Towards an Integrated Assessment of Global Catastrophic Risk

Seth D. Baum and Anthony M. Barrett
Global Catastrophic Risk Institute
http://sethbaum.com * http://tony-barrett.com * http://gcrinstitute.org

Published in B.J. Garrick (Editor), *Proceedings of the First International Colloquium on Catastrophic and Existential Risk*, Garrick Institute for the Risk Sciences, University of California, Los Angeles, pages 41-62.
This version 17 January 2018.

Introduction

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

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

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

Ethics

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

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

In Equation 1, EV(a) is the expected value of an action a that an actor (individual, institution, etc.) could take; {c} is the set of possible consequences of a; P(c) is the probability of consequence c; and V(c,s,t) is the value of consequence c at spatial point s and temporal point t, which is integrated across all points in space and time. V(c,s,t) is in turn defined as:

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

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

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

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

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

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

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

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

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

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

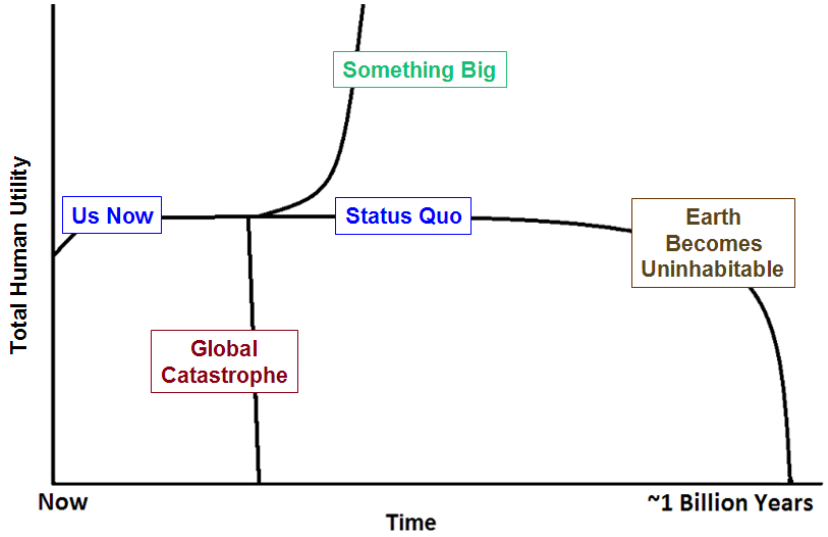

Figure 1. Possible long-term trajectories for human civilization. Adapted from Maher and Baum (2013).

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

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

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

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

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

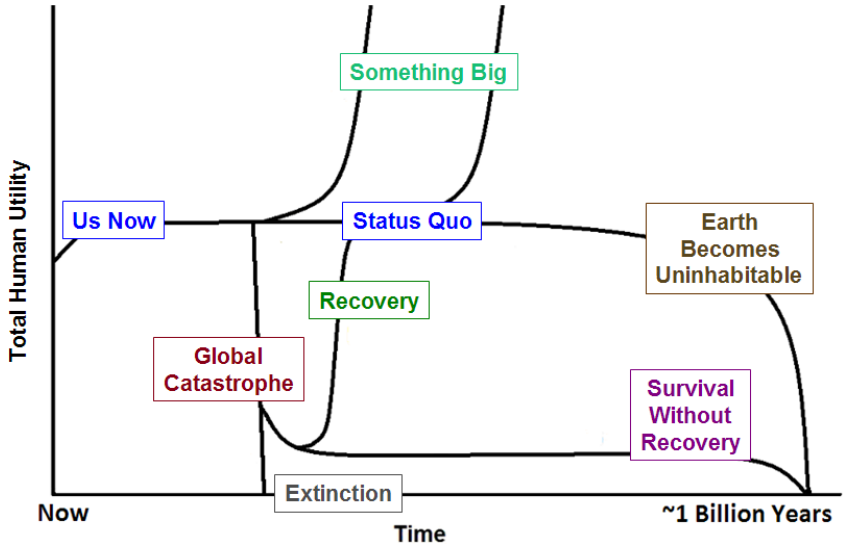

Figure 2. Possible long-term trajectories for human civilization showing different types of global catastrophe. Adapted from Maher and Baum (2013).

Prior Literature

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

After World War II and especially with the buildup of nuclear arsenals, attention went to the prospect of nuclear war. It was commonly believed that a nuclear war with the large arsenals of the day would result in global catastrophe and possibly even human extinction. This led to some novel policy debates. One point of contention was the idea that it would be better to let the other side of the Cold War win than to let nuclear war end humanity. This debate took place in particular between philosophers Sidney Hook and Bertrand Russell under the catchphrase "better red than dead" (Russell 1958a; 1958b; Hook 1958a; 1958b).

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

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

Another line of inquiry into GCR began during the 1970s with the rise of concern about environmental issues. This gave rise to an economics literature on environmental catastrophe (e.g., Cropper 1976), which later led to literatures on the economics of catastrophic climate change (e.g., Gjerde et al. 1999) and on global catastrophes in general (e.g., Martin and Pindyck 2015). This economics literature brought a mathematical sophistication to the analysis of GCR, while continuing to emphasize issues of future generations, discounting, and significance for policy and decision making. However, the economics literature provides a rather crude treatment of the future, consisting mainly of simple mathematical assumptions extrapolated into the distant future with little regard for empirical considerations about what the future might actually look like.

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

One common confusion in the GCR literature is to underestimate the importance of smaller catastrophes. An extreme case of this confusion is found in a much-cited passage of Parfit (1984, p.453-454) that argues that human extinction is vastly more important than catastrophes killing 99% of the population, and indeed that the difference between extinction and 99% is much larger than the difference between 99% and 0 (i.e., no catastrophe). The problem with this logic is that it assumes that the surviving 1% would quickly recover back up to the status quo no-catastrophe

state with no long-term loss in utility. However, as Figure 2 illustrates, this assumption does not necessarily hold, and indeed there is reason to believe that it often will not hold, in which case a 99% catastrophe could be of comparable loss as human extinction.

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

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

Terminology and Definitions

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

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

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

At present, the two terms in widest use are "global catastrophe" and "existential catastrophe". A shortcoming of the term "existential catastrophe" is that it implies some sort of loss of existence, which could be the loss of the human species (i.e., human extinction) or the loss of human civilization. (The term is also found in other contexts, for example in business in reference to corporations that take on enough financial risk to threaten their ongoing solvency.) However, recalling Figure 2 and the surrounding discussion, what ultimately matters is not the existence of the species or the civilization but instead the long-term trajectory. Indeed, Bostrom (2002) defines existential catastrophe as an event that causes human extinction or permanently

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

The term "global catastrophe" does not suffer from the same semantic problem. The words can readily refer to the full range of catastrophes one might care about as per Figure 2. However, the term "global" is a spatial term that on its own does not capture the important temporal dimension of the consequences of catastrophes. Additionally, there is no clear threshold for what makes a catastrophe global. Even small catastrophes can be global—for example, a terrorist attack at a tourist venue killing one tourist from each continent is catastrophic to the deceased and their families across the globe. The GCR literature has assumed a higher severity for global catastrophe. Atkinson (1999) defines global catastrophe as an event in which at least one quarter of the human population dies; Bostrom and Ćirković (2008) set a minimum threshold for global catastrophe in the range of 10^4 to 10^7 deaths or $\$10^9$ to $\$10^{12}$ in damages. But these thresholds are arbitrary and do not signify any deeper reason for concern. Baum and Handoh (2014) define global catastrophe as an event that exceeds the resilience of the global human system, resulting in a significant undesirable state change. This is a more meaningful definition, though it does not speak to long-term effects.

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

Integrated Assessment

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

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

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

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

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

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

Risk Analysis

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

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

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

Identifying the GCRs is relatively straightforward; and the standard tools of risk analysis offer promise for analyzing them (Garrick 2008), but fully quantifying them is not so easy. The GCRs are large, complex, and unprecedented, making for an unusually difficult risk analysis challenge (Baum and Barrett 2017).

Asteroid Collision

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

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

An important limitation of Reinhardt et al. (2016) is that it uses a physical definition of event severity: the amount of energy released. The same limitation applies to many other asteroid risk analyses and analyses of other GCRs. (For elaboration in the context of environmental GCRs, see Baum and Handoh 2014.) However, recalling the above discussion of ethics, what matters is not the physical severity but the human impacts.

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

Artificial Superintelligence Takeover

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

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

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

The fault tree contains two main branches:

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

(2) The ASI uses its capacity for takeover in a way that results in catastrophe. This occurs if three further subconditions all hold: (2a) humans fail in any attempts to design the goals of the ASI to not cause catastrophe, (2b) the ASI does not set its own goals to something that does not cause catastrophe, and (2c) the ASI is not deterred in carrying out its goals, whether by (2c1) humans, to the extent that human actions might be able to deter an ASI, (2c2) another AI, including another ASI if this ASI is not the first, or (2c3) something else.

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

While the model can also be used to quantify risk parameters as well as the total risk, such quantifications will often be uncertain due to the inherent ambiguity of ASI risk. This ambiguity poses a challenge for attempts to calculate optimal decision portfolios for minimizing GCR, such

as in the knapsack problem described above. However, some of this challenge is attenuated by the details of the decision options themselves, to which the paper now turns.

Risk Reduction In Research

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

Table 3 lists some examples of actions that can reduce risk for each of the four GCR categories that were introduced in Table 2. These actions show the value of grouping the GCRs into these categories: the same actions are often applicable across multiple GCRs within the same category:

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

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

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

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

GCR Category	Examples of GCR Reduction Actions
Environmental change	Clean energy, clean agriculture
Technology disasters	Safe technology designs, technology governance
Large-scale violence	Arms control, improved international relations
Natural disasters	Disaster prevention, societal resilience

Table 3. Examples of GCR reduction actions for each of the four GCR categories.

Risk-Risk Synergies: Societal Resilience

The risk reduction action of increasing societal resilience is an important one and worth discussing in further detail. It was brought up in the context of natural disaster risk, but it is applicable across a wide range of GCRs. Indeed, the only GCRs for which societal resilience is not helpful are those in which humanity goes extinct from the initial disaster. Only a small portion of GCRs would result in immediate extinction; these include physics experiment disasters, which could destroy the astronomical vicinity, and ASI, which might kill all humans in pursuit of its goals regardless of any human resistance. But for most GCRs, the risk can be reduced by increasing societal resilience. Actions to increase societal resilience thus have strong risk-risk synergy for GCR: the same action can reduce multiple GCRs.

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

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

The second way to increase societal resilience to GCRs is to increase local self-sufficiency to aid survivors in the event that global human civilization fails. Again, the measures that can be taken often apply widely across GCRs. For example, several GCRs pose direct threats to global agriculture, including nuclear war, asteroid collision, and volcano eruption, each of which block sunlight ("nuclear winter", "impact winter", and "volcanic winter"). Other GCRs threaten global food supplies in other ways, for example by disrupting supply chains. In the face of food supply catastrophes, local self-sufficiency can be enhanced via food stockpiles and alternative methods for growing food locally (Denkenberger and Pearce 2014; Baum et al. 2015).

Both ways to increase societal resilience to GCRs feature extensive synergies across risks: the same action will reduce the risk of many different GCRs. And resilience measures are not the

only ones to have this feature. Some other such measures (discussed above) are clean energy and agriculture, which reduce risk from several environmental GCRs. These synergies reduce some of the pressure on quantifying the risk: if an action reduces the risk for two different risks, the relative size of these two risks is less crucial. That said, the size of the risks remains important for comparing the value of different actions.

Risk-Risk Tradeoffs: Artificial Superintelligence Takeover

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

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

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

Systemic Integrated Assessment

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

Risk Reduction In Practice

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

The Direct Approach

The direct approach involves presenting the results of risk analysis directly to decision makers, who then take the analysis into account in their decision making so as to reduce the risk. The direct approach is perhaps the most familiar one for risk management, and the most idealistic in the sense that it describes an ideal risk management process.

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

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

Another reason that the direct approach may not work is that people do not always want to hear the findings of risk analysis. People may be motivated by cultural, political, or economic factors to ignore risk analysis or reject its findings. Indeed, there is a growing cultural tendency to dismiss all types of expert analysis as elitist, unnecessary, or otherwise unwanted (Nichols 2017). In the context of GCR, this phenomenon can be seen, for example, in the rejection of the scientific consensus on climate change, which is a major impediment to advancing climate policy, and in the rejection of expert advice to use vaccines, which could enhance the spread of pandemics.

The Indirect Approach: Mainstreaming

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

Mainstreaming has been successful for natural hazards, and it can also be successful for GCR (Baum 2015). For example, the 2014 Ukraine crisis brought increased interest in relations between the United States and Russia. This created opportunities to draw renewed attention to

nuclear war risk. The risk was a major focus of attention throughout the Cold War, but since then had largely faded from the spotlight. It was commonplace to believe that nuclear war risk ended with the end of the Cold War, but sure enough, the weapons still exist in large number, and United States-Russia tensions had not been fully resolved. The Ukraine crisis exposed this, creating an opportunity for discussion of a wide range of nuclear weapons issues, including those not directly related to the United States-Russia relationship. Additional opportunity is created by the alleged intervention by Russia in the 2016 United States election. There is a growing sense that the Cold War is back, which, for better or worse, means improved opportunities to draw attention to nuclear weapons issues.

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

The Very Indirect Approach: Co-Benefits

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

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

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

stance on climate change—it also depends on their stances on other GCRs, and perhaps on other factors as well.)

Stakeholder Engagement

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

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

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

Conclusion

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

Acknowledgments

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

References

Atkinson A, 1999. *Impact Earth: Asteroids, Comets and Meteors—The Growing Threat*. London: Virgin.

Barrett AM, 2017. Value of GCR information: Cost effectiveness-based approach for global catastrophic risk (GCR) reduction. *Decision Analysis*, 14(3), in press.
Barrett AM, Baum SD, 2017. A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2), 397-414.
Baum SD, 2014. The great downside dilemma for risky emerging technologies. *Physica Scripta*, 89(12), article 128004, doi:10.1088/0031-8949/89/12/128004.
Baum SD 2015. The far future argument for confronting catastrophic threats to humanity: Practical significance and alternatives. *Futures*, 72, 86-96.
Baum SD, Handoh IC, 2014. Integrating the planetary boundaries and global catastrophic risk paradigms. *Ecological Economics*, 107, 13-21.
Baum SD, Denkenberger DC, Pearce JM, Robock A, Winkler R, 2015. Resilience to global food supply catastrophes. *Environment, Systems, and Decisions*, 35(2), 301-313.
Baum SD, Barrett AM, 2017. The most extreme risks: Global catastrophes. In Bier V (ed.), *The Gower Handbook of Extreme Risk*. Farnham, UK: Gower, forthcoming.
Bostrom N, 2002. Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9(1).
Bostrom N, 2013. Existential risk prevention as global priority. *Global Policy*, 4(1), 15-31.
Bostrom N, Ćirković M, 2008. Introduction. In Bostrom N, Ćirković M (eds.), *Global Catastrophic Risks*. Oxford: Oxford University Press, 1-29.
Centeno MA, Nag M, Patterson TS, Shaver A, Windawi AJ, 2015. The emergence of global systemic risk. *Annual Review of Sociology*, 41, 65-85.
Cropper ML, 1976. Regulating activities with catastrophic environmental effects. *Journal of Environmental Economics and Management*, 3(1), 1-15.
Denkenberger D, Pearce J, 2014. *Feeding Everyone No Matter What: Managing Food Security After Global Catastrophe*. Waltham, MA: Academic Press.
Garrick BJ, 2008. *Quantifying and Controlling Catastrophic Risks.* Burlington, MA: Academic Press.
Gjerde J, Grepperud S, Kverndokk S, 1999. Optimal climate policy under the possibility of a catastrophe. *Resource and Energy Economics*, 21(3-4), 289-317.
Good IJ, 1965. Speculations concerning the first ultraintelligent machine. In Alt FL, Rubinoff M (eds.), *Advances in Computers*. New York, NY: Academic Press, 31-88.
Halal W, Marien M, 2011. Global megacrisis: A survey of four scenarios on a pessimism-optimism axis. *Journal of Futures Studies*, 16(2), 65-84.
Hertsgaard M, 2000. Mikhail Gorbachev explains what's rotten in Russia. *Salon.com*, 7 September, http://www.salon.com/2000/09/07/gorbachev.
Hook S, 1958a. A free man's choice. *The New Leader*, 26 May, 10-12.
Hook S, 1958b. Bertrand Russell retreats. *The New Leader*, 7-14 July, 25-28.
Konopinski EJ, Marvin C, Teller E, 1946. *Ignition of the Atmosphere with Nuclear Bombs*. Report LA-602, Los Alamos Laboratory.
Koopmans TC, 1974. Proof for a case where discounting advances the doomsday. *Review of Economic Studies* 41, 117-120.
Maher TM Jr, Baum SD, 2013. Adaptation to and recovery from global catastrophe. *Sustainability* 5(4), 1461-1479.
Martin IWR, Pindyck RS, 2015. Averting catastrophes: The strange economics of Scylla and Charybdis. *American Economic Review*, 105(10), 2947-2985.

Ng Y-K, 1991. Should we be very cautious or extremely cautious on measures that may involve our destruction? *Social Choice and Welfare*, 8(1), 79-88.
Nichols T, 2017. *The Death of Expertise: The Campaign Against Established Knowledge and Why it Matters*. Oxford: Oxford University Press.
Parfit D, 1984. *Reasons and Persons*. Oxford: Oxford University Press.
Persson I and Savulescu J, 2012. *Unfit for the Future: The Need for Moral Enhancement*. Oxford: Oxford University Press.
Reinhardt JC, Chen X, Liu W, Manchev P, Paté-Cornell ME, 2016. Asteroid risk assessment: A probabilistic approach. *Risk Analysis*, 36(2), 244-261.
Russell B, 1958a. World communism and nuclear war. *The New Leader*, 26 May, 9-10.
Russell B, 1958b. Freedom to survive. *The New Leader*, 7-14 July, 23-25.
Sagan C, 1983. Nuclear war and climatic catastrophe: Some policy implications. *Foreign Affairs*, 62, 257-292.
Schlesinger AM Jr, 1965 (reprint 2002). *A Thousand Days: John F. Kennedy in the White House*. Boston: Houghton Mifflin.
Seidel P, 2003. 'Survival research': A new discipline needed now. *World Futures*, 59(3-4), 129-133.
Tonn, BE, 1999. Transcending oblivion. *Futures*, 31, 351-359.
Tonn B, Stiefel D, 2013. Evaluating methods for estimating existential risks. *Risk Analysis*, 33(10), 1772-1787.
Wilson G, 2013. Minimizing global catastrophic and existential risks from emerging technologies through international law. *Virginia Environmental Law Journal*, 31, 307-364.