthropogenic catastrophic risks of concern are out of control technologies such as artificial intelligence, pandemics, climate change, and nuclear wars.
We introduce a framework for what we mean by risk and quantitative risk assessment. To be sure this is not the only framework for doing quantitative risk assessment, but it has been extensively applied world-wide with considerable success for several decades. The concept has been generally adopted by much of the quantitative risk assessment community, including the U.S. Nuclear Regulatory Commission.

The framework is known as the triplet definition of risk (Kaplan & Garrick, 1981) and it was conceived to apply to any type of risk, subject only to different boundary conditions. An essential property of the definition is that it be rigorous in its form. Its premise is when we ask the question “what is the risk,” we are really asking three questions: what can go wrong with the system of concern ($s_i$), how likely is it if it does go wrong ($L_i$), and what are the consequences ($x_i$).

Question 1 is answered by a structured set of scenarios ($s_i$), thus the framework is often referred to as the scenario approach to quantitative risk assessment. Question 2 requires the calculation of the “likelihoods” $L_i$ of each of the scenarios $s_i$. Each of the “likelihoods” $L_i$ can be expressed as a “frequency,” a “probability,” or a “probability of frequency.” Question 3 is answered by describing the “damage states” or “end states” $x_i$ resulting from the risk scenarios. Thus, the risk, $R$, is represented by a complete set of scenarios, likelihoods, and damage states (consequences) as follows:

$$R = \{(s_i, L_i, x_i)\}^c$$

where the inner bracket represents the “triplet set,” the outer bracket as the “set of,” and the subscript “c” meaning the “complete set.”

A key element in the process is knowing what constitutes the “as-planned” or “stable” scenario as a reference for what constitutes
an upset condition. How the triplet results are assembled into compact, coherent, and comprehensive measures of risk is well established (Garrick, 2008) and briefly discussed later. First, it is important to define and interpret some of the critical terms.

### III.2.4 INTERPRETATION OF CRITICAL TERMS

#### III.2.4.1 QUANTIFICATION

In rare event work, where risks are measured in terms of their likelihood, that is their probabilities, quantification is the rigorous linking of the probabilities to the supporting evidence. The risk is not a single number, but a range of numbers rigorously distributed probabilistically. Often we hear the expression, “there is just not enough evidence, that is, data,” to quantify the risks. Besides being a misinterpretation of what is meant by “quantification,” it is most often incorrect. The truth is analysts seldom make use of all the data that is available because it is hard work and sometimes costly to develop. Furthermore, data goes way beyond event rates and includes evidence based markers that may be geological or biological, which is the reason for using the term “evidence” over “data.” More on this point later.

#### III.2.4.2 PROBABILITY

The probability and statistics literature usually defines three probabilities: classical, relative frequency, and axiomatic. The classical definition is based on the ratio of a particular event and the space of all such events. Relative frequency probability is based on a limiting frequency of events. Axiomatic probability is based on a set of axioms. All three have their limitations. The first two require the outcome of the events to be equally likely and the number of trials required is vague. While the axiomatic definition is preferred by many, it too has its issues without some modification especially for the case of limited data and the representation of uncertainty, the principal contributor to the risk of rare events. A more flexible narrative of what is meant by probability enables addressing such issues as limited data, variable outcomes, and other frequently mentioned limitations such as the quantification of the uncertainties involved.
For rare event work where the outcomes are generally not equally likely and where uncertainty must be represented, the so-called classical definitions of probability are too limiting. Thus, the need is for a probability definition and interpretation more suited to more demanding conditions; conditions of limited data, high levels of uncertainty, and multiple outcomes of the events. The benefit is definitions that enable increased knowledge about the occurrence of events, even though there might be some deviation from traditional figures-of-merit for measuring risk.

The definition of probability with which there has been considerable success is an extension of the subjectivist view of probability offered by Jaynes (Jaynes, 2003). In particular, Jaynes notes, “a probability assignment is ‘subjective’ in the sense that it describes a state of knowledge rather than any property of the ‘real’ world, but is ‘objective’ in the sense that it is independent of the personality of the user. Two rational beings faced with the same total background of knowledge must assign the same probabilities.”

Taking the Jaynes interpretation a step farther, probability is synonymous with “credibility.” Probability is defined as the credibility of a hypothesis based on the totality of the supporting evidence. It obeys Bayes theorem and the fundamental postulates of probability and plausible reasoning. This is a contemporary view of probability that greatly enhances assessing the risk of rare events; events about which there is little or no human experience, but there is evidence of their occurrence.

III.2.4.3 EVIDENCE

Evidence as used here is any information that can impact the likelihood or probability of events. Evidence can be classified as either direct or indirect. Here, direct evidence is taken to be that observed and recorded by humans. Indirect evidence is evidence not directly observed by humans but consists of markers, traces, or signatures of events. Such markers can be biological, geotechnical, or in space inside or outside our solar system. Indirect evidence is the primary basis of most global catastrophic risks.
We choose the word “evidence” over “data” because of its broader meaning, at least in the minds of most people. While they may be considered the same in many respects, the words often stimulate different images in our minds; evidence often being viewed as a legal term with data being more associated with statistics. For example, to many of us in the physical and social science worlds, data brings up such images as observed failure rates and event rates, that is, the frequency of something. To be sure such information is indeed evidence. Of course we know that evidence goes much beyond actual observance of events by humans. After all we know there have been some 15 extinction events, 5 of which involved the extinction of 50% or more of all species. How do we know this?

Primary sources of indirect evidence are near-misses, identifiable threats and their impact should they occur, known precursors to global events, societal conditions and actions, and expert knowledge. Direct near-miss evidence greatly reduces the uncertainties in the event sequences and provides important insights on what has to happen for a particular global event to occur. For an issue such as the global risk of nuclear war, there is considerable near-miss evidence in the “direct” category. The result is direct evidence of known precursors to a global nuclear event. This information together with the political tensions and stress that might exist provides a backdrop for the behavioral scientists to estimate the actions that the decision makers might take—a critical factor on the matter of “intent.” For example, a primary source of evidence that the Cuban missile crises threat of 1962 might turn into a nuclear war was President Kennedy’s action of an evacuation order to the families of the White House staff. Additional evidence that a nuclear war could be imminent was Defense Secretary McNamara’s fear that he would not live out the week (Hellman, 2008).

One major source of indirect evidence of rare events is expert knowledge. Many advancements have been made not only on the eliciting of information from expert groups, but the methods of processing the resulting information, especially with respect to natural disasters. The nuclear power field has advanced the
methods considerably. Expert elicitation has become very structured and disciplined in terms of putting together the expert teams, accounting for anthropic bias, phrasing the questions, and processing the information. Repeatability, consistency, and Bayesian inference have become major components of the process. An important and extremely effective strategy is to bypass the experts and directly consider the evidence on which the experts' knowledge is based (Kaplan, 1992). The result is less dependence on calibrating the expert and a means of greatly expanding the evidence base. Each expert processes their evidence a certain way, also an important source of information.

Clearly, the resources for developing and evaluating evidence involves both the physical and social sciences. Depending on the risk, many other disciplines may be required, including economics and legal. Logic and the uncertainty sciences are foundational to the development of evidence. Consider the role of “thought experiments.” For example, for a credible scenario, suppose it occurred N times such as 10, 100, 1000 or more rather than just once; what is the likelihood that “M” out of “N” times behavioral science indicates a high likelihood of the wrong decision being made under a prescribed set of conditions? While we can’t do the experiment, the exercise sharpens the question in a manner that subject matter experts are more able to respond. The uncertainties may be extensive, but evidence is generated and the evaluation basis is much more robust than had we not gone through such a process, including the actual quantification of the uncertainties.

Table 1 highlights types of evidence and risk indicators for selected global risks. Part of it came from the book edited by Bostrom and Ćirković (2008). Some liberties were taken to tailor the table to the specific topic of this paper as well as adding the Unintentional Nuclear War event (UNW). The distinguishing feature of the UNW event besides being anthropogenic is the very different nature of the evidence. There has never been in all time an unintentional nuclear war. Meanwhile, there is strong evidence that all the others on the list while never recorded or observed by

---

humans have actually occurred. The difference in the “hazard” evidence and the “risk” evidence for the UNW event is the actions that are being taken, that is, the dynamics of the activities associated with the hazards. A hazard is considered an agent that can potentially render damage only if certain events transpire.

Table 1. Evidence and Risk Indicators of Selected Global Risks

<table>
<thead>
<tr>
<th>EVENT</th>
<th>EVIDENCE OF HAZARD</th>
<th>EVIDENCE OF RISK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earth Impacts</td>
<td>Number and distribution in space of asteroids, comets, and space debris</td>
<td>Number and distribution of impact craters and debris, and shock glasses</td>
</tr>
<tr>
<td>Super-Volcanism</td>
<td>Number and distribution of geophysical “hot spots,” and their activity</td>
<td>Number and distribution of calderas, volcanic ash, and ice cores</td>
</tr>
<tr>
<td>Supernovae and Gamma-Ray Bursts</td>
<td>Distribution of progenitors and their motions near the solar system</td>
<td>Geochemical trace anomalies, distribution of remnants</td>
</tr>
<tr>
<td>Unintentional Nuclear War</td>
<td>Nuclear weapons, delivery systems, military and political posture, history of aggression</td>
<td>Nature of nuclear weapons activities, leadership instability, nuclear command center vulnerability</td>
</tr>
</tbody>
</table>

III.2.4.4 RISK MEASURE

Ideally, it would be desirable to measure rare event risk in the form of a direct probability of the event. Of course, in principle, the probability of any event can be calculated regardless of the strength of the evidence if uncertainty is taken into account as there is always evidence. For example, not experiencing an event over a prescribed period of time is certainly evidence about that event. The problem with choosing probability in a time space is the difficulty of characterizing the time at which rare events occur, that is, the difficulty of developing time data. In most cases it is easier to develop data on occurrence intervals that is frequency
data, which has a time element to it, for example, when using nuclear and other dating methods on the byproducts of the event.

The occurrence and reoccurrence of rare events is generally supported with geologic and biologic markers. If the global event of interest is a supernova that could impact the earth, then the markers of interest may be remnants of the explosion inside or outside our solar system. Of course, there is uncertainty in the occurrence intervals, that is, the frequencies, which leads us to using Bayes theorem to construct a probability distribution of the frequency of occurrence based on the totality of the evidence. This is discussed later and is referred to as the “probability of frequency” representation of risk. This form of quantifying risk has been effective as an alternative to direct probabilities of the events. It does provide some insight into when such events occur. And while a “probability of frequency” measure may be a hedge from the actual probability that the event will occur, it is an example of doing the best that can be done, given an extremely limited amount of evidence.

III.2.5 QRA METHODOLOGY

A methodology of quantitative risk assessment has been evolving for almost 5 decades based on the triplet definition of risk. Event trees, fault trees and Bayesian networks (Figure 1) are all a part of what contemporary investigators have labeled the Hybrid Causal Logic methodology (Wang, 2007), a multi-layered structure that integrates various logic models into what are referred to as Bayesian networks. This allows the most appropriate modeling techniques to be applied in the different individual domains of the system or process and the environment. Additional causal information can be added by additional layers of modeling.

A Bayesian network (lower right hand corner of Figure 1) is a combination probabilistic and graphic representation of the variables and their causal dependencies. The variables are linked forming what is referred to as a directed acyclic graph (DAC) as shown. Bayesian networks provide the flexibility to indicate cause-effect relations that are ‘soft’ or uncertain such as the impact of the human-system interface. Bayesian networks are
very useful for modeling complex organizational and human factors. Such networks sometimes replace or complement event trees and fault trees as they provide some advantages of efficiency and the consideration of human interaction with the system. This combination of analytical tools allows the most appropriate modeling techniques to be applied in the different individual domains of the system or process and the environment. These are the tools for implementing the risk triplet methodology, that is, the scenario approach to quantitative risk assessment. The event trees are central to structuring the scenarios, the fault trees facilitate the quantification of the split fractions of the event tree and the Bayesian networks provides “soft” causal links and causal dependencies of variables in the event tree and fault tree models.

Figure 1. An Example of Multi-layered Hybrid of Event Tree, Fault Tree, and Bayesian Network

The core input to the triplet or scenario based approach to risk assessment is the scenarios that map initiating events to consequences. The primary logic diagram for structuring scenarios is the inductive logic diagram known as an “event tree.” The goal is the development of a so-called complete set of scenarios that can emanate from a particular upset condition, that is, a particular initiating event. Each initiating event has its own event tree. Consider Figure 2. The top events A, B, etc., are protective systems, procedures, or human actions intended to prevent system degradation due to the initiating event.
Figure 2. The Event Tree as a Scenario Quantification Tool

If the top event performs its intended function, the convention is that the sequence continues in a straight line towards a “success” end state. If the top event fails to do its intended function, then the path is downward to a new path to be impacted by subsequent top events until the sequence terminates into various success or damage states.

Each of the paths through the event tree constitutes a scenario that could result from a particular initiating event.

Figure 3. Quantification of the Scenarios using Bayes Theorem

Logic models such as fault trees and Bayesian networks are used to quantify the “split fractions” determining the likelihoods of the events in the scenarios (event sequences). The scenario is defined by the scenario Boolean equation which is transformed into a frequency equation as shown in Figures 2 and 3.

Figure 3 is a schematic of the process of quantifying the individual event sequences, that is, the scenarios. Convoluting the $P(\varphi)$ split fractions, that is, doing the necessary probability arithmetic propagates the uncertainties into the $P(\varphi)$ curve of the scenario, thus quantifying the scenario. This process is repeated for
different initiating events as necessary to develop a complete set of quantified scenarios. A complete set is the full set of scenarios having an impact on the risk being quantified.

Figures 4 and 5 show the most frequently used forms for representing the risk. Figure 4 shows the uncertainty in the frequency of the risk, often referred to as a $P(\phi)$ curve. Such a metric can be used to represent the risk of individual events, the total risk of a scenario, or the cumulative risk of all the scenarios leading to a particular consequence. For example, in nuclear plant risk work this is the usual form for representing core damage frequency.

![For A Specific Consequence](image)

**Figure 4.** Probability of Frequency Curve

Figure 5 is a more comprehensive representation of risk and is considered by many practitioners as the classical risk curve as it integrates the complete set of scenarios providing the frequency, $\phi$, as a function of consequence $X_i$ or greater. The probability, $P_i$, is the parameter of the model represented by a family of curves. This graphic representation is the “frequency of exceedance curve” or the “complementary cumulative distribution function.” For example, if $P_1$ were the 5th percentile and $P_3$ the 95th percentile, the 90% confidence interval would be that the damage level of $X_1$ or greater has a frequency of occurrence between $\phi_1$ and $\phi_2$.

Presenting the risk assessment results in the form of Figure 5 enables all of the scenarios contributing to the risk to be...
represented, including the propagation of their uncertainties to the total risk of the system being analyzed. But there is much more decision making information that falls out of the analysis. Among the results are the contributors to the risk and their rank order. Thus, the decision makers, that is the risk managers, have a quantitative basis for prioritizing the risk management process.

![Complementary Cumulative Distribution Function](image)

**Figure 5.** Complementary Cumulative Distribution Function

Properly presenting the risk measures, the contributors to risk, and their context is essential to making the right decisions on how to prevent catastrophes from happening or at least how to reduce and manage their consequences.

### III.2.6 CONCLUSION

Quantitative risk assessment, also known as probabilistic risk assessment, and probabilistic safety assessment, has been evolving for several decades. With this evolvement has been efforts to generalize the methods to apply to essentially any kind of risk including global catastrophic risks. Indeed, the methodology has been successfully applied to complex facility risk such as nuclear power plants for several decades and many other complex facilities and natural hazards. There is a high level of confidence that the methodology will apply equally well to global catastrophic risks when accounting for the differences in the search for evidence to support the assessment. Supporting this view is a follow-on paper, Part III of this presentation series, applying the methodology to one of the most feared current global risks, namely the inadvertent initiation of a nuclear war.
REFERENCES


ABSTRACT

The quantification of catastrophic events such as inadvertent nuclear war is a challenging task as the evidence to date is that no such event has occurred. It is important to gain an understanding of the nature of such events because their consequences could potentially impact the survival of humankind and their probability of occurrence is nonzero. This study uses quantitative risk assessment (QRA) methods to identify scenarios that could lead to inadvertent nuclear war, including initiation events and potential system or human failure mechanisms. A set of scenarios for inadvertent nuclear war initiated by a false detection of an ICBM launch is examined in detail starting with the potential sources of false alarms. The false alarm is then propagated through each level of decision making until it is either detected or a decision to launch a preemptive strike is made. The scenario model uses three levels of modeling: event sequence diagrams at the top level to capture the risk scenario context, fault trees to model the potential causes of system failures or human failure events, and Bayesian networks to further examine the factors that influence human behavior. The likelihood of inadvertent nuclear war, including uncertainties, is calculated using the Hybrid Causal Logic Analyzer (HCLA) with data from a wide range of open sources and processed using Bayesian estimation techniques.

III.3.1 INTRODUCTION

On several occasions during the cold war, the United States and Russia came dangerously close to nuclear conflict. While all of these events were made more tense by the underlying pressure from the world affairs, several were initiated not by deliberate escalation, but by human error, miscommunications, system false alarms, and other inadvertent causes. It is a known fact that such mistakes have occurred and will occur, making it critical to
understand how one of these inadvertent mistakes could lead to a nuclear launch and how likely it is that it will happen.

This study comes naturally after the first two presentations in which Seth Baum exposed his reflections on the quantification challenges of the global risk of nuclear war using risk analysis and B. John Garrick outlined the high-level principles and concepts of quantitative risk assessment (QRA) applicable to quantifying global catastrophic risks. Two additional ingredients are adopted for a complete modeling of these scenarios, namely the human reliability analysis and Bayesian estimation techniques for data analysis.

It could be argued that a more relevant and challenging question would be to look at the intentional initiation of nuclear war; however most if not all the direct evidence we have is for cases that could have led to unintentional initiation of a first strike believed to be retaliation. This does not mean that the intentional initiation of nuclear war could not be modeled and quantified as we have the right tools for such analyses. For more details, see Garrick’s previous discussion on evidence. A follow-up study could focus on this set of scenarios.

### III.3.2 SCOPE, LIMITATIONS, ASSUMPTIONS

The scope, data limitations, analysis limitations, and assumptions (SLA) involved in this analysis are given in this section.

The scope of this study is narrowed to cover only the case of a false alarm from an early warning system which propagates to become a cause for first strike out of a belief of retaliation as shown in Figure 1.
The scope of this activity was constrained by the limited research team, and especially access to data. It was intended to showcase the principles and concepts of quantitative risk assessment with what publicly accessible information we could find. A rigorous literature search will be necessary in order to boost the confidence in the results; however, this was not the main scope of our research. During the period we had at our disposal, we were barely able to scratch the surface of a comprehensive analysis. Moreover, we have relied on information that we can openly access, most of which was from the 80s. Therefore, the systems we have characterized are most probably dated and the results should reflect these limitations. It goes without saying that this methodology could reach its full potential in a classified project with access to all the data, procedures, and guidelines that are currently in place. Data for this project came from a variety of open sources, including declassified historical information, testing information, reliability databases, expert knowledge, and empirical data from related systems and conditions.

Nevertheless, a small example on how to perform the analysis on a much smaller scale was set up. At a high level, the event
sequence diagram (ESD) given in Figure 2 summarizes the full range of scenarios developed in this activity.

![Diagram of Event Sequence Diagram of Initiation of Inadvertent Nuclear War](image)

**Figure 2.** Event Sequence Diagram of Initiation of Inadvertent Nuclear War

This high-level ESD was set up based on actual historical events, none of which fully show up in its sequences. The details of these historical events should show up in the cut sets after every pivotal event is further decomposed logically into its constituent elements or influencing factors. Of course, fortunately, all the historical events end up in a “No Launch Response” state.

One of our main goals was to characterize every pivotal event using the appropriate modeling technique (i.e., fault tree (FT) or Bayesian network (BN)) in order to showcase their strengths as it will described later on.

To fit this scenario to the existing data and systems information, the following assumptions were made:

- **Launch Under Attack (LUA) Strategy:** This assumption simplifies the ESD by requiring that the Missile Attack Conference (MAC) can only make the decision to suggest a launch if the threat has been confirmed either by the Threat Assessment Conference (TAC) or by Dual Phenomenology from Radar. The alternate strategy, a Launch on Warning (LOW) strategy allows a decision to launch after only a single detection, which greatly increases the chances of an inadvertent launch\(^{11}\).

---

\(^{11}\) This is an important assumption that is heavily debated among experts. Regardless of the current practices, a comparative quantitative study between the two would be of great value to the decision makers.

Page | 69
*Intercontinental Ballistic Missile (ICBM):* Assuming an ICBM allows enough time for each phase to properly complete before the “Use Them or Lose Them” point. This, as in the case of the first assumption, provides a lower risk estimate than the alternative, namely a Submarine Launched Ballistic Missile (SLBM) that has a much shorter flight time and would likely reach a critical decision point before proper analysis can be completed.

*Context: US-Russia During Cold War (70s-80s):* Due to the lack of public information on both the physical systems, frequency of events, and procedures, the sequences and systems shown in this system are based on those which have been declassified, and which are primarily from the Cold War. Therefore, the sequences represented may be out of date and should be taken as a guideline for how to design such an analysis instead of an accurate representation of current systems and guidelines.

*Procedure-Driven Decision-Making Process:* In this study we have assumed that the decision-making process is mainly procedure-driven. We recognize this assumption may not be applicable to all cases as we have seen situations in which humans had to rely on their knowledge-based judgment instead of written procedures. For more details on how knowledge-based decision-making can be modeled see Li’s work (2013).

Other more specific assumptions will be given in the appropriate section in order to provide a clear context.

### III.3.3 METHODOLOGIES AND TECHNIQUES

As already hinted in the Garrick presentation, the main QRA methodology employed in this study is the Hybrid Causal Logic (HCL) methodology. HCL grew out of the need in QRAs of technological systems to consider not only the elements associated with ‘hard’ causes such as hardware systems, but also the factors generated by ‘soft’ causes such as human activities, physical environment, and socio-economic environment (Wang, 2007; Diaconeasa, 2017). It is a methodology consistent with the general framework of QRA known as the triplet definition of risk (Kaplan & Garrick, 1981) that was developed to be applicable to any type of risk. This methodology proposes a multi-layered
modeling approach so that each individual domain of the system is modeled with the most appropriate technique. The three layers modeled in HCL are:

1. The Event Sequence Diagram (ESD) layer - it is used to model the risk scenario context, which is consistent with the QRA framework based on the triplet definition of risk as highlighted in the Garrick presentation.
2. The Fault Tree (FT) layer – it is used to model the physical systems’ behavior and quantify their impact on their corresponding linked events in the ET.
3. The Bayesian Network (BN) layer – it is used to model the causal relations between ET or FT events that have ‘soft’ root causes.

It is an integrated model with the corresponding sets of taxonomies, analytical, and computational procedures. The interactions of the different HCL layers are shown in Figure 3. The combinations of ESDs, FTs, and BNs are defined as HCL diagrams. One of the main assumptions made in FT analysis is that the basic events are independent. In reality, this may not be the case. In these situations, BN models can be created that link common causal factors to each basic event and introduce explicit dependencies between the basic events. Moreover, in an HCL model, BNs can be used to model the events in the ESDs. Overall, the combination can be made at any level to build an HCL diagram. Additionally, the FTs and BNs can also have common nodes.

For quantification, the connections between the BNs and ESD/FT logic models are formed by binary variables in the BN that correspond to basic events in the FTs or initiating events and pivotal events in the ESDs. Thus, the probability of the linked events is evaluated in the BN. In order to quantify the complete HCL model with all its dependencies, it is necessary to convert the three types of diagrams into a set of models that can be compatible mathematically. This is accomplished by converting the ESDs and FTs into Reduced Ordered Binary Decision Diagrams (ROBDD). A hybrid ROBDD/BN is created to include the BNs enabling the probability of one or more of the ROBDD variables to be given by a linked node in the BN.
Figure 3. a) Schematic of Hybrid Causal Logic Methodology; b) Break Down of a Human Behavior Model Linked to an Event Sequence Diagram. Key Human Activities Can Be Broken Down
into Information Gathering (I), Decision Making (D), and Action Taking (A) (Ekanem 2013)

The other key ingredient of this study is human behavior modeling that employs an adapted version of the Phoenix methodology (Ekanem, 2013). In the Phoenix methodology human behavior is broken down into the following three cyclic stages: information gathering (I), decision-making (D), and action taking (A) as is shown in Figure 3. Each of these events has a probability of failure that is influenced by a set of factors developed later in this paper. Additionally, these events are conditional; for example, a failure of information gathering will increase the likelihood of making a poor decision.

For the quantification of probability of events that have not yet occurred, such as the probability of an inadvertent initiation of a nuclear war, expert opinion techniques are heavily used as a way to compensate for test and field data limitations and to address uncertainties about various models and predictions (Mosleh and Apostolakis, 1984). The Bayesian method for the use of expert opinion adopted here has already been tested and successfully applied to a variety of problems. It is worth mentioning that these are not the only techniques for Bayesian estimation; however, the ones discussed here should be sufficient for our application to inadvertent nuclear war.

In general, expert opinion can be treated as a piece of evidence about an unknown quantity $x$. To update the belief about this particular variable we can use:

$$\pi(x|x^*) = \frac{L(x^*|x) \cdot \pi_0(x)}{\int L(x^*|x) \cdot \pi_0(x) \, dx}$$

Given $x^*$ is the expert’s judgment about the value of $x$ and $L(x^*|x)$ is the likelihood function. There are two well-known models that can be used to build the likelihood function based on expert opinion:

- Additive error model: The expert’s assessment, $x^*$, is expressed as the sum of the true value of $x$, and an error $E$ so that: $x^* = x + E$, where $E$ is the error term and is
distributed normally $E \sim \mathcal{N}(b, \sigma)$, and $b$ is called the bias term. Therefore, the likelihood function takes the form of the normal distribution:

$$L(x^*|x) = \frac{1}{\sqrt{2\pi \sigma}} \exp \left( -\frac{1}{2} \left( \frac{x^* - (x - b)}{\sigma} \right)^2 \right)$$

- Multiplicative error model: The expert’s assessment, $x^*$, is expressed as the product of the true value of $x$, and an error $E$ so that: $x^* = x \cdot E$. By taking the logarithm of this equation we obtain, $ln(x^*) = ln(x) + ln(E)$, where the quantity $ln(E)$ can be assumed to be normally distributed $ln(E) \sim \mathcal{N}(b, \sigma)$. Therefore, the likelihood function is now a lognormal distribution:

$$L(x^*|x) = \frac{1}{\sqrt{2\pi \sigma x^*}} \exp \left( -\frac{1}{2} \left( \frac{ln(x^*) - (ln(x) - ln(b))}{\sigma} \right)^2 \right)$$

As the HCL models can contain many variables which have a wide degree of factors influencing them, it is important to consider the uncertainty of each event. Using uncertainty propagation techniques, the uncertainties are propagated through the layers all the way to the end states. Monte Carlo techniques are used to propagate the uncertainty of each event through the entire model (Figure 4). The actual sampling and quantification are performed on the minimal cut sets.

Figure 4. Uncertainty Propagation using Monte Carlo (Garrick, 2008)

Aside from the uncertainty in the frequency of the risk as described in the Garrick presentation, risk metrics are obtained in the form of time-dependent or independent importance measures that are useful in ranking the scenarios based on their importance.
III.3.4 MODELING THE SEQUENCE OF EVENTS

As part of the HCL methodology, an ESD was made for the given scenario, with individual events linked to FTs and BNs. Before any data could be used, a theoretical understanding of the procedures and systems was modeled from public information on how the US processes alarms.

The decision-making process after the detection of an alarm is split into five phases: investigation of the alarm by the military command post (in this case NORAD), evaluation by the Missile Display Conference (MDC), evaluation by TAC, evaluation by MAC, and finally evaluation by the President (Figure 5). Each phase is assigned a time limit to take a decision. It is assumed that if the time limit is exceeded within a certain phase, the alarm is promoted into the next level of decision-making. As a result, within each phase there are two possible ways for the alarm to be promoted: through a deliberate choice to call forth the next phase after information gathering and decision-making, or through a timeout. The branches going down from these events (i.e., pivotal event false, for example NORAD phase does not reach timeout) all lead to a transfer to a second ESD which covers potential decisions made by lower level officials.

To be noted that in this model we assume a false alarm is triggered that needs to be investigated and correctly identified as false by the decision-makers in order to prevent an inadvertent initiation of a first strike believed to be in retaliation.

![Figure 5. Overview of the Event Sequence Diagram Showing Branching to Either a Timeout or an I-D-A Sequence](image)

To understand the factors that would impact the human activities in each phase, it is important to know the nature of these conferences (Halloran, 1983):
• **NORAD**: Once a sensor picks up a signal, it is sent to the NORAD command post, the Missile Warning Center, and a Space Computation Center, all in buildings near each other that are meant to be resistant to damage from a nearby nuclear attack.

• **Missile Display Conference**: The commanding officer of NORAD gives operators at each sensor site one minute to confirm that the sensor is working. This process occurs several times a day and is considered routine.

• **Threat Assessment Conference**: If the threat is considered credible, the NORAD commander would then call senior military officers, and an assessment would be made about the credibility of the threat. During this phase, a confirmation from one of many radar sensors is needed for dual phenomenology. This assessment is sent to the national Military Command Center in the Pentagon and the President is alerted.

• **Missile Attack Conference**: If the Generals at the Pentagon decide that an attack has been launched, a conference, which would include the President, would be held to determine whether an attack should be made. No MAC has ever been held in the U.S. to date.

Starting from the overview event sequence diagram, a compact event sequence diagram was obtained (Figure 6). The full event sequence diagram can be provided by request as it is too large to fit on paper. The pivotal events shown in blue follow the general pattern given in Figure 5 including the conditional events. Also, the end states designating “Unintended initiation of nuclear war” are shown in red, while the end states representing “No Launch” end states are given in green”.

![Figure 6. Event Sequence Diagram Showing Branching to either a Timeout or an I-D-A Sequence](image)

Page | 76
Given that this example is for a case with the initial false alarm generated by an early warning system, the nature of these systems was modeled using fault trees to identify potential causes of false alarms (Figure 7). Two early warning systems were modeled: infrared satellite and radar, each of which could have a false alarm generated either by an object falling within that system's expected range for a missile, or from a system error due to either hardware or software failures.

**Figure 7. Fault Trees Representing Satellite False Alarm Detection and Dual Phenomenology**

In this model, it is assumed that the first detection must come from a satellite, following the scenario from the overlapping false alarm scenario described by Sennott (1988). During the decision-making process, the initial false alarm may overlap with a second
false alarm from a radar source, increasing the validity of the threat through dual phenomenology. Since dual phenomenology requires that the radar false alarm occur while the initial false alarm is still unresolved, the likelihood of a radar false alarm is coupled with the likelihood of an unresolved false alarm, which is calculated based on the minimum time before radar detection of a real attack after a satellite detection as well as an average 3.5-minute resolution time for a false alarm. This 3.5-minute resolution time is an important assumption, since resolution times can vary greatly and the probability of a false alarm leading to escalation increases drastically with increasing resolution time.

**Table I. Potential Sources of Threshold-based False Alarms in Satellites**

<table>
<thead>
<tr>
<th>Satellite Threshold</th>
<th>Random Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Structures</td>
</tr>
<tr>
<td></td>
<td>Terrain</td>
</tr>
<tr>
<td></td>
<td>Coastline</td>
</tr>
<tr>
<td></td>
<td>Wave Reflection</td>
</tr>
<tr>
<td>Immobile or Background Sources</td>
<td>Plane</td>
</tr>
<tr>
<td>Mobile Sources</td>
<td>Cloud</td>
</tr>
<tr>
<td></td>
<td>Non-hostile Rocket (i.e. Weather Rocket)</td>
</tr>
<tr>
<td></td>
<td>Bird(s)</td>
</tr>
<tr>
<td></td>
<td>Ship</td>
</tr>
</tbody>
</table>

For the evaluation of threshold sources, some information could be obtained about the nature of these sources from historical events and older specifications for aircraft-detecting satellites.
The ten sources used in the example model are listed in Table I. Each source was further expanded to account for the automatic methods for filtering out false alarms (Figure 8). The filters, shown for the case of random noise in the figure below, were taken from a study on fire detecting satellites which also use infrared detection (Arrue et al., 2000).

**Figure 8.** Breakdown of Conditions for False Alarm Detection Representing Safeguards Built within the Satellite or Radar

For the radar threshold case, the sources were assumed to be similar to those for satellites, with a heavier focus on objects in the sky or in orbit, and less focus on terrain and other ground-based sources.

<table>
<thead>
<tr>
<th>Radar Threshold</th>
<th>Random Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural or Manmade Objects</td>
<td>Plane</td>
</tr>
<tr>
<td></td>
<td>Satellite/Debris</td>
</tr>
<tr>
<td></td>
<td>Non-hostile Rocket</td>
</tr>
<tr>
<td></td>
<td>Bird(s)</td>
</tr>
<tr>
<td></td>
<td>Weather and moon cycles</td>
</tr>
</tbody>
</table>

For both satellite and radar, the hardware and software systems were modeled primarily at a high level to show the breakdown into...
subsystems and identify some failure modes that could potentially create false alarms (Figure 9).

Figure 9. Satellite and Radar Subsystems

A more thorough analysis could further expand on each type of failure to identify the failure rates of parts and software in the system. An example of this breakdown was carried out for the circuit in the Command and Data Handling (C&DH) subsystem, using commercial off-the-shelf parts. This break down is shown in Figure 10.

Figure 10. Satellite Hardware Deeper Examination into C&DH with Breakdown into Circuit
Three fault trees modeling human failure events (Figure 11) were adapted from the Phoenix HRA methodology developed primarily for nuclear power plants operation (Ekanem, 2013). Fault trees for information gathering, decision making, and action taking were linked for each phase of the event sequence diagram and, through connections to BNs, are able to account for changes in the environment and mindset as the alarm propagates.

**Figure 11.** Human Failure Fault Trees for a) Information Gathering, b) Decision Making and c) Action Taking Phases
The Bayesian Network used for quantification and as a basis for the qualitative modeling of the phases of decision making was taken from Ekanem (2013) and it is shown in Figure 12. This network considers failure modes from each of the I-D-A phases and uses the following first layer of performance influencing factors (Figure 12 in red): human system interface, procedures, resources, team effectiveness, knowledge/abilities, bias, stress, task load, and time constraint.

Figure 12. General Bayesian Performance Influencing factors
Reproduced from Ekanem (2013)

Modified BNs were then constructed for the NORAD, MDC, TAC, MAC, and President phases (Figure 13) with additional PIFs: local political or economic status, national alliances, tunnel vision, fitness for duty, environment condition (time of day, crisis or peace, etc.), system complexity, delegation of authority, perception of distant risk, fear of release to the public, economic vs. national security, emotional state.

The first two BNs, that is for NORAD and MDC, cover mainly failures to accurately collect and report information regarding the validity of the alarms. At this stage, events occur routinely and are handled procedurally.

In addition to the above PIFs, two new crew failure modes are added to represent the case of politically or emotionally motivated
human failure events. The first failure mode represents cases where information is not reported correctly, not due to a miscommunication, but due to biased, emotional, or intentionally incorrect reporting. This failure mode would be most relevant in the early stages of the decision-making process while information is being gathered and communicated. The second crew failure mode captures irrational or emotionally charged decisions that could be made by officials later in the decision-making process when stress, fear, and political motivation would be strongest.
III.3.5 MODEL QUANTIFICATION

Two primary methods were used for quantifying the system fault trees linked to the initiating event: top-down quantification using historical data on the frequency of false alarms, and bottom-up quantification using failures in time (FIT) rates for individual parts obtained from the Texas Instruments Reliability Database. Overall, the probability of failure of the systems in a year were taken directly from historical data, but the bottom-up quantification was still shown as an example of how a thorough analysis of new satellite or radar systems could be carried out with complete information on the design of these systems.

Data on the frequency of false alarms was taken from a published table from NORAD (Table III). Since there is consistently more than one event per year, we assumed a probability of 1.0 for an initial false alarm, which is used as the top gate for the satellite false alarm detection fault tree. The frequency of these events will later be used to determine the frequency of inadvertent initiation of nuclear war, given that the probability obtained from the event sequence diagram is on a per false alarm basis. It should be noted that even if in Table III the number of false alarms is on the order
Table III: NORAD Data on Number of False Alarms Per Year, 1977-1984

<table>
<thead>
<tr>
<th>YEAR</th>
<th>FALSE ALARMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1977</td>
<td>1567</td>
</tr>
<tr>
<td>1978</td>
<td>1009</td>
</tr>
<tr>
<td>1979</td>
<td>1544</td>
</tr>
<tr>
<td>1980</td>
<td>3815</td>
</tr>
<tr>
<td>1981</td>
<td>2851</td>
</tr>
<tr>
<td>1982</td>
<td>3716</td>
</tr>
<tr>
<td>1983</td>
<td>3294</td>
</tr>
<tr>
<td>1984</td>
<td>2988</td>
</tr>
</tbody>
</table>

of thousands per year, consideration should be given to the severity of the alarms. Thus, although the several events occurred each day, only about five per year were serious enough to make it past routine checks.

The joint conditional probabilities of the crew failure modes obtained by quantifying the BNs (Ekanem, 2013) while setting each performance influencing factor’s influence at 0%, 50%, and 100% being true (Table IV).

Table IV. Joint Conditional Probabilities of the Crew Failure Modes

<table>
<thead>
<tr>
<th>Name</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key Alarm Not Responded to</td>
<td>8.47E-05</td>
</tr>
<tr>
<td>Data Not Obtained (Intentional)</td>
<td>4.52E-02</td>
</tr>
<tr>
<td>Data Discounted</td>
<td>4.52E-02</td>
</tr>
<tr>
<td>Decision to Stop Gathering Data</td>
<td>1.73E-02</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Data Incorrectly Processed</td>
<td>6.72E-03</td>
</tr>
<tr>
<td>Reading Error</td>
<td>1.49E-03</td>
</tr>
<tr>
<td>Information Miscommunicated</td>
<td>3.95E-03</td>
</tr>
<tr>
<td>Wrong Data Source Attended to</td>
<td>6.60E-04</td>
</tr>
<tr>
<td>Data Not Checked with Appropriate Frequency</td>
<td>1.31E-02</td>
</tr>
<tr>
<td>System State Misdiagnosed</td>
<td>4.10E-02</td>
</tr>
<tr>
<td>Procedure Misinterpreted</td>
<td>3.19E-03</td>
</tr>
<tr>
<td>Failure to Adapt Procedure to the Situation</td>
<td>9.32E-03</td>
</tr>
<tr>
<td>Procedure Step Omitted (Intentional)</td>
<td>6.60E-03</td>
</tr>
<tr>
<td>Deviation from Procedure</td>
<td>6.60E-03</td>
</tr>
<tr>
<td>Decision to Delay Action</td>
<td>6.60E-03</td>
</tr>
<tr>
<td>Inappropriate Strategy Chosen</td>
<td>9.20E-03</td>
</tr>
<tr>
<td>Incorrect Timing of Action</td>
<td>8.77E-03</td>
</tr>
<tr>
<td>Incorrect Operation of Component/Object</td>
<td>1.87E-03</td>
</tr>
<tr>
<td>Action on Wrong Component/Object</td>
<td>2.47E-03</td>
</tr>
</tbody>
</table>

The set of data used in the computation of human failure events was based on available theories of cognitive sciences and psychology, experimental results, operation of US and German nuclear power plants experience, and expert opinion.

Next, an example Bayesian estimation calculation performed with the Bayesian Estimation Web Application\(^\text{12}\) is given for how to process expert opinion information with a log-normal prior model and log-normal likelihood model. We used the log-normal distribution to specify the source to source variability of human failure.

\(^{12}\) https://www.risksciences.ucla.edu/software/
failure events probability estimates since it can be generally used to express orders of magnitude variations in the estimates of the quantity of interest. In Table V, the prior model parameters, the prior mean and error factor (EF), are given. It should be noted that these parameters are also uncertain, thus their uncertainty needs to be specified as well as the EF of the mean and EF of the EF. The EF is obtained by the ratio between the 95th and 50th percentiles. In general, the EF is used to quantify the uncertainty around the estimate. Lower values are assigned to more certain estimates. For expert opinions values between 3-5 are considered appropriate.

Table V. Prior Parameters for Bayesian Estimation

<table>
<thead>
<tr>
<th>Value</th>
<th>Mean</th>
<th>EF of Mean</th>
<th>EF</th>
<th>EF of EF</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>5.1803</td>
<td>10</td>
<td>5.1803</td>
<td></td>
</tr>
</tbody>
</table>

The likelihood model parameters that were estimated by four experts, including the uncertainty (i.e., EFs), are shown below in Table VI. A Monte Carlo simulation was performed to obtain the posterior distribution. The obtained posterior distribution parameters are given in Table VI.

Table VI: Likelihood and Posterior Values

<table>
<thead>
<tr>
<th>Expert</th>
<th>Mean</th>
<th>EF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1</td>
<td>0.03</td>
<td>3</td>
</tr>
<tr>
<td>Expert 2</td>
<td>0.05</td>
<td>5</td>
</tr>
<tr>
<td>Expert 3</td>
<td>0.04</td>
<td>5</td>
</tr>
<tr>
<td>Expert 4</td>
<td>0.01</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0552</td>
<td>0.0118</td>
</tr>
</tbody>
</table>

With this method, we have generated a variability distribution for each of the model parameter estimates in order to obtain the posterior distribution as shown in Figure 14.
III.3.6 RESULTS AND DISCUSSION

Given the initiating event of a false alarm with a probability of 1.0, the probability of inadvertently initiating a nuclear war given the described scenarios in the model was calculated using the HCLA Command-Line Tool. The uncertainties were propagated using Monte Carlo sampling techniques. Moreover, by multiplying this probability with the frequency of false alarms per year given in the model quantification discussion, the probability of frequency distribution from Figure 15 was obtained with a median value of 7.65E-4 per year. To put it in perspective, the core damage frequency of the U.S. nuclear power plants was estimated at 2.0E-5 per reactor-year (Gaertner, 2008).

It should be reiterated that this value represents only the expected frequency for nuclear war initiated by false alarms generated from an early warning system. It is meant to be taken as a demonstration of the methodology and not an accurate determination of the likelihood of nuclear war. Moreover, as important as the quantitative results, are the qualitative results in the form of cut sets. Given the complexity of the systems and various failure modes in which the human can fail, important scenarios may be easily overlooked. For example, the two scenarios shown in Figures 17-18 that show up in the list of cut sets have already happened.
On September 26, 1983, the initial false alarm was caused by sunlight reflection off of clouds that was picked up by a Soviet early warning satellite as a launch of five missiles. The responsibility to assess the validity of this alarm and to pass along the information came to Lt. Stanislav Petrov, who broke protocol and used his own judgment to determine that this was likely a false alarm since it was unlikely that the U.S. would launch only five missiles. This example is graphically shown below in Figure 16, with the red nodes indicating failures, green nodes indicating successes, and yellow nodes indicating a flag. It is notable that the event in which the sensor operator reports the false alarm is considered a failure. Though this is counterintuitive, failure in these types of scenarios represents any event in which the false alarm is not identified and leads to further progression towards nuclear retaliation.

**Figure 15.** Final PDF for Initiation of Nuclear War

**Figure 16.** Example Cut Set Highlighting the Petrov Incident
Figure 17. Example Cut Set Highlighting the Norwegian Rocket Incident

On January 25 of 1995, a weather rocket launched by a team of Norwegian and U.S. scientists was mistaken for a missile launch by a Russian radar system. Though Russia had been alerted of the missile’s launch, they failed to identify the false alarm until after the nuclear keys had already been unlocked and the decision fell to President Yeltsin. This example, shown in Figure 17, shows a further progression of a false alarm through each stage. The false alarm was finally identified by new information in the final stage of decision making, and no further escalation occurred.

Other historical events of note captured in the set of scenarios analyzed are:

- On 24 November of 1961, all communication links went dead between the Strategic Air Command (SAC) HQ and NORAD, and so cut SAC HQ off from the three Ballistic Missile Early Warning sites (BMEWS) at Thule (Greenland), Clear (Alaska), and Filingdales (England).
• On 24 October of 1962, a Russian satellite entered its parking orbit, and shortly afterwards exploded making the U.S. believe that the USSR was launching a massive ICBM attack. NORAD command post logs remain classified.
• On November 1965, the power failed resulting in faulty bomb alarms. During the commercial power failure in North Eastern United States in November 1965, displays from all the bomb alarms for the area should have shown yellow. Moreover, two of them from different cities showed red because of circuit errors.
• On November 9, 1979 at 3:50 AM a computer exercise tape simulating a Soviet ICBM attack on the United States is accidentally loaded, causing NORAD and other command centers to be told that a nuclear attack is in progress.
• The warning displays at the Command Centers included displays that normally showed “0000 ICBMs detected” and “0000 SLBMs detected.” On 3 June of 1980 these indicators started showing random numbers of missiles detected.

III.3.7 CONCLUSIONS
This study used quantitative risk assessment (QRA) methods to identify scenarios that could lead to the initiation of an inadvertent nuclear war stemming out of false alarms. A set of scenarios for inadvertent nuclear war initiated by a false detection of an ICBM launch was examined in detail starting with the potential sources of false alarms. The false alarm was then passed through each level of decision making until it was either detected or a decision to launch a preemptive strike was made. The scenario model used three levels of modeling: event sequence diagrams at the top level to capture the risk scenario context, fault trees to model the potential causes of system failures or human failure events, and Bayesian networks to further examine the factors that influence the crew failure modes. The likelihood of inadvertent nuclear war, including uncertainties, was calculated using the Hybrid Causal Logic Analyzer (HCLA) with data from a wide variety sources and processed using Bayesian estimation techniques. Although a value of 7.65E-4 per year was obtained for the frequency of inadvertent initiation of nuclear war triggered by a false alarm, a more complete analysis should be performed to gain confidence in the numerical results. Moreover, the
qualitative results obtained in the form of cut sets are as important as the quantitative results as overlooked scenarios may be discovered given the high complexity of the systems and procedures designed as barriers to prevent the inadvertent initiation of nuclear wars.

REFERENCES


Li, Y., “Modeling and Simulation of Operator Knowledge-Based Behavior.”, *PhD Dissertation, University of Maryland, College Park, Maryland* (2013)


III.4 REPRODUCED: CLASSIFYING GLOBAL CATASTROPHIC RISKS – JULIUS WEITZDÖRFER

This presentation was based on the following paper:

Classifying global catastrophic risks

Shahar Avin, Bonnie C. Wintle, Julius Weitzdörfer, Seán S. HÉigeartaigh, William J. Sutherland, and Martin J. Rees

The paper is attached in these proceedings with the first author’s permission obtained on December 10, 2018.
Classifying global catastrophic risks

Shahar Avin*, Bonnie C. Wintle, Julius Weitzdörfer, Seán S. Ó hÉigeartháigh, William J. Sutherland¹, Martin J. Rees

Centre for the Study of Existential Risk, University of Cambridge, UK

ARTICLE INFO

Keywords:
- Global catastrophic risk
- Existential risk
- Foresight
- Planetary boundaries
- Anthropogenic risk
- Classification
- Interdisciplinary
- Risk assessment
- Prevention and mitigation

ABSTRACT

We present a novel classification framework for severe global catastrophic risk scenarios. Extending beyond existing work that identifies individual risk scenarios, we propose analysing global catastrophic risks along three dimensions: the critical systems affected, global spread mechanisms, and prevention and mitigation failures. The classification highlights areas of convergence between risk scenarios, which supports prioritisation of particular research and of policy interventions. It also points to potential knowledge gaps regarding catastrophic risks, and provides an interdisciplinary structure for mapping and tracking the multitude of factors that could contribute to global catastrophic risks.

1. Introduction

In our uncertain times it is good to have something we can all agree on: global catastrophes are undesirable. As our science advances we gain a better understanding of a broad class of global catastrophic risk (GCR) scenarios that could, in severe cases, take the lives of a significant portion of the human population, and may leave survivors at enhanced risk by undermining global resilience systems (Baum & Tonn, 2015; Bostrom, 2002; Bostrom & Ćirković, 2008; Posner, 2004; Rees, 2003; Tonn & MacGregor, 2009). Much progress has been made in identifying individual GCR scenarios, and in compiling lists of the scenarios of greatest concern, but there is currently no known methodology for compiling a comprehensive, interdisciplinary view of severe global catastrophic risks. While a fully complete list of GCRs may remain beyond reach, we present here a classification framework designed specifically to draw on as broad a knowledge base as possible, to highlight commonalities between risk scenarios and identify gaps in our collective knowledge regarding global catastrophic risks.

To date, research on global catastrophic risk scenarios has focused mainly on tracing a causal pathway from a catastrophic event to global catastrophic loss of life (Asimov, 1981; Bostrom & Ćirković, 2008; Coburn et al., 2014; Cotton-Barratt, Farquhar, Halstead, Schubert, & Snyder-Beattie, 2016; Turchin, 2015). Such research has been fruitful in identifying and assessing a range of such GCR scenarios. Some severe GCR scenarios have posed a persistent threat to humanity since our emergence as Homo sapiens (e.g. impact by a 10km astronomical object, or a volcanic super-eruption of 1000 km³ of tephra). Other scenarios have increased in likelihood following human population expansion and the accompanying increase in resource demands (e.g. natural pandemics or ecosystem collapse). In addition, novel GCR scenarios can accompany new technologies: some of these are relatively well established (e.g. “nuclear winter” or an engineered pandemic); others are more speculative (e.g. accidents in or weaponisation of advanced artificial intelligence, or environmental shocks from ill-judged geoengineering efforts aimed at mitigating climate change).

* Corresponding author.
E-mail address: s478@cam.ac.uk (S. Avin).
1 William J. Sutherland is professor of conservation biology in the Department of Zoology, University of Cambridge, UK.

https://doi.org/10.1016/j.futures.2018.02.001
Received 26 May 2017; Received in revised form 26 January 2018; Accepted 12 February 2018
Available online 23 February 2018
0016-2377/ © 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Fig. 1. Classification of Critical Systems aimed at identifying Global Catastrophic Risk scenarios. Systems are grouped at different levels, arranged from “lower level” to “higher level” in a clockwise fashion starting with the “Physical” group on the top right.

However, compiling a comprehensive list of plausible GCR scenarios requires exploring the interplay between many interacting critical systems and threats, beyond the narrow study of individual scenarios that are typically addressed by single disciplines. The classification framework presented here breaks down the analysis of GCR scenarios into three key components: (i) a critical system (or systems) whose safety boundaries are breached by a potential threat, (ii) the mechanisms by which this threat might spread globally and affect the majority of the human population, and (iii) the manner in which we might fail to prevent or mitigate both (i) and (ii). For example, a major astronomical impact may lead to a global catastrophe if we lack the technology to deflect it (mitigation failure), and it raises a cloud of dust that spreads around the world (global spread mechanism), and that cloud of dust blocks sunlight for a sufficient length of time to undermine the global food system in a manner that we cannot overcome (critical system affected). Other scenarios will have different combinations of one or more mitigation failures, one or more global spread mechanisms, and one or more critical system breaches.

In order to gain a holistic picture of potential global catastrophes, knowledge about each of the three system components needs to be explored and shared. By first constructing a classification from the broad range of known critical systems, global spread mechanisms, and prevention and mitigation failures, and then by classifying known GCR scenarios according to these dimensions, we aim to: (i) showcase the GCR relevance of a variety of scientific disciplines, (ii) highlight how commonalities between threat scenarios have research and policy implications, and (iii) highlight areas where there are potential gaps in our knowledge of global catastrophic risks. We also propose concrete steps for coordinating the broad-based, interdisciplinary research required to meet the challenges highlighted by the framework.

2. Critical systems

We define a “critical system” as any system or process that, if disturbed beyond a certain limit or scale, could trigger a significant reduction in humanity’s ability to survive in its current form (see Fig. 1).

Building on the “life support systems” outlined in the research on so-called planetary boundaries (Rockström et al., 2009; Steffen et al., 2015) (many of which appear in our biogeochemical group), and their potential links to GCRs (Baum & Handoh, 2014), we identify critical systems and processes that, if disrupted, would affect human ability to survive. While we aim for comprehensiveness and minimal overlap, we acknowledge that different systems overlap. For example, while the processes affecting ocean acidity have direct effects on ecosystem stability and thus human life, there is significant overlap (causally, structurally and academically) with the global water cycle, carbon cycle and sulphur cycle systems.

In our classification framework, critical systems are grouped at different levels in a hierarchy, such that “higher-level” systems rely on the functioning of those at a “lower-level”. Thus, the framework builds up from the stability of life-supporting physical
systems, through cellular and other systems, right up to species-wide ecological and sociotechnical systems. “Lower-level” systems are directly linked to human survival (which relies on functioning anatomical systems, which in turn relies on cellular systems, etc.). “Higher-level” systems, especially technology-enabled ones such as the food and health systems, help maintain the human population at its current size, and provide resilience. If these “higher-level” systems were to be disturbed significantly in some scenario, e.g. through a severe and prolonged disruption to utilities networks (such as water and electricity), or through shock effects (such as social unrest), these could cause more harm than the system disturbance itself.

Identification of critical systems, and their cross-links, could also come from historical and archaeological study of more limited instances of human population collapse. For instance, the collapse of the Easter Island civilisation shows how excessive resource extraction (of palms for the making of canoes) led to ecological degradation, undermining primary production and food chains, which in turn led to failure of the Easter Island society’s food system (Morrison, 2006). Further study of each critical system requires specialised expertise, often in more than one domain, as there is no one-to-one mapping from scientific disciplines to critical systems. Future work, conducted with collaboration with the wider scientific community, could lead to the demarcation of safe operating bounds for each critical system, following the example of Rockström et al. (2009).

3. Global spread mechanisms

For many critical systems, a failure of some instances of the system, e.g., regional crop failure, would fall far short of posing a GCR. In severe GCR scenarios, the failure of critical systems is coupled with some mechanism by which this failure spreads globally, thus potentially threatening the majority of the human population. In the framework, we separate the analysis of global spread mechanisms from the analysis of critical systems (Fig. 2). This separate focus on global spread allows us to identify relevant mechanisms (and means to manage or control them) as targets of study merit further attention, and highlights interesting commonalities.

A critical system failure can spread globally without human intervention: some astronomical objects or events are sufficiently massive to have direct global effect, while other threats can spread through the dynamic systems of the natural environment, such as the air- and water-based dispersal systems. Dust and toxins could be spread naturally even if they do not replicate, though of course a self-replicating threat (e.g. a virus that affects multiple species of fish) could couple with a dynamic system (e.g. ocean currents) to achieve much faster spread.

In addition to natural spread, many risk scenarios, and especially emergent risk scenarios, rely on the highly connected nature of our species, both materially and conceptually. A modern pandemic can spread through airports and other mass-transit hubs of the globe-encircling transit network, thus coupling a biological replicator (this might be, e.g., a bacterium itself, or a biological vector, e.g. a mosquito) to a highly connected anthropogenic network. A cyber attack can escalate through global critical systems at the speed of digital communication, shutting down health and security systems, and undermining resource extraction and utilities by disrupting mines and power plants (a digital replicator, such as a computer worm, could spread the up speed rate and reach).

Access to information can play a more abstract, but no less important, role in the spread of critical system failure. The widespread, and growing, access of individuals and groups across the globe to ideas, schematics, and manufacturing capabilities (e.g. Do-It-Yourself, or DIY, biology) through digital and cultural exchanges (e.g. online fora), enables novel hypothetical GCR scenarios. Such a scenario could start with, say, the accidental or malicious release of a home-grown pathogen, or the one-sided deployment of geoengineering efforts in an attempt to mitigate climate change. Some ideas encourage their own spread, e.g. schematics for communication devices, or ideas that encourage further sharing of those ideas (e.g. ideologies or viral videos), coupling cultural replicators with human interaction networks.

Table 1 illustrates how analysis of critical systems and analysis of global spread mechanisms might be combined into a single classification framework. The table presents a mapping from eight hypothetical GCR scenarios to the critical systems that are most likely to be undermined in each scenario, for each type of global spread mechanism. We have chosen a selection of severe GCR scenarios that are (i) familiar, (ii) considered plausible, and (iii) cover both natural and anthropogenic threats. This is far from a comprehensive list of scenarios, as the very framework presented here aims to help explore possible scenarios.

4. Prevention and mitigation failures

Analysing GCR scenarios along the dimensions of critical systems and spread mechanisms draws significantly on our understanding of the natural world and technical systems, and complements existing endeavours to classify risks of a smaller scale (BRDR, 2014). Holistic risk management, however, must take into account the human elements that moderate GCR through prevention and mitigation efforts, and how these efforts might fail. The challenge of preventing global catastrophes thus requires integration of the
Table 1
Classification of hypothetical global catastrophic risk scenarios by global spread mechanisms and critical systems affected. Letters represent eight examples of risk scenarios: asteroid impact (a), volcanic super eruption (v), pandemic (natural) (p), ecosystem collapse (o), nuclear war (n), bioengineered pathogens (b), unpowered artificial intelligence (w), geoeconomic termination shock (g). Cell colour represents number of catastrophic scenarios potentially compromising the critical system globally via the spread mechanism (grey: no likely disruption, light pink: one scenario, dark pink: two scenarios, red: three or more scenarios). Critical systems with an identical vulnerability profile to these risk scenarios have been omitted for brevity, indicated by ellipses (see Fig. 1 for the full list of systems). (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.)

<table>
<thead>
<tr>
<th>Global spread mechanism</th>
<th>Natural global scale</th>
<th>Anthropogenic networks</th>
<th>Replicators</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Socioeconomic</td>
<td>Geoeconomic</td>
<td>Biological</td>
</tr>
<tr>
<td>Physical</td>
<td>Stable separation</td>
<td>b</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Complex organic molecules</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>Radiation &amp; temperature levels</td>
<td>a</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Atmospheric</td>
<td>Carbon, Oxygen cycle</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Nitrogen cycle</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Phosphorus cycle</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Sulphur cycle</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>Water cycle</td>
<td>a</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Water cycle</td>
<td>a</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Cellular</td>
<td>Contraction, Signaling</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Mitochondria</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Endoplasmic Reticulum</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Nucleus</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Lysosomes</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Mitochondria</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Endoplasmic Reticulum</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Nucleus</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Lysosomes</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>Whole organism</td>
<td>Coordination, learning</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Creativity, Reproduction</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>Ecological</td>
<td>Decomposition</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Water cycle</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Food cycle, nutrition</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Climate cycle</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Food, Resources withdrawal</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Security</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>Society</td>
<td>a</td>
<td>a</td>
</tr>
</tbody>
</table>

work and expertise in and between the natural and the social sciences, on a global scale.

A particularly comprehensive existing risk management framework with such integrative characteristics and international scope is the Sendai Framework for Disaster Risk Reduction (SFDRR), adopted by 187 UN member states in 2015 (UNISDR, 2015). Although developed for natural rather than technological disasters, it considers many of the potential human factors that influence resilience and vulnerability to an unfolding disaster. We take a similar approach here, and identify potentially fragile areas in the global risk prevention and mitigation system (Fig. 3). Rather than aiming for comprehensiveness or exclusivity, it highlights that understanding these interdependent and complex human factors requires input from a wide range of disciplines beyond the natural sciences.

For instance, short-term thinking and a limited focus constitute cognitive biases affecting risk perception and management on the
work and expertise in and between the natural and the social sciences, on a global scale.

A particularly comprehensive existing risk management framework with such integrative characteristics and international scope is the Sendai Framework for Disaster Risk Reduction (SFDRR), adopted by 187 UN member states in 2015 (UNISDR, 2015). Although developed for natural rather than technological disasters, it considers many of the potential human factors that influence resilience and vulnerability to an unfolding disaster. We take a similar approach here, and identify potentially fragile areas in the global risk prevention and mitigation system (Fig. 3). Rather than aiming for comprehensiveness or exclusivity, it highlights that understanding these interdependent and complex human factors requires input from a wide range of disciplines beyond the natural sciences.

For instance, short-term thinking and a limited focus constitute cognitive biases affecting risk perception and management on the global scale.
individual and institutional level (as studied in psychology and behavioural economics); unresolved political conflicts and competing ethical notions of justice undermine international cooperation and burden-sharing on the institutional and supra-institutional level (as studied in e.g. law, philosophy and political science).

Some risks (e.g. natural pandemics) are already the focus of well-developed institutional systems (e.g. the World Health Organization), robust research activity and technical know-how. For GCRs from emerging technologies, however, the institutional risk and a research agenda are only just becoming established. Conventional disaster response (e.g. recovery and compensation), and even newer, comprehensive strategies (e.g. the “build back better” principle adopted in some countries post-disaster) are inadequate for addressing threat scenarios where there is limited reaction time and no second chance. For these cases, we need a novel framework that is at least as interdisciplinary as the SFDRR, but moves away from uni-dimensional, natural hazards and instead addresses complex, anthropogenic risks, which are far more likely to cause a severe global catastrophe (Rees, 2003). In particular, we have to focus on the prevention and mitigation of multidimensional risk scenarios that involve cascades of socio-technological, natural-technological (“natech”) and technological-natural disasters.

As we confront emergent technological GCR scenarios, lessons can be learnt from previous smaller disasters. An instructive recent case of a multi-dimensional disaster scenario, albeit of local scope, is the Fukushima Dai-ichi nuclear accident, which laid bare failures at the interface of natural, scientific, technological, socioeconomic, legal and political realms. One such failure was the supervision of Japan’s nuclear industry by the very same authorities that were to promote nuclear technology. Such an institutional setup, aggravated by cognitive biases (e.g. groupthink) in a sector with revolving doors to the regulator, was lacking adequate incentive structures, and was destined to result in conflicts of interest and regulatory capture. The international science and policy community therefore has the opportunity and the responsibility to co-create better risk prevention and mitigation systems, by engaging with researchers in the social sciences and humanities.

In principle it is possible to create a table that would expand on Table 1 to include the third dimension described here, i.e., prevention and mitigation failures. Such a table is, however, difficult to produce in practice, as the scenarios it helps us distinguish between are more fine-grained than those classified in Table 1. They are subcategories of these scenarios. For example, in Table 1 we classified “natural pandemic” as a single scenario, yet from a disaster policy and risk reduction perspective there is a clear difference between a pandemic that emerged due to underinvestment in veterinary surveillance, and a pandemic that emerged due to accidental release from a research laboratory. These scenarios can be further subdivided through the precise failures that allow the pandemic risk to materialise. If we consider just the accidental release scenario, we would start from the grid items occupied by ‘p’ in Table 1, which highligh intersections of the critical systems undermined by pandemic, such as anatomical systems, and the spread mechanisms for pandemic, which naturally include biological replicators but are also affected by anthropogenic networks as well as air- and water-based dispersal. To these we would add a third dimension, that would highlight all the prevention and mitigation failures potentially involved in accidental release, from failures of individual skill or risk perception, through institutional failures including malformed incentives, or insufficient staffing and resources, to supra-institutional failures of insufficient monitoring and enforcement.

5. Intended use of the classification system

In this section, we illustrate three key ways the classification system could potentially be used, although more may be discovered as the system is expanded and updated.

The first potential use is to prioritise risk reduction efforts. As can be seen in Table 1, scenarios with significantly different primary causes could manifest their GCR potential through a similar mechanism. For example, asteroid impact-, volcanic super-eruption-, and nuclear war scenarios all feature a risk of significant reduction of inbound solar radiation, disrupting food security and potentially leading to mass starvation. Not only does this draw attention to systems that are vulnerable to multiple hazards, but it also suggests there is value in considering these scenarios together in research and policy contexts, rather than thinking about them in isolation. For example, if accounting for volcanic super-eruptions, asteroid impacts and nuclear wars together, one might seriously consider risk management strategies that are robust to all scenarios, such as alternative food production systems to withstand the multi-year “winter” that might follow (Denkenberger & Pearce, 2015). While this does not preclude investment in nuclear disarmament or asteroid deflection, it demonstrates that alternative food policies may warrant more attention than first thought.

In addition to the challenge of securing food under reduced solar radiation, the classification framework highlights other areas that warrant further attention as potentially occurring from a range of threats. These include how to manage the proliferation of potentially dangerous technologies, how we would function if human contact was restricted during a pandemic spread, and how we might make critical digital systems resilient to disruption by error or malware. The value of the classification system in highlighting potentially compatible risk reduction strategies is visualised in Table 2.

While expansion of this table into the third dimension of prevention and mitigation failures is beyond the scope of the current paper, we foresee that the creation of such an expansion, in a dynamic and collaborative fashion as described below, will have the same benefits as Tables 1 and 2. That is, it could be used to focus attention on prevention and mitigation failure categories that affect a range of GCR scenarios (e.g. better risk communication tools). While policy relevance to multiple risks does not directly entail higher priority for an intervention (as matters of probability, effectiveness and cost need to be taken into account), it could indicate the value of a comprehensive cross-risk analysis, to paint a more complete picture of the value of a proposed intervention.

The second potential use for the classification system lies in creating a live reference list of expertise for different risk scenarios. Our attempt to carve out categories in each dimension based on different academic domains should provide a quick index of the academic disciplines that are essential to “have at the table” when researching a specific risk scenario. Such an index could prove useful for policy makers who take responsibility for certain risk domains, or when an emerging risk is unfolding and an
interdisciplinary team needs to be assembled in a hurry. This potential use underscores the importance of including the third dimension, which points to relevant academic disaster management expertise outside the natural sciences.

The third potential use for the system is as a tool to highlight highly uncertain or neglected corners of the GCR possibility space, and guide research efforts towards these corners, in the hope of discovering unknown unknowns. The combinatorial nature of the classification systems provides a natural way of progressing from well-known systems and mechanisms to a vast and as-yet largely unexplored space of possible GCRs. Admittedly, even an exhaustive exploration of all possible GCR scenario configurations within the current classification system would not provide a guarantee against "black swans", but it can certainly foster a fuller understanding of the threats we face.

6. Where to next?

The classification framework presented above is dynamic, spanning a broad range of disciplines and reflecting a dense web of interacting variables along three dimensions: where critical systems are vulnerable to GCRs, how threats might spread globally, and how attempts to prevent or mitigate these threats might fail due to human factors. To successfully maintain awareness and organise the plethora of knowledge around GCRs we need to meet the following challenges:

1. collect, aggregate and digest information from highly distributed knowledge networks, overcoming communication barriers and delays;
2. update regularly the classification of GCR scenarios as knowledge advances, and as technology shapes—or is poised to shape—the relevant domains.

Meeting these challenges requires a combination of strategies. It would be sensible to populate a classification framework using a group elicitation approach, calling on experts in different critical systems, global reach mechanisms and mitigation approaches to produce short summaries containing signposts to evidence in their fields that would be relevant to GCRs. Such summaries would then be aggregated in a central repository. A group of multi-domain experts could serve as editors to make sure efforts are coordinated, language is harmonised and appropriate for an interdisciplinary audience, and credit is attributed appropriately. Similar, successful repositories for other disciplines already exist and could provide inspiration (Wolfrum, 2017; Zalta, 2016). The evolving classification system, when part of a knowledge synthesis effort, could offer a visual way to communicate the current state of knowledge (McKinnon, Cheng, Garside, Masuda, & Miller, 2015).

As the frontiers of knowledge and innovation expand, so too does the horizon of our possible futures. The framework outlined here could both inform, and be informed by, different foresight tools (Cook, Inayatullah, Burgman, Sutherland, & Wintle, 2014). It may be a useful tool for generating scenarios that help us explore and prepare for new risks, emerging trends and key uncertainties. Scenarios can then be characterised in more detail and monitored using horizon scanning (Amanatidou et al., 2012; Sutherland &
interdisciplinary team needs to be assembled in a hurry. This potential use underscores the importance of including the third dimension, which points to relevant academic disaster management expertise outside the natural sciences.

The third potential use for the system is as a tool to highlight highly uncertain or neglected corners of the GCR possibility space, and guide research efforts towards these corners, in the hope of discovering unknown unknowns. The combinatorial nature of the classification systems provides a natural way of progressing from well-known systems and mechanisms to a vast and as-yet largely unexplored space of possible GCRs. Admittedly, even an exhaustive exploration of all possible GCR scenario configurations within the current classification system would not provide a guarantee against “black swans”, but it can certainly foster a fuller understanding of the threats we face.

6. Where to next?

The classification framework presented above is dynamic, spanning a broad range of disciplines and reflecting a dense web of interacting variables along three dimensions: where critical systems are vulnerable to GCRs, how threats might spread globally, and how attempts to prevent or mitigate these threats might fail due to human factors. To successfully maintain awareness and organise the plethora of knowledge around GCRs we need to meet the following challenges:

1. collect, aggregate and digest information from highly distributed knowledge networks, overcoming communication barriers and delays;
2. update regularly the classification of GCR scenarios as knowledge advances, and as technology shapes—or is poised to shape—the relevant domains.

Meeting these challenges requires a combination of strategies. It would be sensible to populate a classification framework using a group elicitation approach, calling on experts in different critical systems, global reach mechanisms and mitigation approaches to produce short summaries containing signposts to evidence in their fields that would be relevant to GCRs. Such summaries would then be aggregated in a central repository. A group of multi-domain experts could serve as editors to make sure efforts are coordinated, language is harmonised and appropriate for an interdisciplinary audience, and credit is attributed appropriately. Similar, successful repositories for other disciplines already exist and could provide inspiration (Wolfrum, 2017; Zalta, 2016). The evolving classification system, when part of a knowledge synthesis effort, could offer a visual way to communicate the current state of knowledge (McKinnon, Cheng, Garside, Masuda, & Miller, 2015).

As the frontiers of knowledge and innovation expand, so too does the horizon of our possible futures. The framework outlined here could both inform, and be informed by, different ‘foresight’ tools (Cook, Inayatullah, Burgman, Sutherland, & Wimble, 2014). It may be a useful tool for generating scenarios that help us explore and prepare for new risks, emerging trends and key uncertainties. Scenarios can then be characterised in more detail and monitored using horizon scanning (Amanatidou et al., 2012; Sutherland &
III.5 SUMMARY: THE VARIETIES OF UNCERTAINTY QUANTIFICATION – STEPHEN UNWIN

Crucial to the use of quantitative analysis in decision-making is a respect for uncertainties inherent in the methods, models, data and interpretations. While this may be intuitively self-evident across all technical communities, there is perhaps less concurrence on the means of accounting for uncertainty. Often, the prevailing philosophy is one of conservatism in which decisions are based on worst-case assumptions. But just as often, the price of excessive conservatism is too great, or consequence trade-offs make less clear what assumptions can
be deemed conservative. In such circumstances, quantitative metrics of uncertainty and their underlying analytical formalisms have substantial value. However, diverse approaches have been adopted to uncertainty quantification using methods based on fundamentally disparate concepts, uncertainty metrics, and means of analysis.

In this presentation, we provided an overview of some of the optional methods for uncertainty characterization, gave examples of their deployment, and assessed the strengths and weaknesses of each. A taxonomy of uncertainty models was first introduced comprising of:

- Qualitative/bounding methods
- Classical statistical methods
- Bayesian probabilistic methods
- Non-Bayesian quantitative methods

and each was then described.

An example of qualitative uncertainty analysis is a simple bounding approach in which model input parameters are varied over their ranges, and the corresponding ranges of the output parameters are determined. There is no weighting metric on the parameter ranges, and so the major shortcoming of such approaches resides in the limited means of consistent input/output interpretation. For example, if the mental model of a parameter input range is a 5th to 95th percentile – perhaps characterized as reasonably high confidence that the true value lies in that range – then a similar interpretation cannot be placed on the output ranges because probabilistic mechanics play no role in propagation. The upside of the method is there is no need to define and defend input probability distributions, but the price of poor output interpretability is high.

The use of statistical models to produce classical confidence intervals was discussed next, and the conclusion reached that it is a practical method only for an extremely narrow problem set; specifically, when a likelihood function is available to interpret the data, and where the model logic is sufficiently simple to allow propagation of the confidence intervals. Few real-world problems have such features.
Next, Bayesian methods were discussed, beginning with the conceptual revolution in which probability becomes interpreted as a measure of partial belief, thus allowing probabilities to be attached to general propositions and not only to statistical data. This put probabilities at the center of a theory of evidence, substantially broadening the potential domains of application, and examples of Bayesian analysis were described along with the decision/policy/regulatory questions to which they have been applied.

Next, common criticisms of Bayesianism as a theory of evidence were discussed, often reflecting the opinion that probability theory does not intuitively represent the way evidence influences belief. For example, if belief is withheld from a proposition, then in probability theory that belief must be re-assigned to a combination of competing propositions; that is, probabilities must normalize. The implication is that withholding belief from a proposition is equivalent to attaching belief to competing propositions. But is this really the way we interpret evidence? Two alternatives to probability theory were described:

- Dempster-Shafer (D-S) theory, which is a generalization of probability theory.
- Possibility theory, which is based in the algebra of fuzzy logic.

In each of these approaches, belief is represented by two distinctive measures: Support versus Plausibility in the case of D-S theory, and Necessity versus Possibility in the case of Possibility theory. The former measure in each case reflects the degree to which evidence points towards the truth of a proposition, while the latter measure reflects the degree to which evidence fails to point away from the proposition. In this way, both theories propose remedies to the argument against the probabilistic model.

Finally, the paper assessed the advantages and disadvantages of each theory of evidence, with a focus on comparison of Bayesian methods to D-S and possibility theory. While the nonprobabilistic methods were assessed to have some conceptual advantages (as outlined), on balance, we proposed that Bayesian probability theory is the most robust and defensible
approach to uncertainty characterization in real-world applications. This conclusion was supported by several factors:

- The ability to standardize/interpret probabilities through analogy to statistical frequentist concepts. D-S and possibility theory have no such analogy.
- The uniqueness of the algebra of probabilities: competing methods have no unique algebras and combinatorial rules often need to be tuned to ensure reasonable results.
- The practicality of application: non-probabilistic methods can, for complex problems, be computationally intractable.
- There has grown a wealth of methods and tools for probabilistic uncertainty analysis and associated decision-making.
- There is extensive experience with the use of probabilistic uncertainty analysis across multiple application domains – commercial, regulatory, and R&D – with the availability of standards and technical criteria for their use.

The presentation concluded that Bayesian probability theory is the most robust, transparent and defensible candidate to support uncertainty characterization in emerging models of catastrophic and existential risk.

III.6 ENVIRONMENTAL AND NATURAL DISASTERS: “SURPRISE” VERSUS “DEGRADATION” – HOLLY BUCK, PETER KAREIVA, CLAIRE HIRASHIKI

III.6.1 USING ECOLOGICAL COLLAPSES AND DISASTERS TO DEVELOP AN EVIDENCE-BASED FRAMEWORK FOR MITIGATING EXISTENTIAL ENVIRONMENTAL RISK

The only reason to create research centers dedicated to “existential risk” is to learn something that might reduce the risk. Iconic existential risk – the end of the human species – is only going to happen once, and hence we are not going to be able to forecast our extinction by studying past terminations of the human
species. This is why most discussions of existential risk come in the form of computer simulations: discussions of incredibly improbable events with serious challenges for quantifying how improbable they are, or a mix of science fiction narratives and scenario building. This is not satisfying.

There is, however, one empirical approach. While we may not be able to study the end of our species, we can examine other sorts of catastrophes and ask of them: what went wrong? The key here is selecting the sorts of catastrophes that no one would find acceptable, and for which “avoidance at any cost” would be society’s goal (as it surely would be for the extinction of our species).

To pursue this approach we focus on environmental disasters. We do this because we are environmental scientists, and because environmental disasters often do reach the sort of “magnitude” that elicits at the minimum, “avoid even if the cost of risk mitigation is high”. Secondly, although environmental catastrophes do not end the human species, they can create problems that range from migratory crises, to failed states, to massive famines, to violent conflict. These are indeed, outcomes that everyone wants to avoid. They are also rare events. The interesting feature of many types of environmental catastrophes is that societies do plan emergency responses, and societies and governments often do develop policies to reduce them. Yet they still occur. And sometimes their magnitude and impact is staggeringly unanticipated. By studying these catastrophes we can learn what system properties make them more likely; we can learn whether they are truly a surprise, or are recognized as a substantial risk but not dealt with because of failed governance. We can also ask if these catastrophes are the opposite of a surprise—but instead are almost an inevitable consequence of existing economic and geopolitical systems.

III.6.2 WHAT CLIMATE CHANGE TEACHES US ABOUT EXISTENTIAL RISK

Within the category of systemic global risks typically referred to as environmental (global warming, biodiversity loss, pollution, depletion of available fresh water, etc.), global warming or climate
disruption stands out as the one environmental risk that approaches potentially being a true existential risk. If we do not curb emissions, our current understanding of climate systems suggests a reasonable probability of massive famines, disease outbreaks, frequent battering by severe storms and floods, massive wildfires, and the great cities of the world being underwater. The interesting thing about climate change is we know all of these horrific impacts could happen and we know how to avoid them – simply stop burning fossil fuels. Yet emissions continue to grow. If global warming is to be the undoing of Homo sapiens, it reflects either a failure of government and politics, or it reflects an endpoint the planet is almost inevitably moving towards because of fundamental systemic flaws in the prevailing economic and social norms (commitment to relentless GDP growth, short-termism, and senseless consumerism). But no one would say that global warming is a surprise.

Or would they? There is one element of surprise associated with a potential global meltdown due to global warming: that is the possibility of runaway warming as a result of positive feedback loops, and connections among systems and processes that are neglected. The surprise would be this: yes we know global warming is a problem and we are taking action to address it (policies to reduce emissions, development of renewable energy, etc). But we think we have a window of opportunity on the scale of 20 to 50 years. What if we are wrong? What if the warming is accelerating, and what if the scope of damage is also growing because of interactions with other processes? What if we have only five years (as opposed to 10 or 20 years) to establish resilient communities, and to turn off emissions?

These “what if” scenarios are not matters of whimsy. Scientists have already identified several positive feedback loops in our Earth system that could accelerate global warming. Methane releases from melting permafrost could trigger a vicious cycle of further warming. Wildfires, which are exacerbated and made more frequent and severe by warming, are an especially dramatic positive feedback loop. Global warming creates more wildfires, which releases more CO2 into the atmosphere, which then creates more global warming.
California wildfires are especially interesting in this light. California is widely hailed as a state with an enlightened energy policy that is reducing emissions. Unfortunately, that narrative about California and its climate leadership is going up in smoke as a result of all these wildfires. Specifically, emissions from wildfires are undoing all the gains made by energy policy, and in fact are swamping out those gains. For example, from 2015-2016, emissions due to California’s energy demands, as well as by California’s general population dropped by some 12 million tons of CO2 (CARB, 2018). That seems like a victory. It is not an uncontested victory — because one megafire of the sort that California has seen several examples of in 2017 and 2018 typically emits well more than the 12 million tons of savings due to California’s energy policy. In other words, the surprise of mega-wildfires means that what California thought it had achieved in terms of emissions reduction simply is not there. The scary possibility is that there may be positive feedback loops we have not yet discovered. For example, as much as 30 to 40% of the world’s carbon is estimated to be in soils. If warmer temperatures somehow alter soil processes in a way that enhances emissions from soils, then that would be a whole new positive feedback loop. Climate change could also exceed all expectations in terms of the havoc it creates because of the “double whammy” effect — a climate extreme or a climate stress interacts with another agent of disaster (e.g. tornadoes) to create record-breaking damage. This is exactly what happened when wildfires in Central America, whose magnitude and spread is increased by global warming, produced tiny smoke and soot particles that changed air masses over the Gulf of Mexico, influencing convection and contributing to the deadly 2011 tornado outbreak in the southeastern US (Lefevre, 2016). This is an example of a teleconnection in the climate system (a link among meteorological or other environmental phenomena that occur over long distances). It is only in the last decade that scientists have begun to understand teleconnections, and who knows how many teleconnections remain to be discovered.

In general, we hypothesize that “existential surprises” are likely to be increasingly common because we operate in narrow disciplines, and fail to see how risk factors outside our own discipline might link to and heighten risk factors within our
discipline. For example, consider the three disciplines of civil engineering and water treatment, ecosystem science and landscape ecology, and climate science. Ecologists know that mining, over-grazing and land conversion is a global problem for rivers because the result can be excessive sediment run-off. Civil engineers know that too much sediment will force a shut-down of water treatment plants. And climate scientists know that heavy rains are increasing due to climate change. Putting all these together, severe rains that cause cities to be without water for 4 or 5 days (as recently happened in Chubut province in Argentina) should not be a surprise – but the cities were surprised, and had no back-up plans.

Meanwhile the frequency and intensity of heatwaves has been growing. Heat waves are a climate feature that potentially touches every individual. The surprise occurs in how people respond. Many scholars would have predicted a link between heat waves and human behavior. But who could have expected that riots following a heat wave would include attacks on electricity substations as happened in a 2014 heat wave in India? (Banerjee, 2014). We should expect the list of surprising links between climate stress and other risk factors (or double whammies) to only grow.

That climate change is a threat multiplier — with cascades of entangled social and natural impacts — is widely accepted. Recently, scientists, scholars, and journalists have also begun to speak about climate change as a potential existential risk. One example from the grey literature is a report entitled *What Lies Beneath?: The Understatement of Existential Climate Risk*, published by an Australian think tank. The forward to this report, which is written by the climate scientist Hans Joachim Schellnhuber, makes the point that the Intergovernmental Panel on Climate Change (IPCC) has been focused on *probabilities*, which are incredibly difficult to estimate within even an order of magnitude when it comes to rare events. Schellnhuber suggests more attention should be given to evaluating potential risky *possibilities*. A 2018 high-profile article in PNAS, the so-called “Hothouse Earth” paper by Will Steffen and colleagues, does just this, reviewing the literature to identify potential biogeophysical feedbacks and tipping cascades. They point out that there is a
pathway to avoid these risks, which they call Stabilized Earth, and which requires a coordinated, deliberate effort to build resilience, with the following features: “(i) maintenance of diversity, modularity, and redundancy; (ii) management of connectivity, openness, slow variables, and feedbacks; (iii) understanding social–ecological systems as complex adaptive systems, especially at the level of the Earth System as a whole; (iv) encouraging learning and experimentation; and (v) broadening of participation and building of trust to promote polycentric governance systems.”

### III.6.3 FRAMEWORKS FOR DEALING WITH POSITIVE FEEDBACK LOOPS AND DOUBLE WHAMMIES

Given Steffen et al’s five-point recipe for addressing climate change (and many other similar recipes), it is fair to conclude we know what needs to be done to avert climate disaster. However, even with increasing attention given to the potential existential risk posed by global warming and related environmental catastrophes, governments and the global community remain blithely paralyzed. For example, there is still no early warning system for keeping watch on potential tipping points. The key question is: Can we do better at anticipating environmental existential risks, including climate risks? We approach this question by asking: Have we in fact done better in the past in anticipating and alleviating disaster and environmental risks?

We proceed in three parts. First, we compare the top ten environmental disasters to a random selection of environmental disasters – all drawn from a common global database. We ask if there are any distinctive features of the massive disasters compared to the general population of disasters. Second, we look at species extinctions; and thirdly, we examine ecosystem collapses. Implicit in this exercise is the idea that massive disasters, species extinctions, and the collapse of entire ecosystems are all singularly drastic outcomes that nations and societies everywhere would like to avoid. To be clear, we do not argue that local disasters or extinctions and ecosystem collapses “add up” to a global existential risk — rather, we simply believe that there are insights to be gained by examining what seems to
push a species, an ecosystem, or a disaster across some critical threshold.

In analyzing these three analogues, we are particularly interested in whether the disasters, extinctions, and ecosystem collapses were surprises. Surprises can take the form of risks we are unaware of, i.e. “unknown unknowns”, or in the parlance of Shackle, “unexpected events” which had never been imagined (in Bier et al, 1999). Another form of surprise is what Shackle calls “counterexpected events”, or events that had been imagined but regarded as impossible (Bier et al, 1999). And a third form of surprise is the “double whammy”, or a surprise that comes from combinations of great technological uncertainties that yield shocking possibilities. Positive feedback loops play into both of these forms of surprises.

**III.6.3.1 ANALOG 1: DISASTERS**

For this analysis, we drew upon the EM-DAT database, developed by the Centre for Research on the Epidemiology of Disasters in collaboration with: the International Federation of Red Cross and Red Crescent Societies (IFRC); the United Nations International Strategy for Disaster Reduction (UNISDR); and the U.S. Agency for International Development (USAID). Any disaster that fulfills any of the following criteria is input into the database: a) ten or more people are reported killed, b) one hundred or more people are reported affected, c) there is a declaration of a state of emergency, and/or d) there is a call for international assistance. The data has high granularity, even down to the latitude and longitude for specific location events.

Free access to the EM-DAT database is limited to a download of 8000 rows, so we needed to pare down our dataset. Although the data spans from 1900 to 2018, we decided to choose a 10-year period from 2007 to 2017, in order to account for climatological changes. Of the categories provided, we selected epidemic, climatological (incl. drought and wildfire), hydrological (incl. flood and landslide), meteorological (incl. extreme temperature, fog, and storm), and technological, since we felt that these categories best described environmental disasters that are induced or exacerbated by human activity on the environment.
We selected the top ten disasters based on fatality count (Table 1), and compared this to a random sample of 100 from our 2007 to 2017 data subset (Table 2). For each randomly drawn disaster, we required a certain minimum amount of coverage in the news or scientific literature. This allowed us the flexibility to drop any randomly drawn disaster (and draw again from the data base) that lacked significant information and focus on disasters with substantial evidence for the study.

**Table 1.** Top ten disasters by death toll, 2007-2017.

<table>
<thead>
<tr>
<th>Location</th>
<th>Date</th>
<th>Disaster Type</th>
<th>Total Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haiti</td>
<td>2010</td>
<td>Earthquake (7 Richter)</td>
<td>222,570</td>
</tr>
<tr>
<td>Myanmar</td>
<td>2008</td>
<td>Tropical cyclone (215 kph)</td>
<td>138,366</td>
</tr>
<tr>
<td>China</td>
<td>2008</td>
<td>Earthquake (7.9 Richter)</td>
<td>87,476</td>
</tr>
<tr>
<td>The Russian Federation</td>
<td>2010</td>
<td>Heat wave (40 °C)</td>
<td>55,736</td>
</tr>
<tr>
<td>Somalia</td>
<td>2010</td>
<td>Drought</td>
<td>20,000</td>
</tr>
<tr>
<td>Japan</td>
<td>2011</td>
<td>Earthquake (9 Richter)</td>
<td>19,846</td>
</tr>
<tr>
<td>Nepal</td>
<td>2015</td>
<td>Earthquake (7.8 Richter)</td>
<td>8,831</td>
</tr>
<tr>
<td>The Philippines</td>
<td>2013</td>
<td>Tropical cyclone (315 kph)</td>
<td>7,354</td>
</tr>
<tr>
<td>Haiti</td>
<td>2010</td>
<td>Bacterial disease (vaccinated)</td>
<td>6,908</td>
</tr>
<tr>
<td>India</td>
<td>2013</td>
<td>Riverine flood (131,743.41 km²)</td>
<td>6,054</td>
</tr>
</tbody>
</table>

**Table 2.** Randomly drawn disasters, 2007-2017.
<table>
<thead>
<tr>
<th>Location</th>
<th>Date</th>
<th>Disaster Type</th>
<th>Total Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chile</td>
<td>2011</td>
<td>Transport accident</td>
<td>21</td>
</tr>
<tr>
<td>Martinique</td>
<td>2011</td>
<td>Tropical cyclone</td>
<td>1</td>
</tr>
<tr>
<td>Haiti</td>
<td>2013</td>
<td>Riverine flood</td>
<td>6</td>
</tr>
<tr>
<td>Chile</td>
<td>2015</td>
<td>Flash flood (154,773.2 km²)</td>
<td>178</td>
</tr>
<tr>
<td>Ukraine</td>
<td>2016</td>
<td>Flood (19,889.49 km²)</td>
<td>3</td>
</tr>
<tr>
<td>Taiwan</td>
<td>2013</td>
<td>Tropical cyclone (220 kph)</td>
<td>3</td>
</tr>
<tr>
<td>The Democratic Republic of the Congo</td>
<td>2009</td>
<td>Viral disease</td>
<td>17</td>
</tr>
<tr>
<td>Brazil</td>
<td>2009</td>
<td>Riverine flood</td>
<td>0</td>
</tr>
<tr>
<td>Togo</td>
<td>2017</td>
<td>Flood</td>
<td>0</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>2015</td>
<td>Landslide</td>
<td>52</td>
</tr>
</tbody>
</table>

Next, we analyzed the factors that went into these disasters, and categorized them into the following:

A. **Unsolvables:** Event is not a surprise, but we simply do not know what to do about it, or how to reduce it

B. **Failures:** We know how to reduce the risk, but we do nothing

C. **Surprises:** We are caught off-guard

Of the top ten disasters ranked by death toll, just one was
arguably a surprise: the cholera epidemic in Haiti, following the Jan. 12, 2010 earthquake, a 7.0 quake that hit 25km from the capital, Port-Au-Prince. This killed about 300,000, and displaced a million, damaging economic, administrative, and health care infrastructure. Conditions were dire before the earthquake, with half of people living in abject poverty (less than $1 per day), chronic malnutrition afflicting between 18-30% of children, and 4 out of 5 people lacking access to drinking water (Echevin, 2011). After the earthquake, NGOs flooded into Haiti, many lacking skillsets. Response was slow and arduous, and several months later, a cholera outbreak occurred. The outbreak continued for years, with 800,000 eventual cases and around 9,000 deaths. It was a surprise, as cholera had not been seen in the region for over a century, and there was a lack of institutional knowledge about how to deal with it — the source of the outbreak was traced back to UN peacekeepers years later. Aside from this disaster, the other nine of the top ten disasters could be characterized as failures. They were characterized by a high death toll due to inexperience or poor emergency preparedness, typically as a result of governmental negligence.

The randomly chosen sample, however, was characterized by mainly “unsolvables”: these disasters were the result of accepted, understood hazards that were sufficiently addressed by basic risk preparation. For non-technological events, there were fewer deaths due to less extreme weather events and/or sufficient emergency response given the greater predictability of the events. From this, we learn that the worst disasters aren’t unsolvables, but genuine failures of something: governance, anticipation, preparedness, institutions, etc. And true surprises are relatively rare.

III.6.3.2 ANALOG 2: EXTINCTIONS

When a species ends, is it a surprise? Our categories, again, were (A) Unsolvables: We can see the extinction coming, but just do not have a practical solution that we can see working; (B) Failures: There is a practical solution, but extinction seems inevitable as a failure of leadership, governance, and society; and (C) Surprises: extinctions that no one saw coming. For this initial exploration, we chose a few illustrative examples chosen from the
twentieth century and beyond, which is when we start to have reliable information about the causes of extinctions.

**Heath hen:** This species' initial decline was due to habitat loss and hunting. When people realized it needed protection, a 1600 acre reserve was established on Martha’s Vineyard, and this seemed to initially be a success: the population grew from 50 in 1907 to 2000 in 1915. However, a cascade of challenges befell the heath hen starting in 1916. A wildfire in 1916 burned much of the vegetation and the habitat the heath hen relied on; in 1917, an extraordinarily cold winter and an invasion of predatory goshawks caused a spike in heath hen mortality; and in 1920, a poultry disease from turkeys swept through the population, knocking it down to 13 birds. Each of these events were low probability events. Having them all happen within five years of one another, when they are in fact independent events, was highly unlikely and would never have been predicted. If they had been spaced out over a much longer time period, there might have been opportunities for the heath hen population to recover before getting hit again. Once the population hit 13, the species’ fate was sealed due to inbreeding.

**African elephant:** While not yet extinct, this fate seems likely, in spite of significant resources devoted to saving it. The African elephant’s initial decline was due to habitat loss and hunting. Now, it is protected, but criminal ivory poaching is slaughtering elephants. Between 2007-2014, 30% of savanna species and 65% of forest species was wiped out. Measures aimed at stopping illegal ivory poaching include demand reduction in Asia, incentivizing communities to steward populations, arming anti-poaching patrols, and controlling ivory movement via UN action (CITES). Yet in spite of this, extinction still looms. This is a case where we know what needs to be done – stop poaching – but do not know how to do so.

**Vaquita:** This little porpoise lives in one region: the Gulf of Mexico. Between 2008 and 2015, 80% of the world’s vaquita vanished. The decline is due to by-catch, as the little porpoise is netted when fishing for Totoba. The initial policy response to the vaquita decline was to ban Totoba fishing in areas where there were vaquita. Because Totoba was not a major commercial
fishery this should have worked. However a surprising teleconnection between a related fish off the coast of China and the Totoba doomed the Vaquita. Specifically, a different but evolutionarily related fish species off the coast of China had been harvested for centuries by the Chinese because of presumed medicinal properties of its fish bladder. The local fish population off the coast of China was so depleted it could not meet the demand for fish bladders. Somehow (it is not known how) it was discovered that the Totoba in the Gulf of Mexico had fish bladders that could replace the Chinese counterpart and the merchants shifted to Totoba fish bladders. These fish bladders garner extraordinary prices in China—as much as $100,000 for a single large bladder (the price increases with bladder size). What the Mexican fishery managers thought was a fish worth only a few dollars per fish surprisingly turned out to be an exotic commodity harvested for rare large fish bladders, and hence unsuited to a fishing ban (Bessessen, 2018). One hundred thousand dollars swimming in body of a marine fish is going to attract a lot of individuals willing to break the law, and that is what has happened—rampant illegal harvest of Totoba, supposedly conducted by Mexican cartels as opposed to commercial fishers.

**Passenger pigeon:** Once the most abundant bird on the planet, the passenger pigeon suffered a double-whammy. They were hunted and their habitat was destroyed. Those twin threats drove down the numbers of hundreds of species in North America, but did not cause their extinction. But the Passenger Pigeon was unique in having an unusual social breeding system whereby hundreds of nests would congregate in one tree (presumably as an anti-predation strategy). In contrast, other pigeon species would typically have one nest per tree. As passenger pigeon numbers declined, there were not enough birds to form dense congregations of nests, and breeding failure ensued.

Looking at examples of the heath hen, African elephants, the vaquita, and passenger pigeons, we found the following: extinctions were generally not due to just one thing, that criminal behavior is hard to overcome, and that a highly unlikely run of bad luck can undermine what might seem to be successful effort. In modern times, rarely is the failure to prevent extinction a reflection simply of governance or social failure. The question this brings
III.6.3.3 ANALOG 3: ECOSYSTEM COLLAPSES

For the final analysis, we used a dataset maintained by the International Union for Conservation of Nature (IUCN), called the Red List of Ecosystems. The IUCN’s hypothesis is that ecosystem risk is a function of the species that compose them, their interactions, and the ecological processes they depend on. Their criteria for assessing the risk of ecosystem collapse include: (A) Reduction in geographic distribution (loss of area); (B) Restricted geographic distribution (narrow habitat range); (C) Environmental degradation; (D) Disruption of biotic processes or interactions; (E) Quantitative analysis that estimates the probability of ecosystem collapse. We looked at ecosystems categorized as Critically Endangered, Endangered, and Vulnerable, and found that for most of these ecosystems, multiple factors contributed to their decline (Figure 1).

**Threatened Categories:** CR – Critically Endangered, EN – Endangered, VU – Vulnerable

**Figure 1. Number of factors contributing to vulnerability in threatened ecosystems.**
Of the 12 global ecosystems that have been analyzed by the IUCN, the Aral Sea is the only ecosystem that is categorized as collapsed. In the IUCN’s documentation for the scientific foundations of the Red List, there are at least eight socio-ecological processes mentioned as playing into this collapse (Keith, 2013). These include (1) an increase in human population increased, causing more water withdrawal; (2) irrigation expanded to desert, causing more withdrawal; (3) a warmer climate means more evapotranspiration; (4) irrigation increased salinity of inflows; (5) non-native species altered native biota; (6) food system impacts for humans due to collapsed fisheries; (7) dust impacts from exposed lakebed; (8) economic decline, which becomes a positive feedback in that it limits the ability to deal with ecological deterioration.

What a first glance at this analog reveals is that multiple factors play into ecosystem collapse — the multiple whammy. As Bier and colleagues write, when system behavior is nonlinear, additive or synergistic, the concatenation of non-extreme events can yield an extreme outcome. They give the example of normal flooding being escalated to extreme flooding due to non-extreme events like soil saturation and heavy rains piling up (1999). A multiple-ecosystem collapse, i.e. an existential ecosystem risk like climate change, may also be made of multiple factors that would be non-extreme when taken in isolation.

III.6.4 RETOOLING OUR THINKING TO ANTICIPATE MULTIPLE WHAMMIES

Together, these three analogies stress the importance of anticipating and dealing with multiple whammies. Many analysts of existential risks have also pointed to these problems, identifying “cross-systemic” dimensions of existential risks (Matus); “intermediate” existential risks (Circkovic et al, 2014); properties of emergence within complex systems (Centeno et al, 2015); “multidimensional risk scenarios” that can look at mitigation failures, global spread mechanisms, and critical system breaches (Avin et al 2018). A second feature that emerges from the three environmental catastrophe analogies that we probed is outright failure: failure due to criminal activity, failure due to ignorance or government malfunction.
Finally, while it does not emerge from an examination of any specific catastrophe, the repeated catastrophes suggest that there could be some “organized irresponsibility” at play. This notion is attributed to the sociologist of risk Ulrich Beck, who described how the old routines of decision, control, and production can facilitate ecological degradation (1996). “Concretely, it is not rule-breaking but the rules themselves which ‘normalize’ the death of species, rivers or lakes,” Beck writes, and this is what he called “organized irresponsibility”: “For dangers are being produced by industry, externalized by economics, individualized by the legal system, legitimized by the natural sciences and made to appear harmless by politics” (1996). An example of organized irresponsibility is the way that assessing the risk of extinction, and the bureaucratic superstructure that carries out these assessments, normalizes the phenomena of vanishing species.

Critiquing society and norms, and calling out narrow siloed thinking, is easy. The challenge is coming up with mechanisms and a framework that could help avert existential disasters.

We make three concrete suggestions. First, teleconnections, double whammies, and positive feedback loops emerge consistently as contributors to disasters. It is normal practice in risk analysis to identify specific classes of existential risk (nuclear war, nuclear terrorism, meteorite impact, etc.), and make some attempt to estimate long-tailed risks. In examining any particular “existential catastrophe” or “extreme disaster” there should be a systematic examination of all possible interacting processes that could exacerbate or make more likely any given catastrophe. Much more will be gained by thoroughly considering all possible interacting processes than is gained by trying to estimate probabilities of singular events. If only the relief response in Haiti had thought about the possibility of disease epidemics. If only those who set up the Heath hen reserve had the idea of setting up more than one reserve in very different places (to spread the risk). When the IUCN reviews the level of extinction threat for species, teams of scientists look at all possible threats to that species. When we examine existential risks, we should consider all possible interacting processes.
Second, the absence of tipping-point monitoring systems is inexcusable. We do not need to know what the exact tipping point is to draw value from monitoring systems. There is a rich theory of tipping points that indicates increased temporal variability commonly precedes a tipping point (see Ashwin, 2012; Carpenter and Brock, 2015; Dakos, 2015; Moore, 2018). Monitoring could detect this. The amount of information available from global satellites has not been tapped and should be. Data from household surveys are currently used in FEWS (Famine Early Warning System) (Braimoh et al, 2018), and there is no reason such data could not address disasters other than famines. There seems to be a misconception that because we cannot agree on what the tipping points are, we need not invest in monitoring for tipping points. The opposite is the case: precisely because we do not know what tipping points exist, we need these monitoring systems of social, ecological, and physical variables.

Finally, the third suggestion concerns systemic failures and irresponsibility. This must be addressed in a serious manner, and not simply as an academic exercise in critique. In order to assess just how much risk is due to systemic flaws, one could study a large population of cities around the world, and examine differences among those cities in terms of magnitude of catastrophes, and recovery from those catastrophes. Cities differ in governance, social norms, degree of community cohesion, commitment to social welfare, and so on. If the cities display different catastrophe regimes and catastrophe recoveries (and they will), it may be possible to estimate the role of systemic failure and irresponsibility. Once one has an estimate of the “systemic failure” effect, there could be a discussion of how to address it. Without such an estimate, there will be weak motivation to seriously reconsider our economic, governance and social systems. The research opportunities are vast, and the value of this research to our species survival cannot be overstated.

This article and this volume is about risk -- albeit it, a very unique form of risk. The science of risk analysis inevitably is drawn to estimation of the probabilities of events or failures. This endeavor is worthwhile for standard risk analyses (e.g., the probability of a nuclear power plant having an accident). But when it comes to
existential risk, for which probabilities are incredibly tiny, it is hard to imagine being able to make a cogent argument for a risk of one in a million versus one in two million chance of the apocalypse. But what we can discover through research is what actions might reduce risk, even if we do not know by how much. And, recognizing the importance of double whammies, teleconnections, positive feedback loops, and ingrained system flaws -- we can invest in research that illuminates the many different possible pathways to an existential crisis. That research will need to be challengingly unbounded, in the sense that if it is nuclear war you are worried about, you may need to also examine droughts, famines, international governance, mass human migrations, and the dominance of corporations in setting global trade agreements. Finally, we hypothesize that global climate disruption is proceeding at such a surprising pace, with impacts that potentially shape every facet of human societies, and should be viewed as the "great agitator" of existential risk in ways we are only just beginning to uncover.

REFERENCES


FEWS NET. 2018. FEWS NET. Available from: http://fews.net/


LeFevre, K. 2016. The Power of Particles. Earthdata, Available at: https://earthdata.nasa.gov/user-resources/sensing-our-planet/the-power-of-particles


B. JOHN GARRICK

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.

SETH BAUM

Dr. Seth Baum is Executive Director of the Global Catastrophic Risk Institute, a nonprofit think tank that he 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 MS in Electrical Engineering from Northeastern University and a PhD in Geography from Pennsylvania State University. He then completed a post-doctoral fellowship with the Columbia University Center for Research on Environmental Decisions.

MIHAI A. DIACONEASA

Mihai A. Diaconeasa, Ph.D., Mechanical Engineering, UCLA; M.S. Nuclear Engineering, MIT, is currently a Postdoctoral Research Scholar at UCLA and Research Staff at the UCLA B. John Garrick Institute for the Risk Sciences. He is also involved in the institute's collaborations with the Southern Company, Pacific Gas and Electric Company (PG&E), and Japan's Nuclear Regulation Authority (NRA) on probabilistic risk assessment projects for nuclear power plants. Moreover, he is involved in the institute’s multiple projects on the reliability and resilience assessment of space systems for extended deep space missions with NASA's Jet Propulsion Laboratory (JPL) and NASA Ames.
THERESA STEWART

Theresa Stewart is a graduate student in Materials Engineering who completed her undergraduate education at Cal Poly San Luis Obispo. At UCLA her focus is on model-based systems engineering. She is working on system reliability modeling for the JPL “Lunar Flashlight” satellite.

JULIUS WEITZDÖRFER

Dr. Julius Weitzdörfer is the Director of Studies in Law and a Charles & Katharine Darwin Research Fellow at Darwin College Cambridge, and an Affiliated Lecturer teaching Environmental and Sustainable Development Law at the Faculty of Law, University of Cambridge. He is a full-time Research Associate at the Centre for the Study of Existential Risk (CSER), where he works on the law and governance of extreme technological risks, with a special focus on Japan.

STEPHEN D. UNWIN

Dr. Unwin’s career has centered on the development of uncertainty and risk-analytic methodologies, their application to multi-domain problems of national and commercial importance, and the founding of businesses on those capabilities. He has developed methods and models for risk-informed decision-making that continue to be applied in numerous sectors, including nuclear energy, oil & gas, power grid infrastructure, renewable energy, national security, climate adaptation, the chemical process industries, and the fossil energy sector. Before joining Pacific Northwest National Laboratory (PNNL) in 2006 he founded Brookhaven National Laboratory’s Safety Integration Group, SAIC’s Risk & Reliability Management Division, Battelle’s Integrated Risk Management Group, and Unwin Company – Integrated Risk Management, which is a continuing risk management resource to commercial and government clients. He was an author of the US Nuclear Regulatory Commission’s NUREG-1150 study – a landmark in risk methodology development, he co-led PNNL’s Technosocial Predictive Analytics Initiative, and has contributed substantially to the international literature on risk/uncertainty methods and applications. He holds a bachelor’s degree in physics from Imperial College, London and a doctorate in theoretical physics from the University of Manchester, UK. In his current role he oversees risk management programs in PNNL’s Energy & Environment Directorate.

HOLLY J. BUCK

Holly Jean Buck is a NatureNet Science Fellow at UCLA’s Institute of the Environment and Sustainability. She’s interested in how communities can be involved in the design of emerging environmental technologies. She works at the interface of environmental
sociology, international development, and science and technology studies. Her diverse research interests include agroecology and carbon farming, new energy technologies, artificial intelligence, and the restoration of California’s Salton Sea. At present, she is studying the socio-political feasibility of using solar geoengineering to scale up carbon removal. She has written on several aspects of climate engineering, including humanitarian and development approaches to geoengineering, gender considerations, and human rights issues. Currently, she is focused on how policy for large-scale carbon removal can better address the social implications. Prior to her academic life, she has worked as a foreign affairs analyst, a geospatial technician for remote sensing company, and a creative writing teacher. She holds a doctorate in Development Sociology from Cornell University and a MSc in Human Ecology from Lund University, Sweden.
<table>
<thead>
<tr>
<th>Name</th>
<th>Role</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apostolakis, George</td>
<td>Participant</td>
<td>Head, Nuclear Risk Research Center, Japan; former Commissioner, USNRC</td>
</tr>
<tr>
<td>Baum, Seth</td>
<td>Co-Sponsor Presenter</td>
<td>Executive Director, Global Catastrophic Risk Institute</td>
</tr>
<tr>
<td>Bier, Vicki</td>
<td>Participant</td>
<td>Professor, University of Wisconsin</td>
</tr>
<tr>
<td>Buck, Holly</td>
<td>Presenter</td>
<td>NatureNet Postdoctoral Fellow, Institute of the Environment and Sustainability, UCLA</td>
</tr>
<tr>
<td>Budnitz, Bob</td>
<td>Participant</td>
<td>Lawrence Berkeley National Laboratory, University of California, Berkeley</td>
</tr>
<tr>
<td>Burton, Henry</td>
<td>Participant</td>
<td>Assistant Professor, Civil and Environmental Engineering, UCLA</td>
</tr>
<tr>
<td>Carnesale, Al</td>
<td>Participant</td>
<td>Professor; Chancellor Emeritus, UCLA</td>
</tr>
<tr>
<td>Dhir, Vijay</td>
<td>Participant</td>
<td>Professor, UCLA; former Dean, UCLA Engineering</td>
</tr>
<tr>
<td>Diaconeasa, Mihai A.</td>
<td>Presenter</td>
<td>Postdoctoral Research Scholar, UCLA; Research Staff, Garrick Institute</td>
</tr>
<tr>
<td>Duettmann, Allison</td>
<td>Participant</td>
<td>Research Staff, Foresight Institute</td>
</tr>
<tr>
<td>Garrick, B. John</td>
<td>Co-Sponsor Presenter</td>
<td>Adjunct Professor, UCLA; Founder, Garrick Institute</td>
</tr>
<tr>
<td>Gartzke, Erik</td>
<td>Participant</td>
<td>Professor, Political Science and Director of the Center for Peace and Security Studies, University of California, San Diego</td>
</tr>
<tr>
<td>Hamrick, Barbara</td>
<td>Correspondent Participant</td>
<td>Radiation Safety Officer, University of California, Irvine</td>
</tr>
<tr>
<td>Hirashiki, Claire</td>
<td>Participant</td>
<td>Data Analyst, California Center for Sustainable Communities, UCLA</td>
</tr>
<tr>
<td>Johnson, David</td>
<td>Participant</td>
<td>Consultant, Garrick Institute Affiliate</td>
</tr>
<tr>
<td>Name</td>
<td>Role</td>
<td>Institution/Position</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------</td>
<td>---------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Kao, Tsu-Mu</td>
<td>Participant</td>
<td>Senior Research Fellow, Institute of Nuclear Energy Research, Taiwan</td>
</tr>
<tr>
<td>Katona, Peter</td>
<td>Participant</td>
<td>Professor, Los Angeles Geffen School of Medicine</td>
</tr>
<tr>
<td>Martins, Marcelo Ramos</td>
<td>Participant</td>
<td>Professor, University of São Paulo, Naval Architecture and Ocean Engineering Department, Analysis, Evaluation and Risk Management Laboratory</td>
</tr>
<tr>
<td>Miller, Mark</td>
<td>Participant</td>
<td>Chief Scientist, Agoric; Senior Fellow, Foresight Institute</td>
</tr>
<tr>
<td>Mosleh, Ali</td>
<td>Participant</td>
<td>Professor, UCLA; Director, Garrick Institute</td>
</tr>
<tr>
<td>North, Warner</td>
<td>Participant</td>
<td>President, Northworks</td>
</tr>
<tr>
<td>Peterson, Christine</td>
<td>Participant</td>
<td>Co-Founder and Senior Fellow, Foresight Institute</td>
</tr>
<tr>
<td>Sandberg, Anders</td>
<td>Participant</td>
<td>Senior Research Fellow, University of Oxford</td>
</tr>
<tr>
<td>Sokolowski, Lukasz</td>
<td>Participant</td>
<td>Thermal-Hydraulics Specialist, Risk Analysis, Kiwa Inspecta, Sweden</td>
</tr>
<tr>
<td>Stewart, Jonathan</td>
<td>Participant</td>
<td>Professor, Civil and Environmental Engineering, UCLA</td>
</tr>
<tr>
<td>Stewart, Theresa</td>
<td>Presenter</td>
<td>Graduate Student, Materials Science and Engineering, UCLA</td>
</tr>
<tr>
<td>Stiller, Jan C.</td>
<td>Participant</td>
<td>Senior PSA Expert, Gesellschaft für Anlagen-und Reaktorsicherheit (GRS)</td>
</tr>
<tr>
<td>Torres, Phil</td>
<td>Participant</td>
<td>Affiliate Scholar, Institute for Ethics and Emerging Technologies</td>
</tr>
<tr>
<td>Unwin, Stephen</td>
<td>Presenter</td>
<td>Manager, Nuclear Sector, Pacific Northwest National Laboratory</td>
</tr>
<tr>
<td>Weitzdörfer, Julius</td>
<td>Co-Sponsor Presenter</td>
<td>Affiliated Lecturer at the Faculty of Law, Director Studies on Law at Darwin College and the 2014 Charles &amp; Katharine Darwin Research Fellow</td>
</tr>
<tr>
<td>Youngblood, Bob</td>
<td>Participant</td>
<td>Senior Risk Consultant, Risk Assessment and Management Services, Idaho National Laboratory</td>
</tr>
</tbody>
</table>