



# An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards

Simon Beard<sup>a,\*</sup>, Thomas Rowe<sup>b</sup>, James Fox<sup>c</sup>

<sup>a</sup> Centre for the Study of Existential Risk at the University of Cambridge, United Kingdom

<sup>b</sup> Department of Philosophy, Virginia Tech, United States

<sup>c</sup> Department of Engineering, University of Oxford, United Kingdom

## ARTICLE INFO

### Keywords:

Existential risk  
Methodologies  
Catastrophe modelling  
Subjective opinion  
Structured expert elicitation  
Doomsday argument  
Prediction markets

## ABSTRACT

This paper examines and evaluates the range of methods that have been used to make quantified claims about the likelihood of Existential Hazards. In doing so, it draws on a comprehensive literature review of such claims that we present in an appendix. The paper uses an informal evaluative framework to consider the relative merits of these methods regarding their rigour, ability to handle uncertainty, accessibility for researchers with limited resources and utility for communication and policy purposes. We conclude that while there is no uniquely best way to quantify Existential Risk, different methods have their own merits and challenges, suggesting that some may be more suited to particular purposes than others. More importantly, however, we find that, in many cases, claims based on poor implementations of each method are still frequently invoked by the Existential Risk community, despite the existence of better ones. We call for a more critical approach to methodology and the use of quantified claims by people aiming to contribute research to the management of Existential Risk, and argue that a greater awareness of the diverse methods available to these researchers should form an important part of this.

## 1. Introduction

“Apocalyptic predictions require, to be taken seriously, higher standards of evidence than do assertions on other matters where the stakes are not as great” (Sagan, 1983, 297-8)

How likely is it the next century will see the collapse of civilization or the extinction of humanity, and how much should we worry about different hazards that constitute this risk? These questions seem to bedevil the study of Existential Risk. On the one hand, they are important questions that deserve answers, not only to assess the level of risk facing humanity but as part of an integrated assessment of Existential Risk and opportunities to mitigate it (Barrett, 2017; Barrett & Baum, 2017). On the other hand, the quantification of Existential Risk is extremely challenging. As Carl Sagan pointed out, part of the problem with apocalyptic pronouncements is their theoretical basis and the fact that they “are not amenable to experimental verification – at least not more than once” (Sagan, 1983 298). Not only would a human extinction event be unprecedented, but the risk of such an event tends to emerge at the interaction of complex social, environmental and economic systems that are hard to model, and there is a substantial degree of uncertainty about the second order ‘risk space’ where they might be found. Together, these problems have made the quantification of risk speculative and reliant on new and creative methods for analysing the threats humanity faces (Currie, 2018).

\* Corresponding author.

E-mail address: [sjb316@cam.ac.uk](mailto:sjb316@cam.ac.uk) (S. Beard).

<https://doi.org/10.1016/j.futures.2019.102469>

Received 7 December 2018; Received in revised form 9 September 2019; Accepted 2 October 2019

Available online 03 December 2019

0016-3287/ © 2019 Elsevier Ltd. All rights reserved.

Previous analysis by [Tonn and Stiefel \(2013\)](#) attempted to resolve this tension, by evaluating a range of methods for quantifying Existential Risk from an ideal perspective, and made several important recommendations, most of which we endorse. However, although the number of researchers developing methods for the quantification of Existential Risk has grown, these methods have only been applied in a piecemeal fashion to a limited number of disciplines. For instance, conservation biologists have made important innovations in the use of structured expert elicitation; analysts in the Intelligence Community have developed new forecasting techniques with a high degree of success in assessing the probability of events in the near future and climate scientists, economists and epidemiologists have developed powerful models of complex global systems. These, together with the range of existing techniques from philosophical analysis and opinion polling to toy modelling and fault trees, each have benefits and limitations and we believe it is time they were more widely understood and adopted. The aim of this paper is, therefore, to survey the literature on the quantification of Existential Risk, to introduce different techniques to new audiences and to give an informal assessment of their capabilities, together with some suggestions for how they can be implemented and improved. We hope that this will spur still more methodological diversification and development.

This paper is addressed primarily to a group of scholars we refer to as the Existential Risk research community. This is made up both of researchers who work within institutes that focus on the study of existential and global catastrophic risk (such as the Future of Humanity Institute, the Centre for the Study of Existential Risk, The Global Catastrophic Risk Institute and the Future of Life Institute) and those who are consciously seeking to align their research with the goal of understanding and managing such risks. However, it is written from an awareness that these two groups do not necessarily contain all of those researchers whose work can be expected to contribute to our understanding and management of Existential Risk, and indeed many of the sources we consider were produced by researchers who are in neither of these groups. It is our conviction that one of the key roles of Existential Risk organizations like the above should not only be to support researchers who see themselves as falling into the Existential Risk research community but to forge better connections with, and a better understanding of, all research that is relevant to the understanding and management of these risks.

Finally, it is worth noting that this paper does not seek to consider fully every aspect of the quantification of Existential Risk. The risk arising from a specific threat is given by multiplying the probability of the threat occurring with its expected severity. This paper only sets out to examine the methodologies used to quantify the former. If one held the severity of all threats constant, at the point of human extinction, then this is all one needs to know. However, while some studies in this paper do aim to assess this, others assess threats at a lower point of severity, such as that of causing a global catastrophe or civilizational collapse, while others do not specifically consider the severity of a threat but are included rather because they relate to potential scenarios that have been of interest to scholars of Existential Risk. In these cases, one needs to be mindful that the severity of the event is still to be determined, or at best is only imprecisely defined when considering the overall quantification of that risk.

### *1.1. A brief introduction to different notions of probability*

We begin by noting that while this paper refers to the probability of Existential Risk, there are multiple ways of understanding the notion of probability. The first of these is the frequentist, or objective, notion of probability. According to this approach, probabilities are fundamentally related to the frequencies of events based on past observations. Once an experiment has been repeated many times, the frequency of any observed phenomena indicates its underlying regularity. Therefore, frequentist probability claims are sensitive to the experimental setup and measuring technique, and any new evidence requires probabilities to be reassessed from scratch. The second notion of probability is the Bayesian, or subjective, account, according to which probabilities represent our level of belief that a phenomenon will occur. One begins with a subjective prior belief about the probability of an event and then updates this via Bayes' Theorem (or Bayes' rule), which specifies how additional information affects the probability of an event. Even though these probabilities are subjective, one is required to set out reasons for arriving at them, which allows others to challenge these or update them further.

In this paper, we discuss the two notions of probability interchangeably with only a few passing remarks and we will instead focus on the quality of the methodologies that can be used to produce both kinds of probabilistic statement without prejudice. This is because even though Bayesian notions of probability dominate the field of Existential Risk, there are some areas, most notably in the domain of public health, where frequentist notions of probability are more common.

### *1.2. Conceptual challenges in studying Existential Risk*

The term Existential Risk can be understood in many ways and clarifying its definition is undoubtedly a crucial concern. However, since our aim is merely to study methods of quantifying Existential Risk, which approach this term from multiple angles, we will take a very broad view of how the term should be used, encompassing human extinction, civilizational collapse and any major catastrophe commonly associated with these things. See also [Torres and Beard \(2019\)](#).

Another point is that most studies consider Existential Risk in terms of distinct threats (such as nuclear war, pandemics and climate change). However, global catastrophes tend to involve a combination of multiple factors, including a precipitating catastrophic event, a systemic collapse that spreads this catastrophe to the global scale and a failure to take adequate steps to mitigate this risk ([Avin et al., 2018](#)).

Finally, Existential Risk cannot be studied in a vacuum. Even our assessment of such risks can profoundly affect them. For instance, if we take Existential Threats more seriously, this may lead to greater efforts to mitigate them. Sometimes, risk assessments can take account of human activities, such as when multiple estimates of catastrophic climate change reflect different future emission

paths; however, this is not always possible. Whilst important, we do not see this as a problem that we must solve here, since it is common to many fields of risk analysis and affects all the methods we describe to a greater or lesser extent.

Whilst some futurists may respond to these difficulties by adopting a pluralist conception of multiple futures, in which the goal is to map out the likely consequences of decisions that we face in the present, Existential Risk mitigation must go beyond this. In particular, it is necessary for research and mitigation efforts to be prioritized and for risk-risk trade-offs to be undertaken, such as when assessing dual-use technologies; these require the quantification of risk. We therefore believe that it is imperative to combine such pluralistic future scenarios into an integrated assessment that takes account of factors such as the resilience of global systems and the magnitude of Existential Risk. Given that such assessments are at an early stage, however, we will generally assume that all risk assessments are being made against a 'business as usual' scenario, where people continue mitigating risk roughly as much as they did at the time when that risk was assessed. In general, we suspect that this will overstate the future level of risk because it misses the potential for technological and governance interventions. However, that may not always be the case, as economic development can systematically push global systems into a more fragile state making Existential Risk increase over time.

### 1.3. Four criteria for evaluating methodologies

As well as presenting and discussing the existing methods for quantifying Existential Risk, we will provide an informal assessment of each according to the following four criteria:

**Rigour:** Can they make good use of the, generally limited, available evidence? Three key considerations for this are 1) their ability to access a broad range of information and expertise from across multiple perspectives, 2) the suitability of their means for turning this into a final judgement and 3) the ease of incorporating new information into this judgement or combining different judgements together using the same method.

**Uncertainty:** How well do they handle the, generally considerable, uncertainty in this field? Two key considerations for this are 1) whether they provide opportunities to quantify the level of confidence or uncertainty in their estimates, 2) whether the application of this method tends to systematically ignore or compound sources of uncertainty in the process of forming a final judgement and 3) whether they can help to identify and overcome epistemic bias.

**Accessibility:** Can they be applied by the individuals and, generally small and interdisciplinary, research groups that make up the Existential Risk research community? This is to be assessed in terms of 1) the amount of time required to implement them in a reasonable way, 2) The level of expertise required for a researcher to take a lead or principal role in implementing them and 3) what other barriers exist to their implementation.

**Utility:** Do they provide results that can be used for purposes like policy selection and prioritization and communicated to varied stakeholders? Three key considerations for this are 1) their credibility, both with scientists and non-scientists, 2) their ability to provide useful quantified information and 3) their ability to provide further information and insights about a risk and how to manage it.

For each of these, we assess methods on a four-point scale from Very Low to High and summarize the results of this evaluation in appendix B. We do not mean to imply that these four criteria are of equal importance; however, their relative importance is likely to depend upon the context for which an assessment of Existential Risk is being produced.

## 2. Analytical approaches

Not all methods for attempting to quantify Existential Risk are based on specific evidence for that risk. Given the lack of evidence available, this is perhaps more appropriate than it seems.

By far the most widely discussed of these is the so-called 'doomsday argument,' which was developed by multiple philosophers including Brandon Carter, Richard Gott and John Leslie. It is a statistical argument about the probability that any given human observer will be at a particular place in human history. There are multiple versions of this argument. Some appeal to our current position in the timeline of human history - for instance, it is much more likely that human history thus far represents more than 5 % of all human history than that we have more than 95 % of our history ahead of us. Others appeal to our position in the population of all human observers - it is much more likely that the human population born before me represents at least 5 % of all humans than that 95 % of humans will be born after me.

These arguments are controversial and they have been used to justify very different claims about the probability of Existential Risk, from a less than 2.5 % chance that we will fail to survive for at least another 5100 years (Gott, 1993 – source 3), to a 20 % probability that we will not survive the next century (Simpson, 2016 – source 5). Important factors that determine the difference between these claims include the type of doomsday argument being deployed, one's assessment of the length of human history or the number of humans who will ever exist and one's assumptions about the temporal distribution of Existential Risk (whether it is evenly distributed over time or comes in peaks and troughs). In theory, some of the uncertainty about these could be quantified by providing multiple calculations with differing assumptions; however, in practice, this is seldom done.

Two other analytical arguments are sometimes discussed in relation to their implications for the likelihood of human extinction. The first is known as the Simulation Argument, which connects the probability that humans face imminent extinction to the probability that we are living in a computer simulation (Bostrom, 2003). The second, known as the Great Filter Argument, connects the probability that humans face imminent extinction to the probability that there is intelligent life on other planets (Bostrom, 2008). Several sources cite a 2006 working paper titled 'The Fermi Paradox: Three Models', by Robert Pisani of the Department of Statistics at UC Berkeley, as providing a quantification of Existential Risk based on the Great Filter argument. However, the paper was never

published, and no version is currently available online. Apart from this, these arguments have yet to be used for the purpose of Existential Risk assessment.<sup>1</sup>

Part of the explanation for the interest in, and use of, the doomsday argument and other analytical tools is their high level of accessibility. Very Low across the other three categories. However, this does not mean these approaches should have no role to play in quantifying risk as they can still inform people's prior beliefs about human extinction, i.e. what we might most reasonably guess about the likelihood of human extinction before considering the evidence. A Bayesian account of probability requires such a prior to begin. In most scientific cases, the prior's precision is relatively unimportant as one can continually update it with evidence. However, within the context of Existential Risk, where evidence is sparse, prior probabilities are disproportionately important, meaning that any technique helping produce better priors can still be very useful. This is the approach taken by [Leslie \(2002 – source 6\)](#) and [Wells \(2009 – source 4\)](#).

### 3. Modelling based approaches

Since it is not possible to undertake an empirical study of human extinction or global civilizational collapse, the next set of methods use observable evidence to produce a set of assumptions, or a model, that allows us to study them indirectly.

#### 3.1. Toy models and extrapolation from data

The simplest approach of this kind involves assuming there exists an underlying regularity in the frequency of certain events that have historically already occurred and have the potential to pose an Existential Risk in the future. A frequentist analysis of historical data can then be used to estimate the approximate time interval between such events and hence produce an annual occurrence probability. So far, this approach has been applied to asteroid impacts ([Bostrom, 2006 - source 46](#)), super-volcanic eruptions ([Decker, 1990 – source 50](#), [Harris, 2008a,b – source 51](#) and [Aspinall et al., 2011 – source 52](#)), nuclear wars ([Lundgren, 2013 – source 14](#)), space weather ([Love, 2012 – source 59](#), [Homeier et al., 2013 – source 60](#), [Melott et al., 2004 – source 61](#) and [Gehrels et al., 2003 – source 62](#)), particle physics experiments ([Dar, De Rújula, & Heinz, 1999 – source 63](#), [Jaffe, Busza, Wilczek, & Sandweiss, 2000 – source 64](#), [Tegmark & Bostrom, 2005 – source 65](#) and [Ellis, Giudice, Mangano, Tkachev, & Wiedemann, 2008 - source 66](#)) and the occurrence of extinctions and global catastrophes in general ([Hempsell, 2004 -source 1](#), [Snyder-Beattie, Ord & Bonsall 2019 – source 2](#)). [Yampolskiy \(2018\)](#) also applies this approach to predicting AI failures but without producing a quantified estimate of their probability.

Where no such underlying frequency can be assumed, for instance because an event is historically unprecedented, it is possible to produce simple toy models that can allow one to use historical data to determine the probability of an event occurring in other ways. Firstly, one can assume that the magnitude of impact from certain threats follows a specific distribution, enabling one to estimate the probability of a large impact event taking place from the historical record of smaller impact events ([Hanson, 2008](#)). For instance, [Millet and Snyder-Beattie \(2017 - source 28\)](#) for the number of fatalities from biowarfare or bioterrorism and [Riley \(2012 – source 58\)](#) for solar flares both assume power law distributions while [Bagus \(2008 – source 23\)](#) assumes that the fatalities from influenza pandemics follow an exponential distribution. From this, they can estimate the probability of more extreme events of this kind that have the potential to pose and Existential Risk. Other toy models include [Day, André, & Park \(2006 – source 20\)](#) [Fan, Jamison, & Summers \(2018 – source 22\)](#) and [Millet and Snyder-Beattie \(2017 - source 27\)](#), all of which assess the risks of catastrophic pandemics.

Secondly, one could assume that a currently unprecedented event would occur as a consequence of multiple events with historical precedents. The existential event could be the end result of a chain of precipitating events each a possible consequence of another. In this case the unprecedented event's probability is the product of the conditional probabilities down the chain. [Klotz and Sylvester \(2014 – source 24\)](#), [Lipsitch and Inglesby \(2014 – source 25\)](#) and [Fouchier \(2015 – source 26\)](#) use this method to estimate the likelihood and impact of a global pandemic resulting from Gain of Function influenza research by assuming that such a risk results from the occurrence of two events whose probabilities are easier to determine, a laboratory-acquired infection and a biosecurity failure. Similarly, [Hellman \(2008 – source 12\)](#) estimates the probability of a nuclear war resulting from a 'Cuban Missile Type Crisis' as the product of probabilities of a sequence of four precipitating events. Alternatively, the unprecedented event could arise due to the coincidence of two mutually independent events. This would create the basis for a fault tree, discussed in the next section.

Although these approaches are methodologically clear, clarity does not imply objectivity. For instance, the analysis of the geological record is itself speculative and open to interpretation ([Currie, 2018](#)) so that [Decker's \(1990\)](#) estimate of the probability of supervolcanic eruptions, while widely cited in the Existential Risk community, is often seen as pessimistic amongst volcanologists and is higher than most other estimates. Other instances of disagreement are even greater, [Lipsitch and Inglesby \(2014\)](#) and [Fouchier \(2015\)](#) arrive at probability estimates over seven orders of magnitude apart, despite using the same toy model to evaluate the same risk. This points to the central issue, approaching the same historical data in different ways can result in very different probability estimates. [Love \(2012 - source 59\)](#) comments on the "limited accuracy of statistical estimates" when comparing his result to Riley's for the recurrence probability of Carrington-like geomagnetic storms, saying that we can only conclude that the probability is

<sup>1</sup> We are aware of one other paper that seeks to quantify Existential Risk using the Fermi Paradox, but the probability estimate is given in terms of an arbitrary free variable for the number of civilisations that have reached our level of development in our neighbourhood. Since it does not also attempt to quantify this variable a final estimate cannot yet be produced, so we have not included this in our literature review. See [Miller, J. D., & Felton, D. \(2017\). The Fermi paradox, Bayes' rule, and existential risk management. Futures 86: 44–57.](#)

“somewhere between vanishingly unlikely and surprisingly likely”. Meanwhile, Hellman openly cherry picks evidence so as to not appear ‘alarmist’.

Another problem with these methods is that some events are excluded from the historical record because of the ‘anthropic shadow’ they would leave. Roughly speaking, were any such event to have occurred in the past, this would have led to the non-existence of the present observer. Therefore, even if its probability was extremely high, it must seem to us as if it could not have happened because our very existence depends on it (Ćirković, Sandberg, & Bostrom, 2010). Manheim (2018) looked at risk estimates of natural pandemics and concluded that there is significant uncertainty about the relationship between historical patterns and present risk because of such “anthropic factors and other observational selection biases”. Tegmark and Bostrom (2005 – source 65) also take account of this effect when quantifying the threat from particle physics experiments, but most theories ignore it. A related issue is that the historical record may need to be understood not just by what happened but also what didn't happen. Gordon Woo (2018) argues we should sometimes incorporate “counterfactual analysis” of near-miss events to more accurately model risks because of the likelihood that very rare events will be underrepresented in historical data. For instance, if the true underlying probability of a certain sized asteroid striking earth was 0.004 per year and we have 100 years of data, then it is most probable that this event would not have occurred within this period leading our analysis to underestimate its probability. However, if we had evidence of asteroids of this size passing close to the earth without striking it during this period then we can use this information to arrive at a better estimate for the underlying probability of such a strike. Woo cites numerous “near-misses” in fields from maritime and air disasters to terrorism that can be used to make better predictions about rare kinds of catastrophe (Woo, 2018), while other scholars have incorporated historical near misses into the study of nuclear war (Lundgren, 2013 – source 14, Lewis, Williams, Pelopidas, & Aghlani, 2014, Barret, de Neufville & Baum, 2018).

In principle, using historical data allows us to calculate our degree of uncertainty via simple statistical techniques. Thus, many of the sources that utilize this approach provide ranges for the probability of the events that they study. However, care needs to be taken when combining uncertain statistics in toy models since merely multiplying uncertainty ranges will overstate uncertainty as the coincidence of multiple outliers can be expected to be rarer than any individual outlier. Furthermore, because the modelling approaches we discuss make assumptions about the distribution of the underlying phenomena from which the data has been sampled, they are vulnerable to abnormal (or even discontinuous) changes in these phenomena. For instance, Fouchier points out that when using historical data about biosecurity, one must account for generally increasing safety standards. Similarly, Manheim (2018) notes that the likelihood of global catastrophic biological risks might be higher than historical evidence implies because of contemporary global travel, high population densities in megacities, and closer contact with animal populations due to factory-farms. Therefore, simplistic frequentist approaches may underestimate the appropriate level of uncertainty.

Despite its simplicity and popularity, we believe that toy modelling of Existential Risk has serious shortcomings. We rate these methods Low for rigour, uncertainty and utility, while once again noting their often underutilized potential to quantify and estimate uncertainty. Indeed, the utility of this approach appears to be highest in cases where assessors are overconfident, both regarding the uncertainty surrounding their predictions and the objectivity of the historical record, which is potentially dangerous. We do however rate this method highly in terms of its accessibility and the opportunities it provides for researchers to bring existing evidence to the study of Existential Risk. We see this approach as having limited appeal for quantifying Existential Risk as the field matures, though it may still play a useful role in stimulating further research and in producing estimates that can be taken as ‘objective’ priors for further Bayesian analysis.

### 3.2. Fault trees and Bayesian networks

A more sophisticated modelling technique for studying Existential Risk involves fault trees. Originally developed to model the emergence of system failures in safety engineering, they have now been widely applied in risk analysis. Fault tree models use Boolean algebra to map out in a logic tree, using “and” and “or” gates, how a system failure could arise. Branching backwards from overall system failure at the tree's top, we first write the ways failure could happen as different nodes and then branch further backwards with how this node could fail and so on. If possible, a probability of failure is assigned to each node and then one can sum or multiply probabilities, depending on the Boolean nature of each gate, to give the overall probability of system failure. Importantly, fault trees can also clearly reveal preventative steps that could be taken. This technique has been used to study risks from nuclear war (Barrett, Baum, & Hostetler, 2013 – source 13)<sup>2</sup> and AI safety (Baum, Barrett, & Yampolskiy, 2017 – source 40).

Bayesian Networks are an extension of fault trees and present great promise for Existential Risk analysis (Tonn & Stiefel, 2013). Although the power of Bayesian networks is also rooted in its graphical representation of the possible failure under study, the nodes now represent random variables and it is the edges between these nodes that are quantified. These edges are directed from parent to child nodes and every node comes with a conditional probability table that can represent the causal probabilistic strengths for the edges between connected nodes. As with fault trees, a Bayesian network is first drawn by working backwards from a failure state (or any other outcome one wishes to study) through the conditions that might lead to this state. However, unlike Fault Trees, Bayesian Networks can handle dependencies between different parts of the system and so all conditions can be factored into one's analysis,

<sup>2</sup> Whereas Barrett et al. (2013) model one set of nuclear war scenarios between Russia and the U.S., Baum, de Neufville, and Barrett, (2018) extend this fault tree to model all important nuclear war scenarios between all relevant states, but they stop short of quantifying the overall probability so we have not included this in our literature review. See Baum et al. (2018). A Model for the Probability of Nuclear War. SSRN Electronic Journal. <http://www.doi.org/10.2139/ssrn.3137081>.

including those that are only rarely important. This can make it easier to incorporate information from near misses and other tenuous sources whilst still being rooted in observed system behaviour. Once a network has been created, Bayes' rule can determine the final expected occurrence probability of each node in this network, including that of the final outcome, and this can be calculated dynamically and updated continuously to take account of additional information.

Khakzad, Khan, and Amyotte (2011) showed why fault trees are less suitable for modelling complex systems because Bayesian Networks are far superior at handling dependencies between different parts of the system, common cause failures, and uncertainty. Also, Bobbio, Portinale, Minichino, and Ciancamerla (2001) demonstrated that every fault tree has a corresponding Bayesian network and so the methodology is proven to be generalisable. However, when evidence is sparse, the results from Bayesian Networks can be significantly affected by the modeller's choice of prior probabilities. Bayesian network analysis has not yet been used to study Existential Risk directly; however, it has been applied to a variety of catastrophe models, for instance Li, Wang, Leung, and Jiang (2010) demonstrated how Bayesian Networks can assess catastrophic risks under uncertainty by modelling catastrophic flooding in China. Both fault trees and Bayesian Networks improve on simpler toy models as they can manage more complex system dynamics whilst handling a greater range of data and inputs. We rate both approaches as Medium for utility but note their particular value in providing insights that can be used to study risk mitigation by modelling how changes in components of a system affect its probability of failure. To be useful, the model must be a sufficiently faithful and detailed representation to capture accurately the effect of individual policies or interventions. It is possible that if the model fails on either of these, then it may simply lead decision makers astray and give them false confidence. This is reflected in our higher degree of uncertainty about this particular classification.

Bayesian Networks outperform Fault Trees in terms of both rigour (High/Medium compared to Medium) and uncertainty (High compared to Low), performing especially well in relation to the quantification of uncertainty. However, Fault Trees outperform Bayesian Networks in terms of accessibility (Medium compared to Low), a category in which neither method performs well as both require significant modelling skills and domain knowledge. Fault tree analysis requires considerable familiarity with the underlying system to understand the processes that may lead to it failing. Bayesian networks require an even greater degree of familiarity as the assessor must provide probabilities for all the conditional relationships that may play a role in determining an outcome.

### 3.3. Adapting and applying existing models

Some researchers, usually from outside the Existential Risk community, have also adapted existing models to study globally catastrophic and existential threats. These include using models of pandemic influenza to assess the likelihood and impact of a 'modern Spanish flu' (Madhav, 2013 – source 21), adapting IPCC climate models to assess the probability of catastrophic climate change (Xu & Ramanathan, 2017 – source 35) and using astronomical models of near-Earth objects to assess the likelihood of asteroid impacts (CNEOS – source 43, Harris, 2008a,b – source 47 and National Research Council 2010 – source 48). Other authors infer from existing estimates of the uncertainty in models what the possibility of more severe impacts might be, including: Wagner and Weitzmann (2015 – source 32) for catastrophic climate change and Atkinson et al. (2000 – source 44) for near asteroid impacts. Finally, some identify parameters where they believe the model is mistaken and use this as a justification to 'correct' the final output, as Dunlop and Spratt (2017 – source 34) and King, Schrag, Dadi, Ye, and Ghosh (2015 – source 33) do for the IPCC's predictions about climate change.

In theory, these approaches can build on the underlying rigour and utility of the model being used; however, this depends significantly on the validity of their adaptation or application. Most models, including the IPCC's, are not designed to assess catastrophic risks specifically, so catastrophic outcomes will be outliers. This provokes debate about whether these outcomes should be treated with genuine concern or can be dismissed as model failures (Pindyck, 2013). Modifications must only be made by those with significant skills and a deep understanding of the model's functionality to adapt them in ways that will preserve their virtues. This may explain why the most comprehensive efforts to do this come from scholars like Ramanathan and Madhav who are outside of the Existential Risk community, which, in turn, may mean that these researchers understand less about the nature of such risk.

Despite their very low levels of accessibility, well-executed applications of sophisticated modelling techniques represent a desirable next step in the study of Existential Risk and we rate them highly in terms of rigour, uncertainty and utility. However, we are sceptical of attempts to replicate the success of high profile existing models with fewer resources, by making less considered model adjustments or making concrete predictions based solely upon their current levels of uncertainty and suggest that this approach should be adopted cautiously and only by better resourced groups within the Existential Risk research community.

## 4. Subjective approaches

Of the 66 sources in our literature review, 45 % relied, at least in part, on the subjective opinion of the author or others, without direct reference to specific models, data or analytical arguments. This included all the sources that discussed the potential threat from Artificial Intelligence, which many Existential Risk scholars believe to be the most significant. This is unsurprising given the difficulties that other methods face, and the use of subjective expert opinion is a well-established and successful means for handling uncertainty in many other scientific fields (Aspinall, 2010; Morgan, 2014).<sup>3</sup> However, not all subjective opinion should be treated equally and, in the next two sections, we will consider different approaches.

<sup>3</sup> Although, note that Morgan specifically cautions against the use of subjective opinion for highly ambiguous phenomena for which there may be no reliable expertise available.

#### 4.1. Individual judgements

At one end of the spectrum are individual opinions given without reference to clear reasoning. Examples include Rees (2003) – source 8) and Stern (2006 – source 11), who both consider the overall probability of human extinction, one for a ‘scientists warning’ and the other to determine the correct social discount rate. Bostrom (2002 – source 7) provides a similar judgement, although this appears to be based on updating a prior belief, derived from analytical arguments, to account for ‘the balance of evidence’. Halstead (2018 – source 36) and Chapman (2004 – source 45) both present a considerable degree of evidence and argumentation before offering subjective conclusions about the threat of climate change and asteroid impacts, but without any specific method to connect the two. At best these estimates represent what Tonn and Stiefel (2013) refer to as ‘holistic probability assessments’, in which “the individual probability assessor estimates the holistic extinction risk through informed reflection and contemplation”.

A more sophisticated approach to subjective opinion formation is for probability assessors to break down any risk into a set of mutually exclusive threats and then classify the danger posed by each of these and the probability of their occurrence (together with the likelihood that they would pose an existential risk). Tonn and Stiefel refer to this as the ‘whole evidence Bayesian approach’. This encourages a systematic way of estimating probabilities and is useful to anyone reviewing such an assessment as it makes it easy to update predictions in the light of new evidence or different reasoning. Tonn and Stiefel illustrate what such reasoning might look like; however, we have decided not to reproduce this here as it seems clear that they intended it only as an illustration.

Another approach is what Tonn and Stiefel refer to as ‘Evidential Reasoning’. This involves specifying the effect every piece of evidence has on one’s beliefs about the survival of humanity. Importantly, these probabilities should only reflect the change that this evidence makes, not one’s initial prior beliefs, allowing others to assess them independently. As such, they will only be ‘imprecise probabilities’ that describe a small portion of the overall probability space, where the contribution each piece of evidence makes to one’s belief and its complement need not sum to 1. For instance, one might reason that evidence about the adaptability of humans to environmental changes suggests a 30 % probability that we will survive the next 1000 years, but only a 10 % probability that we will not.<sup>4</sup> Combination functions can then be used to aggregate these imprecise probabilities to return the overall probability of extinction within this period. This method not only helps assessors determine the probability of extinction, but also provides others with information about the sources of evidence that contributed to this decision and the opportunity to determine how additional information might affect this.

A final method listed by Tonn and Stiefel draws on the technique, common amongst futurists, of anchoring assessments in scenario based considerations of what it would take to bring humanity to extinction. Assessors envisage a possible human extinction scenario and then consider how indicative this scenario is of both the space of all possible future scenarios and the space of those in which humanity goes extinct. This exercise is repeated until the assessor judges that they have exhausted all, or at least a substantial portion, of the human extinction scenario space. They can then estimate what proportion of future scenarios involve human extinction, and by extension how likely this is. Advantageously, this focuses on the end result, human extinction, rather than on the processes by which this might be brought about, although the use of scenarios is sometimes frowned upon in other communities. Despite the fact that these three techniques consist of little more than clearly setting out one’s assumptions and reasoning process for others to follow, none of them has so far been implemented well in the literature on Existential Risk.

It has been shown that, with only a few hours of basic level of training using freely available tools, most people can be calibrated to give reliable estimates of their level of uncertainty for their subjective opinions based on their current state of knowledge (Hubbard, 2014). Despite this, few who have conducted subjective probability assessments have indicated that they have undertaken any such calibration or to state their degree of uncertainty. Instead, experts have tended to hedge their bets merely by couching otherwise precise estimates in vague language. Furthermore, individuals routinely suffer from overconfidence and confirmation bias in their subjective estimates, and when individuals have their name attached to a figure, such biases become especially problematic. Eliezer Yudkowsky (2008) discusses the relevance of cognitive biases affecting the judgement of Global Catastrophic Risk including the availability heuristic, hindsight bias, conjunction fallacy, scope neglect and overconfidence.

The popularity of individual subjective opinion is probably because they are especially easy for researchers to apply and are often offered as the basis for further discussion and inquiry in the future. Such estimates can also be well received by media and policymakers alike, especially when they can be linked to a high-profile academic of celebrity and hence become associated with that individual’s perceived authority, potentially enhancing their utility. Indeed, it often appears easier to get people to agree with the single judgement of a known individual than a collective judgement which combines information from that individual with others.

The quality of individual subjective opinion thus depends on both the person providing the estimation and where suitable techniques are used to present and clarify their reasoning and assumptions. We rate this approach as Low/Medium for rigour. Despite the fact that this kind of estimate is well received we rate its utility as Low, reflecting its generally narrow focus and lack of credibility within scientific communities. We also rate it as Very Low for uncertainty, in particular due to its weaknesses in overcoming bias. However, we rate this method as high in terms of accessibility, which probably helps to explain its relative popularity in this field. The kind of robust approaches to clarify one’s thinking would hardly detract from this high level of accessibility, and indeed may make it even easier for assessors to reach a final judgement, so it is disappointing to see them so little used.

<sup>4</sup> These figures are based on an illustration produced by the authors.

#### 4.2. Aggregating expert opinion

Another way of seeking to improve on individual subjective opinions is to pool together the judgements of multiple people to account for a more diverse range of perspectives. There are two reasons why this could improve the quality of judgements.

The first of these relates to the ‘wisdom of crowds’ (Condorcet, 1785; Galton, 1907), which provides an epistemic justification for the aggregation of large numbers of individual opinions to determine the truth of some proposition. It relies upon the assumption that individuals receive some kind of signal pointing to the truth or falsity of that proposition, and that as a result, they are slightly more likely to judge correctly than incorrectly. The theory then states that, so long as individuals are making independent judgements, adding more will increase the probability that the group’s median judgement will tend towards the correct one. The distribution of judgements across the group will effectively cancel out the noise that leads to some incorrect individual judgements and amplify the correct signal.

The second is that whilst individual judgements will be affected by multiple biases, when aggregated over many people, these biases may average out, improving collective judgement’s accuracy. However, this can also be counterproductive as biases are often shared across large groups, or even reinforced by groupthink and the sense that one may be judged by biased peers. This violates the independence of individual judgements and can lead to the predictive power of a group decreasing with its size (Fujisaki et al., 2018). Partly as a response to this, some studies have suggested that aggregation methods that give more weight to outlying opinions outperform straightforward averaging approaches (Tetlock, Mellers, & Scoblic, 2017).

In this section, we will limit ourselves to discussing approaches which simply average expert judgements, whilst in the next section, we will turn to more structured and deliberative approaches.

Simple aggregation is the dominant method for making predictions about the Existential Risk from Artificial Intelligence (Müller & Bostrom., 2016 – source 38 and Grace, Salvatier, Dafoe, Zhang, & Evans, 2018 – source 39), but has also been applied to the prediction of nuclear wars (Project for the Study of the 21st Century, 2015 – source 15) and to quantifying Existential Risk in general (Sandberg & Bostrom, 2008 – source 9). These surveys vary considerably in quality and size, with many showing little concern for the diversity of participants, the statistical rigour of their analysis or uncertainty quantification (the honourable exception being Grace et al., 2018). Most surveys take the median response as their prediction, but Turchin (2019) argues that this is sometimes not optimal for Existential Risk. For AI safety, instead of using the median estimate of AGI creation for risk assessments, we should be concentrating on the earliest possible time of AGI creation and define a ‘minimum acceptable level of AI risk’.

Those who have adopted this approach often acknowledge its limitations. For example, Sandberg & Bostrom (2008) state that “these results should be taken with a grain of salt. [...] There are likely to be many cognitive biases that affect the result, such as unpacking bias and the availability heuristic as well as old-fashioned optimism and pessimism.” Judgement independence is hard to ensure as surveys are often completed at conferences and so it is difficult to guarantee that individual judgements are not influenced by others. Remote participation via an anonymous platform may offer a partial solution to this problem. However, given how close-knit many academic and technical communities are, this still may not secure judgement independence. Finally, although aggregation may improve judgements, it has the effect of making them less well behaved. For instance, if one seeks the median estimate from a group about the probability of superintelligence being developed and the probability of superintelligence leading to human extinction and then combines these figures, this can differ substantially from the median group prediction that humanity will go extinct from superintelligence.

The aggregation of expert opinion has the potential to improve upon individual judgements regarding their rigour and ability to handle uncertainty; however, in practice, this opportunity is, once again, often not taken. This may reflect the fact that extensive, well designed surveys are still out of reach of many small research groups and that people seem to respond equally well, if not better, to overconfident survey results from a small pool of ‘experts’ than to extensive well-designed surveys that express an appropriate degree of uncertainty. We rate this approach as Medium for rigour, Low for uncertainty and utility, and Medium, both with a reduced level of confidence, for accessibility (due to disagreements about the amount of time it takes to conduct surveys in a ‘reasonable’ way)

### 5. Structured and deliberative approaches

The final family of approaches we discuss also use subjective opinion, but seek to combine multiple opinions in more structured ways than simple aggregation. A variety of such methods have been developed by scientists and foresight specialists to aid decision making under uncertainty, although so far these have been sparsely used in quantifying Existential Risk. Some of the techniques we describe form part of proprietary foresight tools such as the Delphi Technique (developed by RAND) and Superforecasting (developed by the Good Judgement Project); however, these can be disaggregated into their constituent parts for the purposes of discussion.

#### 5.1. Weighted aggregation

The first of these approaches weights opinions differently in the aggregation based on an assessment of each individual opinion’s value. For instance, Roger Cooke’s (1991) “classical” approach to expert elicitation gives greater weight to subjective opinion based upon experts’ performance on a series of calibration questions that ask them to predict things that are either known or that can easily be determined. An expert who more often gets closer to the truth has a larger weight in the overall aggregation of judgments. This approach’s prediction accuracy has been shown in multiple studies to outperform simple aggregation (Colson & Cooke, 2018 - but see also Clemen, 2008).

However, this method is not well suited to predicting Existential Risk as the experts’ competency at predicting catastrophic and



Existential Risk cannot be calibrated due to their unprecedented nature. It might be possible to test experts' putative accuracy through their success at predicting more common and nearer term future events; someone's success at predicting short-term AI milestones could reflect the strength of their predictions about the long-term future of AI. However, there is no obvious means for assessing how success at predicting short-term and long-term trends are related.

An alternative means of weighing individual judgements is via peer ratings of respect and reliability. Theoretically, this avoids the problem of needing to calibrate individual predictors based on past performance and could help individuals assess their own beliefs by considering their credence in the beliefs of their peers. However, such weights are often of little, if any, meaning, especially in the context of Existential Risk (Burgman et al., 2011). Weightings can also be generated by repeatedly sampling experts' predictions and weighting those who gave more consistent answers more strongly than those whose answers varied, potentially indicating a lack of evidence based judgement. For instance, Bamber and Aspinall (2013) asked for the same estimate from experts two years apart and those resampled were forbidden from referring to their first estimate to test the stability of individual judgements in determining the risk of a future sea level rise due to climate change. However, this approach is problematic because it would be hard to distinguish between estimates that varied over time because of the randomness arising from a lack of evidence and those which changed because the estimators were successfully updating their predictions to take account of additional information.

The final method of weighting that we consider was developed by the Good Judgement Project on the basis that empirical studies indicate that some individuals (who the project terms 'superforecasters') are substantially better at making predictions about the future than others. The project selected over 2000 individuals and tasked them with assessing the likelihood of various world events. It found a considerable degree of variance amongst participants, with some individuals performing consistently well regardless of the kind of prediction they were being asked to make. Furthermore, it found that individuals who performed consistently highly in making accurate predictions were able to outperform even domain experts and professional intelligence operators. Philip Tetlock, the project's leader, concluded that these individuals had particular psychological traits that led them to make more accurate predictions, including caution about the strength of their beliefs, humility about the extent to which complex processes can be simplified, curiosity about the facts of a case, valuing diverse views and opinions and a belief in the possibility of self-improvement (Tetlock et al., 2017). However, rather than assessing these psychological traits directly, the key to identifying superforecasters has been to keep track of individual performances at making predictions, including people's ability to update these in order to account for new information. This was done by assigning a 'Brier score' to each superforecaster, an assessment of how close their predictions came to actual events (Tetlock & Gardner, 2016). The project found that the most accurate predictions were produced from an aggregation of participants' predictions, but those with the highest Brier scores were weighted more strongly. However, it is worth noting that the success of these superforecasters was found to diminish significantly when they were asked to make predictions more than 12 months ahead. At present it is unclear whether this reflects a limitation of superforecasters' abilities or a general problem with making longer term predictions.

We rate these methods as no better than Aggregated Opinion Surveys for any category, but low for accessibility. Given this, it is hardly surprising that such techniques have not been used in the assessment of Existential Risk thus far.

## 5.2. Enhanced solicitation

Another approach to structured expert elicitation is to seek to improve, rather than simply measure, the quality of experts predictions. Broadly speaking this can be performed prior to solicitations being made, at the point of solicitation or between solicitation and a final judgment being produced.

Pre-solicitation methods of improving the quality of expert judgement focus on training and method selection. We have mentioned a variety of such approaches already in Section 4.1, including calibration of uncertainty and the use of formal methods like evidential reasoning and holistic probability assessments. However, some methods specifically focus on prediction as a structured group activity and these are worth noting here. For instance, the Good Judgement Project found that both natural Superforecasters and those who did not share their psychological traits were able to learn and develop them over time to greatly improve the accuracy of their predictions when they went through a process of probability training, teaming and tracking. Probability training helped correct cognitive biases, teaming allowed for the sharing of information and the public justification of why a probability was given, and tracking encouraged participants to outperform their previous track record and helped develop stronger teams of peers who could learn from one another (Mellers et al., 2014).

However, it is worth noting that a large amount of time and resources go into selecting super-forecasting teams. The Good Judgment Project spent four years assembling their elite super-forecasting team. It is difficult to imagine that such teams could be rolled out more extensively. Nevertheless, the approach itself is quite simple, and several people in the Existential Risk community have attempted to adapt elements of it into their work. For example, just one hour of training in probabilistic reasoning noticeably improved forecasting accuracy (Chang, Chen, Mellers, & Tetlock, 2016).

Whilst no superforecasters have attempted to predict the possibility of Human Extinction; the organization Open Philanthropy commissioned a team of super-forecasters to predict the probability of a nuclear war (Source 16). Lessons from this approach could be incorporated by the Existential Risk community in one of two ways. Firstly, it might be possible to train those who make Existential Risk predictions with the super-forecasters' techniques. Secondly, applying more resources could motivate existing super-forecasting teams to make more relevant predictions of Existential Risk. Both will take considerable work, and it remains unclear how successful they will be.

Efforts to improve the quality of probability estimates at the point of solicitation focus on what questions are asked and how the person soliciting expert opinions engages with them. In a recent, albeit unpublished, solicitation of expert judgements of the

probability of a global catastrophic biological risk, David [Manheim \(2018\)](#) - source 29) used a variety of such approaches to solicit better quality information from experts in infectious disease. He found that these experts were both poorly versed in probabilistic thinking and liable to reject the notion of a global catastrophic biological risk (in this case “a natural infectious disease that kills 1 billion people”); they disputed whether this could ever happen. However, Manheim proceeded to explain to these experts that other natural events could have catastrophic events and challenged them to provide a fundamental reason why such an event was impossible within the infectious disease domain. By then engaging the experts in scenario-based thinking about what properties such a disease would need to have, Manheim was able to solicit useful information with a reasonable degree of consensus between the experts. Only one expert continued to claim such an outcome was utterly impossible, but they now justified this claim, stating that this was a result of their belief that public health responses would always be sufficient to prevent such a pandemic. The experts remained unwilling to be quoted because they perceived a significant reputation risks in even discussing these extreme events. Post solicitation methods focus on deliberation between experts, creating opportunities for experts to offer updated predictions or sometimes requiring them to adjust their judgements to move towards a consensus opinion. The most famous result is the so called ‘Delphi technique’ developed by RAND in the 1950s. This can be applied to a variety of foresight and horizon scanning activities and uses a panel of experts who are asked to respond to a series of questions across two or more rounds. After each round, a facilitator provides an anonymised summary of the results, along with the reasons each expert provided for their answers. Extreme outliers must substantiate their position. Experts can then revise their judgements given the broader knowledge achieved through considering the responses of others leading, hopefully, to experts converging on the “correct” judgement. Delphi studies have been conducted to provide quantitative assessments in many areas of risk analysis, but the technique has not yet been used to provide quantitative estimates of Existential Risk, although [Wintle et al. \(2017\)](#) apply it to identify key emerging risks related to biotechnology in the Existential Risk context. Other forms of structured expert elicitation that are related to, though not identical with, the Delphi technique have been harnessed to assess existential and Global Catastrophic Risk. For instance, [Pamlin and Armstrong \(2015 – source 18\)](#) used a complex multi-layered process of literature review, deliberative workshops and individual subjective judgements to select and assess Global Catastrophic Risks, but without multiple rounds of estimation. Another approach that has recently been developed, in part by Existential Risk researchers, is the IDEA (Investigate, Discuss, Estimate and Aggregate) protocol ([Hanea et al., 2017](#)). This drops the focus on seeking consensus and allows participants to discuss differences of opinion and defend probability estimates directly rather than responding to anonymised statements of reasons. The final independent estimates are given as anonymous submissions and then aggregated.

Whilst the Delphi technique and its relations aim to remove personal bias from predictions, as with all survey methods there may still be bias in the selection of the experts that can potentially lead to self-fulfilling prophecies ([Devaney & Henchion, 2018](#)). Individual biases may influence people’s willingness to update their judgment in light of evidence from the group and thus disproportionately sway the overall groups’ findings. Moreover, some participants may wish to tailor their contributions to ensure that there is a concordant result, rather than rocking the boat with a contribution that throws the group further away in their estimate. Furthermore, the focus on consensus may be at the expense of cultural and other embedded differences in individuals’ perspectives on information ([Ahlqvist & Rhisiart, 2015](#)). Finally, when the aggregation of expert opinion involves additional deliberation between experts, this can lead them to shift away from consensus and towards the most extreme views under discussion; individuals begin to cluster their identities around opposing positions, such as those defined along political or disciplinary lines ([Sunstein, 2000](#)).

These techniques are relatively difficult to implement, requiring technical familiarity and the resources to convene a sufficient number of experts to implement them, but these barriers are lowering with time especially as the Delphi technique has a long track record of use in a variety of scientific and policy contexts. The fact that this approach can harness knowledge and expertise from across disciplinary backgrounds, requires individuals to substantiate judgements and encourages individuals to revise their first estimates in light of new information lends it a considerable degree of rigour, at least relative to many other methods that we have looked at. Whilst potentially controversial, the results are easy to communicate and are given credibility by the structured process through which they are obtained.

[Bamber and Aspinall \(2013\)](#) noted that experts in their study were “exceedingly uncertain about the answer to [the] key question”. They argue that whilst structured expert elicitation can help to quantify uncertainties; it does not overcome them. Such high degrees of uncertainty are often seen as prohibitive for quantitative research, and this may be part of why the Delphi Technique is often reserved for qualitative studies. However, we believe that this feature of the technique should be viewed in a very different light within the field of Existential Risk, where confidence in predictions is often overstated.

We believe that enhanced solicitation techniques have a significant underused potential to contribute to the quantification of Existential Risk. They are more rigorous, useful and able to handle uncertainty than individual or aggregated subjective opinions, although they are also harder to implement. A particular attraction of these techniques is their ability to open up a broad range of knowledge and perspectives on risk and to guide experts in combining this into coherent judgements. We rate these methods as High/Medium for rigour and Medium for uncertainty and utility, although we a higher degree of uncertainty for all three categories. However, we rated these techniques as only Low for accessibility due both to the time and expertise required to implement them well.

### 5.3. Prediction markets

Prediction markets function by providing a platform on which people can make trades based on their different assessments of the probability of an outcome or event. The more accurate any person is, the higher their payout. The price at which people are willing to make these trades depends on their probability assessments and their level of certainty in these assessments. This incentivizes individuals to be as rigorous and accurate as possible and allows for aggregation to take place over a potentially unlimited number of

participants. The prediction market [Metaculus](#) uses trades with in-platform credits allowing individuals to perform actions such as posing their own questions. It has set up a market to establish the probability of human extinction (source 10), although the market clearing price, which will represent its 'final' prediction, will not be available until it closes in 2030. However, as Metaculus notes, this market, unlike its others, will not be able to pay out and users are therefore asked to make trades 'in good faith' only. One proposal to overcome this barrier is to build the markets around trade in a resource that would help individuals survive a global catastrophe, such as access to survival shelters ([Hanson, 2008](#)).

[Page and Clemen \(2012\)](#) argue that prediction markets are relatively well-calibrated when the time to expiration is relatively short, but that prices for the future are significantly biased. One might overcome this barrier by establishing markets for events that would be related to, but not necessarily cause, an Existential Risk. For instance, prediction markets could be used to determine the probability that a large asteroid will pass within lunar orbit, that at least one nuclear weapon will be detonated by a non-state actor or some other 'near miss' event that would help us understand Existential Risk without implying that humanity would actually go extinct.<sup>5</sup>

As with all markets, individuals who have limited information may assume that the market is better informed than they are and therefore not bid away from the current market price. This can cause price biases, where it becomes entrenched and prevents markets fully taking account of changing conditions. Nevertheless, prediction markets have proven success at making predictions even under situations of extreme uncertainty, such as whether CERN will locate the Higgs boson ([Pennock, Lawrence, Giles, & Nielsen, 2001](#)).

Prediction markets currently have a strong track record, and there is considerable interest in their use, both amongst experts and as a means of 'democratising' decision making. However, there are significant barriers to their application for Existential Risk. If a suitable platform could be established where participants were shown to have a clear interest in the long term, and their returns were guaranteed against inflation and loss of investment potential, perhaps via a philanthropic backer, then they might have an important part to play in assessing the probability of existential near misses. We rate prediction markets Low for uncertainty and utility and Medium for rigour and accessibility.

## 6. Discussion and recommendations

In this section, we discuss the relative value of each of the methods that we have described above and make some recommendations for how they should be applied, implemented and evaluated by the community of Existential Risk scholars.

### 6.1. Comparing methodologies

There are many methods currently being used, or with potential to be used, to quantify Existential Risk. Each method comes with its advantages and disadvantages, which we summarize in [Table 1](#):

There appear to be no standout 'winners' from this analysis and every technique is rated Low on at least one criterion. The top scorers from our analysis as a whole are Bayesian Networks, adapting existing models and Enhanced Solicitation Techniques, all of which score Low or Very Low in terms of accessibility. Of the more accessible approaches Toy Modelling and Aggregated Opinion Surveys perform best.

Given this variety of methodological virtues, we conclude that method selection should be understood in context and that the suitability of a method to a researcher's needs and circumstances is more important than its overall performance. At present, methodology choice seems to be strongly related to the nature of the studied threat. Some methods may well lend themselves to specific threats, depending on whether they have already been modelled at the sub existential level or whether there is a past historical record on which to build one's analysis. However, we feel that most of these methods could be applied far more widely and that more appropriate determinants of their use are the resources available to a team, whether the research is being undertaken for scientific or policy purposes and how findings are intended to be used.

[Tonn and Stiefel \(2013\)](#) go further and argue for giving the "results of all methods to a panel of experts to reflect upon before they are asked for holistic assessments" (2013). However, that strikes us as potentially problematic because it leads to the homogenization of quite diverse methodological perspectives and the potential loss of insight and introduction of bias that this entails. Realistically, it also represents a further loss of accessibility and therefore may put researchers off from conducting empirical studies to begin with. Hence, we conclude that it would be better to encourage researchers to focus on the methods that are best suited to their particular context and let a thousand flowers bloom.

### 6.2. Structured and deliberative approaches

We believe that the use of structured approaches, and especially enhanced solicitation techniques has been especially underdeveloped within the field of Existential Risk research and that this deserves more attention. While processes like the Delphi technique and superforecasting are not unproblematic, they have developed a good reputation in many scientific circles for being well suited to both interdisciplinary research and making judgements under uncertainty, two of the greatest challenges facing Existential Risk quantification.

In particular, two areas strike us as prime candidates for employing such techniques. Firstly, given the lack of a transparent

<sup>5</sup> We are grateful to Toby Ord for these suggestions.

**Table 1**  
Informal assessment of each method using our four criteria.

Methodology	Rigour	Uncertainty	Accessibility	Utility	Used for
Analytical approaches	Very Low	Very Low	High	Very Low	X-risk
Extrapolation and Toy Modelling	Low	Low	High	Low	Volcanos, Pandemics, Nuclear, Space, Particle Physics, Asteroids
Fault Trees	Medium	Low	Medium*	Medium*	Nuclear, AI
Bayesian Networks	High / Medium	High	Low	Medium*	None
Adapting large-scale models	High	High	Very Low	High	Pandemics, Climate, Asteroids
Individual Subjective Opinion	Low / Medium	Very Low	High	Low	X-risk, Climate, Asteroids, Nuclear
Aggregated Opinion Surveys	Medium	Low	Medium*	Low	AI, Nuclear, X-risk
Weighted Aggregation	Low / Medium	Low	Low	Low	None
Enhanced Solicitation	High / Medium	Medium*	Low	Medium*	Pandemics, Nuclear, X-risk
Prediction Markets	Medium	Low	Medium*	Low	X-risk

methodology for establishing probabilities in the [Pamlin and Armstrong \(2015 – source 18\)](#) report, the Delphi or IDEAs technique may be an appropriate tool should the Global Challenges Foundation seek to update this research. Secondly, given the prevalence of unstructured surveys in the analysis of Artificial Intelligence as an Existential Threat, we believe that a more structured approach to combining expert opinions in this area would be valuable in providing a more rigorous perspective on a controversial subject.

### 6.3. Improving methodologies

Beyond this, however, our study serves to highlight the significant diversity in approaches to the implementation of these methods. There are examples of both good and bad practice in the literature at present and, regrettably, it is not always the good practice that is driving out the bad in the marketplace of ideas. In particular, many of the methods we considered allow researchers to objectively set out their reasoning process for others to critique and potentially update in light of new evidence and most have techniques for assessing and reporting degrees of uncertainty in a judgement. However, in very many cases such opportunities were not taken or were merely paid lip service despite requiring little, if any, additional effort.

The main reasons for not taking advantage of such opportunities are reputational. If one expresses uncertainty, then others are likely to see your judgements as less credible, and if one clearly sets out one's reasoning process, then others may see it as mistaken. These are not good reasons for bad science, and even if there is some argument to be made for simplification in public facing communication, clear statements of methodology and uncertainty should be produced for the research community.

A good example is set by the IPCC who make use of a clear uncertainty framework in their reports. This combines probability judgments and confidence judgements, with separate terms used to describe each. For instance, terms such as 'likely' present a probability judgment, whilst terms like 'confident' are used to present degrees of certainty. According to the IPCC, authors guidance notes: "A level of confidence provides a qualitative synthesis of an author team's judgment about the validity of a finding; it integrates the evaluation of evidence and agreement in one metric" ([Mastrandrea et al., 2011](#)). A potential strength of this approach is that it can be sensitive to particular limitations within a domain, such as the availability of evidence, the level of disagreement about how to interpret that evidence, the robustness of models and methods that are currently used to evaluate it and the overall level of consensus that has been achieved.

Other good examples tend to be set by studies that come out of the 'hard' sciences, including those relating to pandemics and space weather, or those that are embedded in risk analysis, such as the work of Anthony Barrett and Seth Baum on nuclear war. However, in each of these domains, there remain examples of bad, or even discredited, science that are still repeated by Existential Risk researchers, both in public-facing work and academic papers.

## 7. Conclusion

Despite the challenges involved, the quantification of Existential Risk seems highly likely to continue as a prominent strand of research in this area, for risk communication, research prioritization and policy formation. We believe, however, that it is time that researchers in this field became more aware of how they can, and should, go about this process. There are a wide variety of methods that have been tried thus far, and none of these is definitively best, each having both merits and challenges. More importantly though, any of these approaches can be implemented well or badly and the mere fact that a certain probability assessment has been produced does not mean it is worthy of reproduction or inclusion in further analysis.

This is basic science and common sense. However, it is arguable that within the nascent field of Existential Risk research people have been insufficiently discriminating in this regard. This is not only problematic in that it risks using worse results when better ones are available; it also holds back the development of the field by failing to stimulate scholars to improve the quality of assessments that they produce.

## Acknowledgements

This paper was presented at the Cambridge Conference on Catastrophic Risk 2018 and the authors are grateful to the organizers

and audience for their support and feedback. We would also like to thank Shahar Avin, Adrian Currie, Phil Torres and David Manheim for comments on this manuscript. This project was made possible through the support of a grant from Templeton World Charity Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of Templeton World Charity Foundation. GRANT ID#: TWCF0128/AB82 GRANT TITLE: “Managing Extreme Technological Risk”.

## Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:<https://doi.org/10.1016/j.futures.2019.102469>.

## References

- Ahlqvist, T., & Rhisiart, M. (2015). Emerging pathways for critical futures research: Changing contexts and impacts of social theory. *Futures*, 71, 91–104. <https://doi.org/10.1016/j.futures.2015.07.012>.
- Avin, S., Wintle, B. C., Weitzdörfer, J., Ó hÉigeartaigh, S. S., Sutherland, W. J., & Rees, M. J. (2018). Classifying global catastrophic risks. *Futures*, 102, 20–26. <https://doi.org/10.1016/j.futures.2018.02.001>.
- Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, 463(7279), 294. <https://doi.org/10.1038/463294a>.
- Aspinall, W., et al. (2011). *GFDRR, Volcano Risk Study: Volcano Hazard and Exposure in GFDRR Priority Countries and Risk Mitigation Measures Bristol University Cabot Institute and NGI Norway for the World Bank: NGI Report20100806*.
- Atkinson, H., Tickell, C., & Williams, D. (2000). *Report of the task force on potentially hazardous near Earth objects*.
- Bagus, G. (2008). *Pandemic Risk Modelling*. Chicago Actuarial Association.
- Barrett, A. M., Baum, S. D., & Hostetler, K. (2013). Analyzing and reducing the risks of inadvertent nuclear war between the United States and Russia. *Science & Global Security*, 21(2), 106–133. <https://doi.org/10.1080/08929882.2013.798984>.
- Barrett, A. M. (2017). Value of global catastrophic risk (GCR) information: Cost-effectiveness-Based approach for GCR reduction. *Decision Analysis*, 14(3), 187–203. <https://doi.org/10.1287/deca.2017.0350>.
- Barrett, A. M., & Baum, S. D. (2017). A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2), 397–414. <https://doi.org/10.1080/0952813X.2016.1186228>.
- Baum, S., Barrett, A., & Yampolskiy, R. V. (2017). Modeling and interpreting expert disagreement about artificial superintelligence. *Informatica*, 41(7), 419–428.
- Baum, S., de Neufville, R., & Barrett, A. (2018). *A model for the probability of nuclear war*. Global Catastrophic Risk Institute Working Paper, 18-1.
- Bamber, J. L., & Aspinall, W. P. (2013). An expert judgement assessment of future sea level rise from the ice sheets. *Nature Climate Change*, 3(4), 424.
- Bobbio, A., Portinale, L., Minichino, M., & Ciancamerla, E. (2001). Improving the analysis of dependable systems by mapping fault trees into Bayesian networks. *Reliability Engineering & System Safety*, 71(3), 249–260.
- Bostrom, N. (2019). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9.
- Bostrom, N. (2003). Are we living in a computer simulation? *The Philosophical Quarterly*, 53(211), 243–255. <https://doi.org/10.1111/1467-9213.00309>.
- Bostrom, N. (2008). Where are they? *Technology Review*, 111(3).
- Bostrom, N. (2006). Dinosaurs, dodos, humans? *Review of Contemporary Philosophy*, 8, 85–89.
- Burgman, M. A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., ... Twardy, C. (2011). Expert status and performance. *PLoS One*, 6(7), e22998. <https://doi.org/10.1371/journal.pone.0022998>.
- Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, 11(5), 509.
- Chapman, C. R. (2004). The hazard of near-Earth asteroid impacts on earth. *Earth and Planetary Science Letters*, 222(1), 1–15. <https://doi.org/10.1016/j.epsl.2004.03.004>.
- Ćirković, M. M., Sandberg, A., & Bostrom, N. (2010). Anthropocentric shadow: Observation selection effects and human extinction risks. *Risk Analysis: An International Journal*, 30(10), 1495–1506. <https://doi.org/10.1111/j.1539-6924.2010.01460.x>.
- Clemen, R. T. (2008). Comment on Cooke's classical method. *Reliability Engineering & System Safety*, 93(5), 760–765. <https://doi.org/10.1016/j.res.2008.02.003>.
- Colson, A. R., & Cooke, R. M. (2018). Expert elicitation: Using the classical model to validate experts' judgments. *Review of Environmental Economics and Policy*, 12(1), 113–132. <https://doi.org/10.1093/reep/rex022>.
- Condorcet, M. D. (1785). *Essay on the application of analysis to the probability of majority decisions*. Paris: Imprimerie Royale.
- Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press on Demand.
- Currie, A. (2018). *Existential risk, creativity & well-adapted science*. *Studies in History and Philosophy of Science Part A*. <https://doi.org/10.1016/j.shpsa.2018.09.008>.
- Dar, A., De Rijula, A., & Heinze, U. (1999). Will relativistic heavy-ion colliders destroy our planet? *Physics Letters B*, 470(1-4), 142–148. [https://doi.org/10.1016/S0370-2693\(99\)01307-6](https://doi.org/10.1016/S0370-2693(99)01307-6).
- Day, T., André, J. B., & Park, A. (2006). The evolutionary emergence of pandemic influenza. *Proceedings of the Royal Society B: Biological Sciences*, 273(1604), 2945–2953.
- Decker, R. W. (1990). How often does a Minoan eruption occur? *Thera and the Aegean world III*, 2, 444–452.
- Devaney, L., & Henchion, M. (2018). Who is a Delphi ‘expert’? Reflections on a bioeconomy expert selection procedure from Ireland. *Futures*, 99, 45–55. <https://doi.org/10.1016/j.futures.2018.03.017>.
- Dunlop, I., & Spratt, D. (2017). *Disaster alley: Climate change conflict and risk*. Breakthrough – National Centre for Climate Restoration.
- Ellis, J., Giudice, G., Mangano, M., Tkachev, I., & Wiedemann, U. (2008). Review of the safety of LHC collisions. *Journal of Physics G: Nuclear and Particle Physics*, 35(11), 115004. <https://doi.org/10.1088/0954-3899/35/11/115004>.
- Fan, V. Y., Jamison, D. T., & Summers, L. H. (2018). Pandemic risk: how large are the expected losses? *Bulletin of the World Health Organization*, 96(2), 129. <https://doi.org/10.2471/BLT.17.199588>.
- Fouchier, R. A. (2015). Studies on influenza virus transmission between ferrets: the public health risks revisited. *MBio*, 6(1), <https://doi.org/10.1128/mBio.02560-14.e02560-14>.
- Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, 75(7), 450–451. <https://doi.org/10.1038/075450a0>.
- Gehrels, N., Laird, C. M., Jackman, C. H., Cannizzo, J. K., Mattson, B. J., & Chen, W. (2003). Ozone depletion from nearby supernovae. *The Astrophysical Journal*, 585(2), 1169.
- Gott, J. R. (2019). Implications of the Copernican principle for our future prospects. *Nature*, 363(6427), 315–319.
- Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, 62, 729–754. <https://doi.org/10.1613/jair.1.11222>.
- Halstead, J. (2018). Stratospheric aerosol injection research and existential risk. *Futures*, 102, 63–77. <https://doi.org/10.1016/j.futures.2018.03.004>.
- Harris, A. (2008a). What spaceguard did. *Nature*, 453(7199), 1178.
- Harris, B. (2008b). The potential impact of super-volcanic eruptions on the Earth's atmosphere. *Weather*, 63(8), 221–225. <https://doi.org/10.1002/wea.263>.
- Hellman, M. E. (2008). Risk analysis of nuclear deterrence. *The Bent of Tau Beta Pi*, 99(2), 14.
- Hempell, C. M. (2004). The investigation of natural global catastrophes. *Journal-British Interplanetary Society*, 57(1/2), 2–13.
- Homeier, N. (2013). *Solar Storm Risk to the North American Electric Grid*. Lloyd's.

- Hanea, A. M., McBride, M. F., Burgman, M. A., Wintle, B. C., Fidler, F., Flander, L., ... Mascaro, S. (2017). I nvestigate D iscuss E stimate A ggregate for structured expert judgement. *International Journal of Forecasting*, 33(1), 267–279.
- Hanson, R. (2008). Catastrophe, social collapse, and human extinction. In N. Bostrom, & M. M. Cirkovic (Eds.). *Global catastrophic risks* (pp. 363–377). Oxford: Oxford University Press.
- Hubbard, D. W. (2014). *How to measure anything: Finding the value of intangibles in business*. John Wiley & Sons.
- Jaffe, R. L., Busza, W., Wilczek, F., & Sandweiss, J. (2000). Review of speculative “disaster scenarios” at RHIC. *Reviews of Modern Physics*, 72(4), 1125.
- King, D., Schrag, D., Dadi, Z., Ye, Q., & Ghosh, A. (2015). *Climate change: A risk assessment*. Centre for Policy Research. University of Cambridge.
- Klotz, L. C., & Sylvester, E. J. (2014). The consequences of a lab escape of a potential pandemic pathogen. *Frontiers in Public Health*, 2, 116. <https://doi.org/10.3389/fpubh.2014.00116>.
- Khakzad, N., Khan, F., & Amyotte, P. (2011). Safety analysis in process facilities: Comparison of fault tree and Bayesian network approaches. *Reliability Engineering & System Safety*, 96(8), 925–932.
- Leslie, J. (2002). *The End of the World: the science and ethics of human extinction*. Routledge.
- Lipsitch, M., & Inglesby, T. V. (2014). *Moratorium on research intended to create novel potential pandemic pathogens*. <https://doi.org/10.1128/mBio.02366-14>.
- Love, J. J. (2012). Credible occurrence probabilities for extreme geophysical events: Earthquakes, volcanic eruptions, magnetic storms. *Geophysical Research Letters*, 39(10), <https://doi.org/10.1029/2012GL051431>.
- Lundgren, C. (2013). What are the odds? Assessing the probability of a nuclear war. *The Nonproliferation Review*, 20(2), 361–374.
- Li, L., Wang, J., Leung, H., & Jiang, C. (2010). Assessment of catastrophic risk using Bayesian network constructed from domain knowledge and spatial data. *Risk Analysis: An International Journal*, 30(7), 1157–1175. <https://doi.org/10.1111/j.1539-6924.2010.01429.x>.
- Lewis, P. M., Williams, H., Pelopidas, B., & Aghlani, S. (2014). *Too close for comfort: Cases of near nuclear use and options for policy*. Chatham House: The Royal Institute of International Affairs.
- Madhav, N. (2013). *Modelling a modern-day Spanish flu pandemic*. AIR Worldwide21 February.
- Melott, A. L., Lieberman, B. S., Laird, C. M., Martin, L. D., Medvedev, M. V., Thomas, B. C., ... Jackman, C. H. (2004). Did a gamma-ray burst initiate the late Ordovician mass extinction? *International Journal of Astrobiology*, 3(1), 55–61. <https://doi.org/10.1017/S1473550404001910>.
- Metaculus, Metaculus on-line prediction Market - <https://www.metaculus.com/questions/578/human-extinction-by-2100/>.
- Millett, P., & Snyder-Beattie, A. (2017). Existential risk and cost-effective biosecurity. *Health Security*, 15(4), 373–383. <https://doi.org/10.1089/hs.2017.0028>.
- Müller, V. C., & Bostrom, N. (2016). *Future progress in artificial intelligence: A survey of expert opinion*. *Fundamental Issues of Artificial Intelligence*. Springer, Cham555–572.
- Manheim, D. (2018). Questioning estimates for natural pandemic risk. *Health Security*, 16(6), 381–390.
- Mastrandrea, M. D., Mach, K. J., Plattner, G. K., Edenhofer, O., Stocker, T. F., Field, C. B., ... Matschoss, P. R. (2011). The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups. *Climatic Change*, 108(4), 675. <https://doi.org/10.1007/s10584-011-0178-6>.
- Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., ... Murray, T. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5), 1106–1115. <https://doi.org/10.1177/0956797614524255>.
- Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, 111(20), 7176–7184. <https://doi.org/10.1073/pnas.1319946111>.
- National Research Council (2010). *Defending Planet Earth: Near-Earth Object Surveys and Hazard Mitigation Strategies*. The National Academies Press.
- Pamlin, D., & Armstrong, S. (2015). *Global Challenges: 12 Risks that Threaten Human Civilisation*. Global Challenges Foundation.
- Project for the Study of the 21st Century (2015). *Experts see rising risk of nuclear war: survey* <https://www.scribd.com/document/289407938/PS21-Great-Power-Conflict-Report>.
- Riley, P. (2012). On the probability of occurrence of extreme space weather events. *Space Weather*, 10(2), <https://doi.org/10.1029/2011SW000734>.
- Page, L., & Clemen, R. T. (2012). Do prediction markets produce well-calibrated probability forecasts? *The Economic Journal*, 123(568), 491–513. <https://doi.org/10.1111/j.1468-0297.2012.02561.x>.
- Pennock, D. M., Lawrence, S., Giles, C. L., & Nielsen, F. A. (2001). The real power of artificial markets. *Science*, 291(5506), 987–988. <https://doi.org/10.1126/science.291.5506.987>.
- Pindyck, R. S. (2013). Climate change policy: What do the models tell us? *Journal of Economic Literature*, 51(3), 860–872. <https://doi.org/10.1257/jel.51.3.860>.
- Rees, M. (2003). *Our final hour: A scientist's warning: How terror, error, and environmental disaster threaten Humankind's future in this century – On Earth and beyond*. New York: Basic Books.
- Sandberg, A., & Bostrom, N. (2008). Global catastrophic risks survey. *Civil Wars*, 98(30), 4.
- Simpson, F. (2016). *Apocalypse now? Reviving the Doomsday argument*. arXiv preprint arXiv:1611.03072.
- Snyder-Beattie, A. E., Ord, T., & Bonsall, M. B. (2019). An upper bound for the background rate of human extinction. *Scientific reports*, 9(1), 1–9.
- Stern, N., et al. (2006). *Stern Review: The economics of climate change*, 30. London: HM treasury.
- Sagan, C. (1983). Nuclear war and climatic catastrophe: Some policy implications. *Foreign Affairs*, 62(2), 257–292. <https://doi.org/10.2307/20041818>.
- Sunstein, C. R. (2000). Deliberative trouble? Why groups go to extremes. *The Yale Law Journal*, 110(1), 71–119. <https://doi.org/10.2307/797587>.
- Tegmark, M., & Bostrom, N. (2005). Is a doomsday catastrophe likely? *Nature*, 438(7069), 754. <https://doi.org/10.1038/438754a>.
- Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.
- Tetlock, P. E., Mellers, B. A., & Scoblic, J. P. (2017). Bringing probability judgments into policy debates via forecasting tournaments. *Science*, 355(6324), 481–483. <https://doi.org/10.1126/science.aal3147>.
- Tonn, B., & Stiefel, D. (2013). Evaluating methods for estimating existential risks. *Risk Analysis*, 33(10), 1772–1787. <https://doi.org/10.1111/risa.12039>.
- Torres, P., & Beard, S. (2019). *The validity and usefulness of the scientific concept of existential risk*. Forthcoming.
- Turchin, A. (2019). Assessing the future plausibility of catastrophically dangerous AI. *Futures*, 107, 45–58.
- Wagner, G., & Weitzman, M. (2015). *Climate Shock: The Economic Consequences of a Hotter Planet*. Princeton University Press53–56.
- Wells, W. (2009). *Human survivability. Apocalypse When?*. Praxis: Springer Praxis Books[https://doi.org/10.1007/978-0-387-09837-1\\_5](https://doi.org/10.1007/978-0-387-09837-1_5).
- Wintle, B. C., Boehm, C. R., Rhodes, C., Molloy, J. C., Millett, P., Adam, L., ... Doubleday, R. (2017). Point of view: A transatlantic perspective on 20 emerging issues in biological engineering. *Elife*, 6, e30247. <https://doi.org/10.7554/eLife.30247>.
- Woo, G. (2018). Counterfactual disaster risk analysis. *Var. J.* 2, 279–291.
- Xu, Y., & Ramanathan, V. (2017). Well below 2 C: Mitigation strategies for avoiding dangerous to catastrophic climate changes. *Proceedings of the National Academy of Sciences*, 114(39), 10315–10323. <https://doi.org/10.1073/pnas.1618481114>.
- Yampolskiy, R. V. (2018). Predicting future AI failures from historic examples. *Foresight*, 21(1), 138–152. <https://doi.org/10.1108/FS-04-2018-0034>.
- Yudkowsky, E. (2008). Cognitive biases potentially affecting judgment of global risks. *Global catastrophic risks*, 1(86), 13.