An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards

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 1984, 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, Baum & Barrett, 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 1984 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, 2017).

Previous analysis by Bruce Tonn and Dorian 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 (Forthcoming).

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 5,100 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.[1]

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, 2008 – source 51 and Aspinall et al., 2011 – source 52), nuclear wars (Lundgren, 2013 – source 14), space weather (Love, 2012 – source 58, Homelier et al., 2013 – source 60, Melott et al., 2004 – source 61 and Gehrels et al., 2003 – source 62), particle physics experiments (Dar et al., 1999 – source 63, Jaffe et al., 2000 – source 64, Tegmark & Bostrom, 2005 – source 65 and Ellis et al., 2008 - source 66) and the occurrence of extinctions and global catastrophes in general (Hempsell, CM. 2004 -source 1, Synder-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.

1 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.

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), Lipstich and Inglseby (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, 2017) 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 "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 (Cirkovic, Sanbderg, & 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 13, Lewis et al., 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, Manhiem (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 et al., 2013 – source 13)[2] 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 and 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 to provide 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, 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 et al. (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 et al. (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 et al. (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

2 Whereas Barrett et al. (2013) model one set of nuclear war scenarios between Russia and the U.S., Baum et al. (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, S., de Neufville, R., & Barrett, A. (2018). A Model for the Probability of Nuclear War. SSRN Electronic Journal. http://www.doi.org/10.2139/ssrn.3137081

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, 2008 – 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 et al. (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)[3] However, not all subjective opinion should be treated equally and, in the next two sections, we will consider different approaches.

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 Stiefal (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.

3 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.

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 1,000 years, but only a 10% probability that we will not.[4] 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 Yudkowksy (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.

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

4 These figures are based on an illustration produced by the authors.

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 et al., 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, 2014 – source 38 and Grace et al., 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 8). 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 Superforcasting (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 (1992) "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 2,000 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 nothing 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 et al., 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 17) 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., 2016). 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).

Lionel Page and Robert 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.[5]

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 et al., 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.

5 We are grateful to Toby Ord for these suggestions.

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 the following table:

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

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 methodology for establishing probabilities in the Pamlin and Armstrong (2015 – source 17) 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., 2010). 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.

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.

References (excluding those contained in the appended literature review)

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 .

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. (2017). Towards an Integrated Assessment of Global Catastrophic Risk. In Catastrophic and Existential Risk: Proceedings of the First Colloquium, Garrick Institute for the Risk Sciences, University of California, Los Angeles. 41-62.

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. (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).

Buncic, D. (2016). Superforecasting: The Art and Science of Prediction. By Philip Tetlock and Dan Gardner. Risks, 4(3), 24. https://doi.org/10.3390/risks4030024

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.

Ćirković, M. M., Sandberg, A., & Bostrom, N. (2010). Anthropic 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

Clauset, A., Young, M., & Gleditsch, K. S. (2007). On the frequency of severe terrorist events. Journal of Conflict Resolution, 51(1), 58-87. https://doi.org/10.1177/0022002706296157

Clemen, R. T. (2008). Comment on Cooke's classical method. Reliability Engineering & System Safety, 93(5), 760-765. https://doi.org/10.1016/j.ress.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.

Cousien, A., Obach, D., Deuffic-Burban, S., Mostafa, A., Esmat, G., Canva, V., ... & Mohamed, M. K. (2014). Is expert opinion reliable when estimating transition probabilities? The case of HCV-related cirrhosis in Egypt. BMC medical research methodology, 14(1), 39. https://doi.org/10.1186/1471-2288-14-39

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

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

Galton, F. (1907). Vox populi (the wisdom of crowds). Nature, 75(7), 450-451. https://doi.org/10.1038/075450a0

Gryphon Scientific. 2015. Risk and Benefit Analysis of Gain of Function Research, *Draft Final Report*, http://www.gryphonscientific.com/wp-content/uploads/2015/12/Final-Gain-of-Function-Risk-Benefit-Analysis-Report-12.14.2015.pdf, Accessed 19th July 2017.

Hanea, A. M., McBride, M. F., Burgman, M. A., & Wintle, B. C. (2018). The Value of Performance Weights and Discussion in Aggregated Expert Judgments. Risk Analysis, 38(9), 1781-1794. https://doi.org/10.1111/risa.12992

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 Bostrom, N. and Cirkovic, M. M. (Eds.), Global Catastrophic Risks, Oxford University Press, Oxford, pp. 363-377

Hubbard, D.W. (2014). *How to measure anything: Finding the value of intangibles in business*. John Wiley & Sons.

Inglesby, T. V., & Relman, D. A. (2016). How likely is it that biological agents will be used deliberately to cause widespread harm?: Policymakers and scientists need to take seriously the possibility that potential pandemic pathogens will be misused. EMBO reports, 17(2), 127-130. https://doi.org/10.15252/embr.201541674

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.

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.

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

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), 860-72. 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.

Sagan, C. (1983). Nuclear war and climatic catastrophe: Some policy implications. Foreign Affairs, 62(2), 257-292. https://doi.org/10.2307/20041818

Sokolov, A.P., C.A. Schlosser, S. Dutkiewicz, S. Paltsev, D.W. Kicklighter, H.D. Jacoby, R.G. Prinn, C.E. Forest, J.M. Reilly, C. Wang, B. Felzer, M.C. Sarofim, J. Scott, P.H. Stone, J.M. Melillo and J. Cohen (2005): The MIT Integrated Global System Model (IGSM) Version 2: Model Description and Baseline Evaluation. Joint Program Report Series Report 124, 40 pages (http://globalchange.mit.edu/publication/14579).

Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., ... & Midgley, P. M. (2013). Climate change 2013: The physical science basis.

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

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. (2012). The Race for evolutionary success. Sustainability, 4(8), 1787-1805. https://doi.org/10.3390/su4081787

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. Forthcoming. The Validity and Usefulness of the Scientific Concept of Existential Risk

Turchin, A. (2019). Assessing the future plausibility of catastrophically dangerous AI. Futures, 107, 45-58

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.

Wüthrich, N. (2017). Conceptualizing uncertainty: an assessment of the uncertainty framework of the Intergovernmental Panel on Climate Change. In EPSA15 Selected Papers (pp. 95-107). Springer, Cham.

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

Appendix A - An annotated literature review of probability estimates in existential risk

This supplementary appendix provides an annotated literature review of catastrophic and Existential Risk probability claims as well as the methodologies that were used to produce each of them. It is based on a literature search that focused around three catalogues of literature on Existential Risk. The first was "Resources on Existential Risk", compiled by Bruce Schneier in 2015. The second is the extensive bibliography of the 2015 *Global Challenges: 12 Risks that Threaten Human Civilisation* report authored by Dennis Pamlin and Stuart Armstrong (source 17). The third was an early version of the Doomsday Database, a semi-automated system for finding publications that are relevant to x-risks produced by The Existential Risk Research Network and hosted at x-risk.net. Further sources were identified by searching through works cited and via conversations with colleagues at the Centre for the Study of Existential Risk and elsewhere.

Extinction and Collapse Regardless of Cause

The majority of claims made about the likelihood of existential and global catastrophic risks relates to a specific type of threat or risk. However, a number of attempts have been made to put together an overall estimate of the survival chances of humanity. These include the following:

1. Source: Hempsell, CM. (2004). The investigation of natural global catastrophes. Journal of the British interplanetary Society, 57 (1/2), 2 - 13.

   Probability: "A natural sub-critical catastrophe (that would kill over a billion people) would have a 5% to 10%" chance of occurring in the next century."

   Methodology: This is based on an historical study of events that may have caused the death of more than 10% of the human population. These have a likelihood of causing the collapse of organized society without necessarily causing total human extinction. The historical record for the past 2500 years suggests that there may have been two such events, the Caldera Volcano in the 6th century, which precipitated a global volcanic winter, and the "Little Ice Age" in the 14th Century. Both of these events brought about global droughts, famines, epidemics and wars. This suggests that such events may occur about once per millennium.

2. Source: Synder-Beattie, A., Ord, T., & Bonsall, M. (2019). An upper bound for the background rate of human extinction. *Scientific Reports*.

   Probability: Using only the information that Homo sapiens has existed at least 200,000 years, the probability that humanity goes extinct from natural causes in any given year is almost guaranteed to be less than one in 14,000, and likely to be less than one in 87,000. Using the longer track record of survival for our entire genus Homo, the annual probability of natural extinction likely below one in 870,000.

   Methodology: Data from the archaeological and fossil record about the length of time that humanity has survived so far is used to estimate an upper bound on the extinction rate from all natural sources combined, including from sources for which we remain unaware. The modelling was tested against possible forms of observer selection bias and its conclusions were cross-checked against alternative forms of data, including mammalian extinction rates, the temporal ranges of other hominin species, and the frequency of potential catastrophes and mass extinctions.

3. Source: Gott III, J. R. (1993). Implications of the Copernican principle for our future prospects. Nature, 363, 315-319.

Probability: The prior probability that "humanity will cease to exist before 5,100 years or thrive beyond 7.8 million years" is **5%.**

Methodology: This is based on a version of the 'Doomsday Argument', which uses statistical assumptions about the most likely place of any given observer in the totality of human history to draw conclusions about the most likely length of that history. Note that unlike other versions of the doomsday argument, this only concerns the place of an observer in the timespan over which humanity will exist, and not their place within the human population. Other versions of this argument have led to more pessimistic conclusions.

4. Source: Wells, W. (2009) Human survivability. In: Apocalypse When?. Springer Praxis Books. Praxis https://doi.org/10.1007/978-0-387-09837-1_5

   Probability: In 2009 the annual probability of civilizational collapse was 1%, and the conditional probability of human extinction resulting from this was 30-40%. However, this risk will decrease considerably over time.

   Methodology: Wells begins with Gott's version of the Doomsday argument, which leads to the conclusion that human history can be expected to last millions of years. However, Wells argues that Gott was wrong to suggest that the risk of human extinction and civilizational collapse would be evenly distributed across this time period. Instead, by considering both the kinds of risks that humanity presently faces and statistical evidence about the 'lifespan' of our economic and cultural institutions, he argues that the risk of civilizational collapse (and human extinction) is presently very high, but that it will have a relatively short half-life, implying that the level of risk will rapidly diminish over time. From this, he concludes that humanity has an overall life expectancy that is very long, but also that we currently face a very high risk of extinction.

5. Source: Simpson, F. (2016). Apocalypse now? Reviving the Doomsday argument. arXiv preprint arXiv:1611.03072.

   Probability: "Humanity's prognosis for the coming century is well approximated by a global catastrophic risk of **0.2% per year**."

   Methodology: This paper offers a more pessimistic version of the Doomsday argument. It considers the likely position of a given observer in the population of all past and present human beings and responds to many objections to this argument. Note that 0.2% per year is not the 'conclusion' of this argument, since the risk of human extinction will change over time depending on demographic trends; however, it will remain a good approximation for the next 100 years or so.

6. Source: Leslie, J. (2002). *The End of the World: the science and ethics of human extinction.* Routledge.

   Probability: "The probability of the human race avoiding extinction for the next five centuries is encouragingly high, perhaps as high as **70 percent**".

   Methodology: Leslie starts with a version of the Doomsday argument, which he sees as implying that human extinction is likely to occur soon. However, he then argues that all of the foreseeable routes that lead to human extinction are in-fact unlikely and suggests that our chance of survival is probably higher than this argument on its own suggests.

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

   Probability: The probability that "an existential disaster will do us in" is greater than 25%.

   Methodology: This is a purely subjective judgement based on the balance of the evidence, the methodology for weighing this evidence is not discussed. Note that no time frame for this prediction is given because Bostrom is concerned about existential risk in the context of failing to achieve a state of 'technological maturity'.

8. Source: Rees, M. J. (2003). *Our final century*. Basic Books

   Probability: "The odds are no better than **fifty-fifty** that our present civilization on earth will survive to the end of the present century".

   Methodology: This is explicitly stated as a subjective best guest. Attention is then drawn to the number of decisions and choices that could impact on human survival, indicating that this assessment relates specifically to the probability that people will make the wrong decisions.

9. Source: Sandberg, A. & Bostrom, N. (2008): "Global Catastrophic Risks Survey", Technical Report #2008-1, Future of Humanity Institute, Oxford University: pp. 1-5.

   Probability: There is a **19%** chance of overall extinction by 2100.

   Methodology: Median response of an informal survey of 13 participants at the 2008 Oxford Conference on Global Catastrophic Risk. Participants were surveyed on their estimate of human extinction, the death of more than 1 billion people and the death of more than 1 million people from a list of 8 specific threats. However, this list was not taken to be exhaustive.

10. Source: Metaculus on-line prediction Market - https://www.metaculus.com/questions/578/human-extinction-by-2100/

    Probability: Asked "will there be zero living humans on planet earth on January 1, 2100?" the median responder gave a 1% credence in this proposition whilst the (all time) mean response was **8%**.

    Methodology: These are based on 217 predictions by 195 users in addition to 48 anonymous predictions[6]. The final metaculus prediction will not be released until the question closes on February 9th 2030.

11. Source: Stern, N., et al. (2006). Stern Review: The economics of climate change (Vol. 30, p. 2006). London: HM treasury.

    Probability: The probability that our world will exist with people who will be affected by our current choices can be assumed, for the purposes of discounting future wellbeing, to decline at a rate of **0.1%** per year.

    Methodology: Though much quoted, this is presented as no more than a modelling assumption. Stern asserts that 0.1% 'seems high' on the basis that "if this were true, and had been true in the past, it would be remarkable that the human race had lasted this long." However, he also acknowledges that "there is a weak case for still higher levels."

Nuclear War

The predictions listed below relate to the probability of nuclear war, but were largely produced against the context of understanding this as a potential cause of human extinction. The vast majority of existential risk stems from the likelihood of a nuclear winter and only a, as yet undetermined, subset of possible nuclear wars would cause a nuclear winter.

12. Source: Hellman, M. (2008). Risk analysis of nuclear deterrence. The Bent of Tau Beta Pi, 99(2), 14.
    Probability: The annualised probability of a Cuban Missile Type Crisis (CMTC) resulting in World War III is **0.02% - 0.5%**.

    Methodology: The annual probability of a Cuban Missile Type Crisis is given as:

$$\lambda_{\mathrm{CMTC}} = \lambda_{\mathrm{IE}} P_1 P_2 P_3$$

6 These numbers were taken on the 3rd July 2019.

λ_{IE}: Probability of an "initiating event" (a potential first cause of CMTC) is **0.06**. There have been three possible initiating events in the last 50 years of nuclear deterrence: Cuban missiles in 1962, Naval blockade of Cuba in 1980s, and the deployment of American missiles in Eastern Europe. Taking the average rate of these possible initiating events (3 in 50 years), one obtains an annualized probability of an initiating event of 0.06. The category of events seems to be influenced by the author's wish to not be alarmist, and on this basis the paper avoids some possible initiating events, such as the Berlin Crisis of 1961 and the Yom Kippur War of 1973.

P_1: The probability of an initiating event resulting in a CMTC. This is set at 0.33 based on the fact that only one of these three initiating events actually *was* the Cuban missile crisis, though the other two clearly had the potential to trigger an event of equal severity had conditions been slightly different.

P_2: The conditional probability of a CMTC leading to the use of a nuclear weapon. Since this hasn't happened before, the author relies on the reported *subjective probability estimates* from those involved in the Cuban missile crisis (ranging from 0.01 to 0.5). This is a large range. The author's lower bound of the probability estimate, 0.1, accommodates the fact that the participants stated their estimates before the Russian battlefield nuclear weapons were known in the West. This gives an updated range of 0.1 to 0.5.

P_3: The probability that the use of a nuclear weapon results in full scale nuclear war. The author uses reported estimates for this from John F Kennedy and Robert McNamara to arrive at a probability bound of 0.1 to 0.5.

Combining these probabilities, some of which are frequentist and others subjective, leads to the final overall probability of a Cuban Missile Type Crisis leading to a nuclear war.

13. Source: 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
Probability: 90% confidence that the annual probability of accidental nuclear war between the US and Russia is **from 0.001% to 7%**.

    Methodology: The authors use fault tree analysis to explore possible paths towards the initiation of a nuclear war between the USA and Russia. They then uses a range of sources to establish baseline probabilities for each of the fault tree's nodes to produce an estimate for the annual probability of inadvertent war between the United States and Russia (p. 109). It is assumed that the occurrence of mistaken attack indicators are independent random events (p. 113).

14. Source: Lundgren, C. (2013). What are the odds? Assessing the probability of a nuclear war. The Nonproliferation Review, 20(2), 361-374. https://doi.org/10.1080/10736700.2013.799828

    Probability: "The first sixty-six years of the nuclear age produced a **61 percent** chance of a nuclear war" (p. 371).

    Methodology: Bayesian statistical reasoning is used to assess the occurrence likelihood of a number of past counterfactual events and the implied probability that a nuclear war could have started. The author notes that this is "an especially applicable mathematical method of calculating probabilities where only limited data are available and assured knowledge is not possible" (p. 362).

15. Source: *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

    Probability: There is a "**6.8 percent** probability of a major nuclear conflict in the next 25 years killing more people than the Second World War", which is often taken to imply more than 80 million fatalities. Estimates were also given for the likelihood of specific great power conflicts and for the use of nuclear weapons by a non-state actor.

Methodology: A poll of 50 national security experts from around the world was conducted. Participants were asked to estimate the probability of this and other scenarios. Note that the final prediction was produced by averaging the mean and median response to this question, which is not a recognised statistical technique. Critics have also pointed out that in some cases the median is greater than the mean, which is statistically extremely unlikely.

16. Source: Good Judgment Project (nonpublic data, referenced by Carl Shulman: http://effective-altruism.com/ea/1rk/current_estimates_for_likelihood_of_xrisk/)

    Probability: "A median probability of **2%** that a state actor would make a nuclear weapon attack killing at least 1 person before January 1, 2021. Conditional on that happening...an 84% probability of 1-9 weapons detonating, 13% to 10-99, 2% to 100-999, and 1**% to 100 or more**".

    Methodology: Super-forecasters at the Good Judgment Project were commissioned to assess this as part of a set of questions on the future of nuclear weapons by the NGO OpenPhil.

17. Source: Turchin, A. V. (2008) *Structure of the global catastrophe. Risks of human extinction in the XXI century* lulu.com

    Probability: The risk of extinction due to the consequences of nuclear war, or as a result of a 'Doomsday machine', in the 21st century is in the order of 1%.

    Methodology: The author estimates the current annual risk of nuclear war to be 0.5% meaning a 50% chance of nuclear war this century. The scale of nuclear war might be geographically contained and not globally impactful. Therefore, if a nuclear war happens, the author assigns a 50% chance of global civilization degrading to a 'post-apocalyptic stage' in which humanity's existence will be vulnerable to many other extinction factors. Under such conditions, the probability of human extinction is stated to be 10%. However, after 30 years, the risk from nuclear war is assumed to have been made "irrelevant by even more powerful and dangerous technologies". Therefore, Turchin estimates the risk of extinction due to the consequences of nuclear war in the 21st century to be 0.75%.

    A 'Doomsday Machine' is any device, substance or method which could "destroy all of mankind". The author assumes that the probability of creating and applying a 'Doomsday Machine' is 10 times less than the likelihood of conventional nuclear war. However, the conditional probability of extinction if a 'Doomsday Machine' is created is 10 times higher than the conditional probability of extinction if there is a nuclear war. Therefore, the overall risk of extinction due to a 'Doomsday Machine' is the same from nuclear war.

18. Source: Pamlin, D. & Armstrong, S. (2015). Global Challenges: 12 Risks that Threaten Human Civilisation, Global Challenges Foundation.

    Probability: "Based on available assessments the best current estimate for nuclear war within the next 100 years is **5%** for infinite threshold [and] **0.005%** for infinite impact" (p. 148).[7]

    'Infinite impact' refers to the state where civilization collapses and does not recover, or a situation where all human life ends. 'Infinite threshold' refers to a scenario that has the potential to lead to such a collapse, dependent upon other factors (Dennis & Armstrong, 2015: 11).

    Methodology: This is one of a series of risk specific predictions that resulted from a large,

7 Stated sources include: 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.; Hellman, M. E. (2011). How risky is nuclear optimism?. Bulletin of the Atomic Scientists, 67(2), 47-56.; Lundgren, C. (2013). What are the odds? Assessing the probability of a nuclear war. The Nonproliferation Review, 20(2), 361-374.; and the following two risk analysis websites http://metabiota.com/ and http://instedd.org/.

informal, structured expert elicitation exercise conducted by the Global Challenges Foundation. This constituted an "expert review" of the relevant literature for each risk following which "[two] workshops were arranged where the selection of challenges was discussed, one with risk experts in Oxford at the Future of Humanity Institute and the other in London with experts from the financial sector." Based on all the evidence that was gathered, probability estimates were produced for each risk (p. 12).

19. Source: Sandberg, A. & Bostrom, N. (2008): "Global Catastrophic Risks Survey", Technical Report #2008-1, Future of Humanity Institute, Oxford University: pp. 1-5.

    Probability: **1%** chance of human extinction being caused by nuclear war and **0.03%** chance of it being caused by nuclear terrorism.

    Methodology: See source 7, under Extinction and Collapse Regardless of Cause above. Note that for these predictions no time frame was given.

Pandemics

These predictions relate to pandemics that are seen as having the potential to produce a global catastrophe, whether naturally occurring, a result of an accidental laboratory escape, or due to an intentional release of an engineered pathogen with pandemic potential. Extinction from pandemic pathogens would be extremely unlikely given the spread of the human population and natural constraints on the transmission and potency of pathogens; however, they have a great potential to produce civilizational collapse thus potentially triggering other existential catastrophes.

20. Source: 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.

    Probability: The Probability of a pandemic occurring in any given year is 4%. A conservative estimate of the 95% confidence interval for the yearly pandemic probability is 0.7–7.6%.

    Methodology: This probability is derived from combining 'anecdotal' evidence about the number of influenza pandemics over the past 250 years with more recent data about the expected interval between pandemics emerging.[8] Evidence was combined using a well-defined Bayesian formula set out in an appendix to the paper.

21. Source: Madhav, N. (2013). Modelling a modern-day Spanish flu pandemic. AIR Worldwide, February, 21, 2013.

    Probability**:** There is a **0.5-1%** annual probability of a 'modern day Spanish Flu' event, with similar characteristics to the 1918 pandemic including considerable excess deaths amongst young adults. Such a pandemic would likely cause between 21 and 33 million deaths worldwide.

    Methodology: The AIR Pandemic Flu Model combines demographic and epidemiological, and technological modelling to produce a complete model for pandemic influenza. This model has been extensively peer reviewed.

22. Source: 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

    Probability: The annual probability of a severe influenza pandemic (one that increases global mortality by at least 0.1%) is **1.6%** and the average impact of such pandemics is a global mortality increase of 0.58% (±40 million fatalities). Severe flu pandemics represent 95% of the costs associated with all pandemic influenza.

8 The authors back up this claim with evidence from the following sources:Webster, R. G. (1998). Influenza: an emerging disease. Emerging infectious diseases, 4(3), 436. and Reid, A. H., Taubenberger, J. K., & Fanning, T. G. (2004). Evidence of an absence: the genetic origins of the 1918 pandemic influenza virus. Nature Reviews Microbiology, 2(11), 909.

Methodology: The historical record was used to estimate the total frequency and severity of all influenza pandemics and to generate likely age-specific death rates as a result of a global pandemic. The U.S.' historical age distributions, being the most complete, were used as the template for global age distributions. The authors then model the "expected deaths from pandemic influenza risks" with a highly fat tailed distribution of mortality (meaning that the vast majority of deaths occurring from the most severe pandemics).

23. Source: Bagus, G. (2008) Pandemic Risk Modelling. Chicago Actuarial Association

    Probability: A pandemic of the scale of the Spanish Flu, which caused a ±27% increase in global mortality, occurs around once every 420 years. More severe pandemics, which cause a ±42% increase in global mortality, may have a return rate of 2,700 years.

    Methodology: An 'actuarial model' is constructed in the form of a severity curve based on historical data for the past 420 years of influenza outbreaks. This was found to approximate an exponential curve which was then extrapolated to estimate the probability and severity of more extreme pandemics. The model takes account of shifting demographic features over time but assumes that pandemics have equal severity across all countries.

24. Source: 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

    Probability: The likelihood of a pandemic, through an undetected lab-acquired infection, "could be as high as 27%" over a 10-year research period.

    Methodology: The authors take the annual probability per lab of an escape of a virus through an undetected lab-acquired infection (LAI) to be 2.4%. This statistic is taken from the Department of Homeland Security's risk assessment for a planned National Bio- and Agro-defence Facility in Manhattan, Kansa. They then assume that a research enterprise will comprise of 10 labs working for 10 years to make a virus. So, across this period, the probability of no escape through an LAI will be 0.088. Therefore, the probability of at least one escape from the enterprise through an LAI will be 91%. This is multiplied by the assumed likelihood, a worst-case scenario, of one LAI leading to a pandemic, 30%, to give the overall prediction.

25. Source: Lipsitch, M., & Inglesby, T. V. (2014). Moratorium on research intended to create novel potential pandemic pathogens. https://doi.org/10.1128/mBio.02366-14
    Probability: Each laboratory-year of Gain of Function research into virulent, transmissible influenza virus might have an **0.01% to 0.1%** chance of triggering a global infection via an accidental laboratory escape. Such a pandemic could be expected to kill between 2 million and 1.4 billion people.

    Methodology: The risk of a global pandemic resulting from a laboratory escape of influenza is determined from multiplying two different probabilities. The first is the risk of laboratory incidents and accidental infections in biosafety level 3 laboratories in which such research may be conducted (estimated to be between 0.2%, on the basis that 4 infections have been observed over <2,044 laboratory-years of observation, and 1%, using data from the National Institute of Allergies and Infectious Diseases). The second is the probability that an accidental infection of a working lab could lead to a laboratory escape spreading widely around the world (estimated to be between 5% and 60% according to a range of simulation models, with the authors' own model indicating a 10-20% risk).

    Noting that "readily transmissible influenza, once widespread, has never before been controlled before it spreads globally," the expected severity of such a pandemic is determined by multiplying the historical infection rate of influenza pandemics (24-38%) by possible values for the case-rate fatality of a novel, virulent influenza strain (1-60%). However, it is unlikely that these two figures vary independently and so simple multiplication is likely to be inappropriate.

26. Source: Fouchier, R. A. (2015). Studies on influenza virus transmission between ferrets: the public health risks revisited. MBio, 6(1), e02560-14. https://doi.org/10.1128/mBio.02560-14

    Probability: Each laboratory-year of Gain of Function research into virulent, transmissible influenza virus might have an **2.5×10^{-13} to 3×10^{-12}** chance of triggering a global infection via an accidental laboratory escape.

    Methodology: This paper is a direct response to Lipsitch and Inglesby (2014). It argues that their estimates "were based on historical data and did not take into account the numerous risk reduction measures that are now in place in the laboratories where the research is conducted."

27. Source: 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

    Probability: The annual probability of an existential catastrophe arising from a global pandemic is between **8×10^{-5} and 1.6×10^{-8}**.

    Methodology: The authors construct a toy model to assess this risk, and cite a Gryphon Scientific report (2015) to suggest that the annual probability of a global pandemic arising from an accident with research into Potentially Pandemic Pathogens (PPP) in the US is 0.002% to 0.1%.[9] Next, they note that: "The Gryphon report also concluded that the risks of deliberate misuse were about as serious as the risks of an accidental outbreak, suggesting a twofold increase in risk. Assuming that 25% of relevant research is done in the US as opposed to elsewhere in the world, gives us a further fourfold increase in risk. In total, this eightfold increase in risk gives us a 0.016% to 0.8% chance of a pandemic in the future each year."

    Next, the authors directly estimate the probability that a pandemic will cause an existential catastrophe and combine with this with the previous probability: "For the purposes of this model, we assume that for any global pandemic arising from this kind of research, each has only a one in ten thousand chance of causing an rxistential risk."[10]

28. Source: 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

    Probability: The annual probability of an existential catastrophe resulting from biowarfare or bioterrorism is **0.0000019 (or 1.9×10^{-6})**.

    Methodology: The authors assume that the casualty numbers from terrorism and warfare follow a power law distribution. Previous studies have determined the power law exponent for terrorism using chemical or biological weapons to be -0.5. This means that for every order of magnitude increase in casualties from a terrorist attack, the probability of that attack occurring is multiplied by a factor $10^{-0.5}$, which is approximately 1/3. Assuming one attack per year, the annual probability that an attack kills more than 5 billion people will be $(5 \text{ billion})^{-0.5}$, which is 0.000014 or 1.4×10^{-5}. Historical data gives the power law exponent for warfare to be 0.41. The authors assume 1 new war every other year and that bioweapons are used in 10% of wars. Therefore, the annual probability that a war involving biological weapons kills more than 5 billion people is $0.5 \times 0.1 \times (5 \text{ billion})^{-0.41}$, which is 0.000005 or 5×10^{-6}. The authors assume that of all wars or terrorist attacks that kill more than 5 billion people, 10% of these would lead to extinction. Therefore, the authors reach an annual probability of existential catastrophe from biowarfare or bioterrorism of 1.9×10^{-6}.

9 There is no explicit reference to these particular probabilities in the original report.

10 The authors state that this figure is a "conservative guess". It is not precisely clear whether the authors mean that one in ten thousand pandemics are predicted to *cause* extinction, or whether one in ten pandemics will have a *risk* of extinction. The latter reading is implausible because surely there is at least a risk, however small, that any global pandemic would cause extinction.

29. Source: Manheim D. (2018). Eliciting evaluations of existential risk from infectious disease. Unpublished https://www.openphilanthropy.org/focus/global-catastrophic-risks/biosecurity/david-manheim-research-existential-risk

    Probability: "The risk of global catastrophic risk from natural diseases, killing more than 1 billion people, was 'plausible' but 'very unlikely'. All [experts] who were willing to comment about a numeric estimate agreed that this probability was between 10-9/year and 10-15/year - but were explicitly unwilling to have that claim be attributed to them, or even reported in a published paper that listed their agency's name."

    Methodology:A structured expert solicitation from US government agencies and academics, selected using snowball sampling. The solicitation process involved four steps. Firstly, the definition of a global catastrophic biological risk was introduced and discussed. Secondly, experts were asked to suggest possible scenarios involving such a risk. Thirdly, they were asked to comment on a range of other possible scenarios, to comment on their plausibility and possibility (providing numerical estimates where able) and to consider a possible risk timeline. Finally, the experts offered their views on the most effective mitigation methods and possible roadblocks. The risks of engineered pathogens were also discussed with some of the experts but less well explored.

30. Source: Sandberg, A. & Bostrom, N. (2008): "Global Catastrophic Risks Survey", Technical Report #2008-1, Future of Humanity Institute, Oxford University: pp. 1-5.

    Probability: 2**%** chance of human extinction being caused by an engineered pandemic and **0.05%** chance of it being caused by a natural pandemic.

    Methodology: See source 7, under Extinction and Collapse Regardless of Cause above. Note that for these predictions no time frame was given.

31. Source: Pamlin, D. & Armstrong, S. (2015). Global Challenges: 12 Risks that Threaten Human Civilisation, Global Challenges Foundation

    Probability: "Based on available assessments the best current estimate of a global pandemic in the next 100 years is: **5%** for infinite threshold [and] **0.0001%** for infinite impact" (p. 150).[11]

    Methodology: See source 17, under Nuclear War section above.

Climate Change

Climate change potentially constitutes an existential and catastrophic risk for three reasons. Firstly, it is possible that global average temperatures increase far enough to cause uninhabitable conditions for humanity's continued survival, especially if it leads to us crossing a climatic tipping point. Secondly, climate change may reduce the resilience of global systems making us vulnerable to other kinds of

11 Stated sources include: Bagus, Ghalid (2008): Pandemic Risk Modeling http://www.chicagoactuarialassociation.org/CAA_PandemicRiskModelingBagus_Jun08.pdf; Broekhoven, H. V. & Hellman, A. (2006): Actuarial reflections on pandemic risk and its consequences http://actuary.eu/documents/pandemics_web.pdf; Brockmann, D. & Helbing, D. (2013): The Hidden Geometry of Complex, Network-Driven Contagion Phenomena SCIENCE VOL 342 http://rocs.hu-berlin.de/resources/HiddenGeometryPaper.pdf, W. Bruine de Bruin, B. Fischhoff; L. Brilliant and D. Caruso (2006): Expert judgments of pandemic influenza risks, Global Public Health, June 2006; 1(2): 178193 http://www.cmu.edu/dietrich/sds/docs/fischhoff/AF-GPH.pdf; Khan K, Sears J, Hu VW, Brownstein JS, Hay S, Kossowsky D, Eckhardt R, Chim T, Berry I, Bogoch I, Cetron M.: Potential for the International Spread of Middle East Respiratory Syndrome in Association with Mass Gatherings in Saudi Arabia. PLOS Currents utbreaks. 2013 Jul 17. http://currents.plos.org/outbreaks/article/assessing-riskfor-the-international-spread-of-middle-east-respiratorysyndrome-in-association-with-mass-gatherings-insaudi-arabia/; Murray, Christopher JL, et al.: Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918–20 pandemic: a quantitative analysis. The Lancet 368.9554 (2007): 2211-2218. http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(06)69895-4/fulltext and Sandman , Peter M. (2007): Talking about a flu pandemic worst-case scenario, http://www.cidrap.umn.edu/newsperspective/2007/03/talking-about-flu-pandemic-worstcase-scenario

existential threat. Finally, the efforts we take to mitigate and adapt to climate change, most notably through radical geoengineering, may pose their own risks. The majority of predictions about the existential risk posed by climate change focus solely on the first of these, especiallyon the risk of 'extreme' global warming according to existing climate models used by the IPCC. However, it is worth noting that there exists no consensus over where exactly the threshold for dangerous climate change should be set, from 2 to 10 degrees Celsius above the pre-industrial average.

32. Source: Wagner, G. & Weitzman, M. (2015). *Climate Shock: The Economic Consequences of a Hotter Planet*(pp. 53-56). Princeton University Press

    Probability: On a low-medium emissions scenario, there is **at least a 3% chance** of eventual 6 degrees warming (with significant uncertainty). On a medium-high emissions scenario, the chance could be **around 10%.**

    Methodology: This is based on the authors' assessment of the known levels of uncertainty (as classified by the IPCC) about the climate sensitivity to carbon and their implications for current predictions of climate change and the likely distribution of future scenarios within these. The authors also consider the existential risk from geoengineering but make no comment about the likelihood of this.

33. Source: King, D., Schrag, D., Dadi, Z., Ye, Q., & Ghosh, A. (2015). *Climate change: A risk assessment.* Centre for Policy Research, University of Cambridge.

    Probability: For any emissions pathway, a wide range of global temperature increases is possible. On all but the lowest emissions pathways, the probability of a rise of more than 2°C is greater than 80% in the latter half of this century. On a medium-high emissions pathway, there is a 50% probability of a rise of more than 4°C by 2150. On the highest emissions pathway, a rise of 7°C has a very low probability at the end of this century, but exceeds 50% during the course of the 22nd century and rises to 65% in the following century. A rise of more than 10°C over the next few centuries cannot be ruled out.

    Methodology: These estimates are produced using a model created by the AVOID2 research project which reanalyses existing IPCC data to take account of 1) a more differentiated understanding of uncertainty about climate sensitivity to carbon rather than treating this as a single variable, 2) additional earth feedback systems that were excluded from the IPCCs assessment and 3) New Emissions Pathways. The authors note that "alternative model set-ups may show small differences in these probabilities, but the conclusions will be qualitatively the same."

34. Source: Dunlop, I. & Spratt, D. (2017). *Disaster Alley*: *Climate Change Conflict and Risk*, Breakthrough – National Centre for Climate Restoration

    Probability: **50% chance** of 'catastrophic' climate change, defined as a global temperature increase above 4 degrees Celsius, even if nations meet their commitments under the Paris Agreement. The authors argue that such an increase is "incompatible with an organized global community".

    Methodology: The authors aim to correct what they see as flaws in the IPCC modelling that produce an overly optimistic assessment of the likely implications of policy choices on climate change. This includes the omission of "longer-term" carbon cycle feedbacks, such as permafrost thaw and terrestrial carbon sinks, which they say are now becoming relevant.

    The claim that 4 degrees of warming would lead to a breakdown in world order is based on an assessment of its wider systemic effects, although this is preliminary and unquantified.

35. Source: 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

Probability: "Within eight decades, the warming has a **50% probability** of subjecting the global population to catastrophic (>3 degrees) to unknown risks (>5 degrees) and a **5% probability** of being fully in the unknown risk category, which also includes existential threats for everyone"

Methodology: These predictions, which come from two widely respected climate scientists, are based on an amended version of the IPCCs existing climate models, designed to improve its ability to forecast 'catastrophic' climate change and to better capture the uncertainties in emissions scenarios in the original IPCC data.

36. Source: Halstead, J. (2018). Stratospheric aerosol injection research and existential risk. Futures, 102, 63-77. https://doi.org/10.1016/j.futures.2018.03.004

    Probability: "The probability of existential catastrophe-level warming is **±3.5%**". This is taken to be warming in excess of 10 degrees Celsius.

    Methodology: This is the author's best guess based upon reasonable assumptions about two key factors 1) the probability of different concentrations of greenhouse gases in the earth's atmosphere and 2) the conditional probability, given each of these concentrations, that warming may exceed 10 degrees, which is assumed to be the threshold for climate change to pose an existential risk to humanity. This probability is used to inform a further assessment of the expected costs and benefits of further research into Stratospheric Aerosol Injection as a potential means of mitigating climate change.

37. Source: Pamlin, D. & Armstrong, S. (2015). Global Challenges: 12 Risks that Threaten Human Civilisation, Global Challenges Foundation

    Probability: "Based on available assessments the best current estimate for extreme climate change in the next 200 years is: **5%** for infinite threshold [and] **0.01%** for infinite impact" (p. 146).[12]

    Methodology: See source 17, under the Nuclear War section above.

Artificial Intelligence

In coming decades, there is a high chance that Artificial Intelligence (AI) will surpass that of human intelligence. There are catastrophic and existential risks associated with superintelligent AI severely harming or causing the extinction of the human species.

38. Source: Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In Fundamental issues of artificial intelligence (pp. 555-572). Springer, Cham.

    Probability: Mean credence of experts in the field is that the probability that human level machine intelligence would lead to extinction is **18%**.

12 Stated sources include: Fekete, Hanna, et al. (2013): Analysis of current greenhouse gas emission trends http://climateactiontracker.org/assets/publications/publications/CAT_Trend_Report.pdf; New, Mark G, et al. (2011): Four degrees and beyond: the potential for a global temperature increase of four degrees and its implications http://rsta.royalsocietypublishing.org/content/369/1934.toc; Rogelj, Joeri. (2013): Risk shifts under changing climate sensitivity estimates, http://global-risk-indicator.net/data/pdf_01.pdf; Schneider, Stephen H. (2005): What is the Probability of 'Dangerous' Climate Change? http://stephenschneider.stanford.edu/Climate/Climate_Impacts/WhatIsTheProbability.html; Sokolov, A. P., Stone, P. H., Forest, C. E., Prinn, R., Sarofim, M. C., Webster, M., ... & Reilly, J. (2009). Probabilistic forecast for twenty-first-century climate based on uncertainties in emissions (without policy) and climate parameters. Journal of Climate, 22(19), 5175-5204.http://journals.ametsoc.org/doi/abs/10.1175/2009JCLI2863.1 and World Bank (2013): Turn Down the Heat: Climate Extremes, Regional Impacts, and the Case for Resilience http://www.worldbank.org/content/dam/Worldbank/document/Full_Report_Vol_2_Turn_Down_The_Heat_%20Climate_Extremes_Regional_Impacts_Case_for_Resilience_Print%20version_FINAL.pdf

Methodology: A survey was given to 550 experts with different backgrounds in AI and 170 responded. The percentage results are means.. The four groups that were asked were; (1) participants of the conference on "Philosophy and Theory of AI", Thessaloniki October 2011; (2) Participants of the conferences of "Artificial General Intelligence" 2012 & (3) "Impacts and Risks of Artificial General Intelligence" 2012, both at Oxford and (4) Members of the Greek Association for Artificial Intelligence in 2013.

The following question was asked as part of the survey:
"4. Assume for the purpose of this question that such Human Level Machine Intelligence (HLMI) will at some point exist. How positive or negative would be overall impact on humanity, in the long run? Please indicate a probability for each option. (The sum should be equal to 100%.)" – Respondents had to select a probability for each option (in 1% increments). The addition of the selection was displayed; in green if the sum was 100%, otherwise in red. The five options were: "Extremely good – On balance good – More or less neutral – On balance bad – Extremely bad (existential catastrophe)".

39. Source: 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

    Probability: Asked whether HLMI would have a positive or negative impact on humanity, the "probability was **10%** for a bad outcome and **5%** for an outcome described as 'Extremely Bad' (e.g. human extinction" (p. 4).

    Methodology: A survey was conducted of 352 research 'experts' (individuals who published at two of the "premier venues for peer-reviewed research in machine learning", 2015 NIPS and ICML conferences). Respondents assigned probabilities to outcomes on a 5 point scale. Median probabilities are given.

40. Source: Baum, S., Barrett, A., & Yampolskiy, R. V. (2017). Modeling and interpreting expert disagreement about artificial superintelligence. Informatica, 41(7), 419-428.

    Probability: The probability of a catastrophe resulting from artificial superintelligence is $P \approx 0.25$.

    Methodology: Nick Bostrom and Ben Goertzel are both leading thinkers on the topic of artificial superintelligence (ASI), but they disagree about the risk of ASI catastrophe. A fault tree model individually quantifies the probability, P, of ASI catastrophe resulting from Bostrom's ($P \approx 0.51$) and Goertzel's ($P = 0.1$) arguments. When both experts' arguments are combined, $P \approx 0.25$. However, the authors note that "these numbers come with many caveats and should be used mainly for illustration and discussion purposes."

41. Source: Sandberg, A. & Bostrom, N. (2008): "Global Catastrophic Risks Survey", Technical Report #2008-1, Future of Humanity Institute, Oxford University: pp. 1-5.

    Probabiity: 5**%** chance of human extinction being caused by superintelligent AI

    Methodology: See source 7, under Extinction and Collapse Regardless of Cause above. Note that for these predictions no time frame was given.

42. Source: Pamlin, D. & Armstrong, S. (2015). Global Challenges: 12 Risks that Threaten Human Civilisation, Global Challenges Foundation

    Probability: "Based on available assessments the best current estimate of an impact from AI in

the next 100 years is: **0-10%** for infinite threshold [and] **0-10%** for infinite impact" (p. 158).[13]

Methodology: See source 17, under the Nuclear War section above.

Asteroid Impact

The predictions below consider the possibility of catastrophic asteroid impacts. Of potential asteroid impacts, only a very small subset are considered to be existential risks, primarily due to their short to medium term impact on the global climate.

43. Source: NASA's Centre for Near-Earth Object Studies (CNEOS), https://cneos.jpl.nasa.gov/

    Probability: The real time probability that large space objects may hit the earth is calculated. For instance, in 2017 there was a **1 in 63,000** chance (**0.000016**) that asteroid 2013 TV135 would hit earth. The (expected) impact would have had the kinetic energy of 3,200 megatons of TNT, approximately 60 times that of the most powerful nuclear bomb ever detonated.

    Methodology: Expected trajectories for all near-earth objects are modelled by NASA, which can calculate the probability that they will hit earth based on their expected trajectory and known levels of uncertainty within the model.

44. Source: Atkinson, H., Tickell, C., & Williams, D. (2000). Report of the task force on potentially hazardous near Earth objects.

    Probability: 600m diameter Near Earth Objects strike the earth approximately every **70,000 years**, would have global impacts, and could cause 1.5 billion fatalities. 10km diameter Near Earth Objects strike the earth approximately every **100 million years** and could cause a global mass extinction, including more than 6 billion fatalities.

    Methodology: Estimates are based on the known density and trajectory of all near earth objects that have been detected thus far (400 objects in excess of 1km at the time of writing) and the estimated uncertainty about how many objects are yet to be discovered and studied.

45. Source: 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

    Probability: The chances of a 'civilization destroyer' 2-3km diameter asteroid are "probably < 1 in 100,000 during the next century", (p. 11).

    Methodology: This estimate represents the considered judgement of the author based on the available evidence concerning the distribution of near-earth objects and the impact of a collision. It reflects the fact that few near earth objects of this size are likely to have remained undiscovered, and that all those which have been discovered have been shown to be safe.

46. Source: Bostrom, N. (2009). Dinosaurs, dodos, humans?. Review of Contemporary Philosophy, (8), 85-89.
    Probability: "A meteor or an asteroid would have to be considerably larger than 1km in diameter to pose an existential risk. Fortunately, **such objects hit the Earth less than once in 500,000 years on average**." (p. 3).

13 Stated sources include: Armstrong, S. & Sotala, K. (2012).: How We're Predicting AI—or Failing To. In Beyond AI: Artificial Dreams, edited by Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster, 52–75; Chalmers, D., (2010). The singularity: A philosophical analysis. Journal of Consciousness Studies 17.9-10: 9-10; Muehlhauser, L. & Salamon, A. (2012): Intelligence explosion: Evidence and import. Singularity Hypotheses. Springer Berlin Heidelberg 15-42; Yudkowsky, E.(2013): Intelligence Explosion Microeconomics. Technical report 2013-1 Berkeley, CA: Machine Intelligence Research Institute; Wilson, G. (2012).: Minimizing global catastrophic and existential risks from emerging technologies through international law. Available at SSRN 2179094 and Kruel, A. (2013): Probability of unfriendly and friendly AI.

Methodology: This comes from an extrapolation of historical data about meteorite strikes on the earth, although no precise source is given.

47. Source: Harris, A. (2008). What spaceguard did. nature, 453(7199), 1178.

    Probability: The probability of being killed in an "Impact mass extinction" is 1 in 4.3 million.

    Methodology: Harris derives this probability estimate from the 'Spaceguard Survey', which is a census designed to discover and study near-Earth objects (NEOs), especially those that may impact Earth. The paper notes that "the impact of an asteroid larger than 1 kilometre in diameter has the potential to cause a global climatic perturbation, similar to a 'nuclear winter', and could lead to "billions of deaths worldwide". As of October 2017[14], Spaceguard has identified 872 NEOs larger than 1km in diameter, with an implied survey completion percentage of 93%.

48. Source: National Research Council. (2010). *Defending Planet Earth: Near-Earth Object Surveys and Hazard Mitigation Strategies,* The National Academies Press

    Probability: A 1km diameter near-Earth object that would cause a "potential global catastrophe" is expected to hit the Earth every 700,000 years and a 5km near-Earth object that is "above global catastrophe threshold" is expected to hit the Earth every 30 million years.

    Methodology: The cumulative number of NEOs with diameters greater than D kilometres was found to be $942D^{-2.354}$. The distribution of NEOs is determined from dynamical modelling and comparing the resultant probability distributions with observations. This gives the cumulative number of NEOs with diameters greater than D kilometres as 942D-2.354. The authors use the size and distribution of NEOs to approximate the expected time intervals between different impact events.

49. Source: Pamlin, D. & Armstrong, S. (2015). Global Challenges: 12 Risks that Threaten Human Civilisation, Global Challenges Foundation

    Probability: "Based on available assessments the best current estimate of a major asteroid impact in the next 100 years is: **0.01%** for infinite threshold [and] **0.00013%** for infinite impact" (p. 156).[15]

    Methodology: See source 17, under Nuclear War section above.

Super-volcanic eruption

A super volcanic eruption, known technically as a VEI 8 eruption, is one with an ejective volume in excess of 1,000 km^3. The principle means in which a super-volcanic eruption might lead to human extinction is via the production of a 'volcanic winter' and related climatic impacts. It is worth noting that

14 Harris, A. (2017) *The Population of Near-Earth Asteroids Revisited* American Astronomical Society, DPS meeting #49, id.100.01

15 Stated sources for the estimates include: Jablonski, D. & Chaloner, W. G. : Extinctions in the Fossil Record [and Discussion]. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 344.1307 (1994): 11-17; Collins, G. S., Melosh, H.J., &. Marcus, R. A.: Earth Impact Effects Program: A Web based computer program for calculating the regional environmental consequences of a meteoroid impact on Earth. Meteoritics & planetary science 40.6 (2005): 817-840; BSpace Studies Board.: Defending Planet Earth: Near-Earth Object Surveys and Hazard Mitigation Strategies. National Academies Press (2010). http://neo.jpl.nasa.gov/risks/; Neukum, G., and B. A. Ivanov. (1994): Crater size distributions and impact probabilities on Earth from lunar, terrestrial-planet, and asteroid cratering data. Hazards due to Comets and Asteroids 359; Chodas, P., and D. Yeomans. (1999): Orbit determination and estimation of impact probability for near-Earth objects; Chapman, R. C. & Morrison, D. (1994): Impacts on the Earth by asteroids and comets: assessing the hazard. Nature 367.6458 and Chapman, R. C., Durda, D. D., & Gold, R.E. (2001): The comet/asteroid impact hazard: a systems approach. Office of Space Studies, Southwest Research Institute, Boulder CO 80302.

at least 5 VEI 8 eruptions have occurred since the emergence of hominids, of which only the largest, Lake Toba, with an ejective volume of 2,800km^3, is thought to have posed any significant threat (although this is highly contested). Previous mass extinction events are associated with highly active periods of volcanism that produce Large Igneous Provinces millions of cubic kilometres in volume.

50. Source: Decker, R. W. (1990). How often does a Minoan eruption occur? *Thera and the Aegean world III*, 2, 444-452.

    Probability: A VEI 8 eruption occurs about **once every 50,000 years**.

    Methodology: This study was based on geological evidence for 8 previous super volcanic eruptions. However, it has subsequently been argued that this analysis was over subjective and overstated the magnitude of 6 of these eruptions. This is therefore no longer a widely accepted claim amongst volcanologists, but the claim nevertheless continues to circulate in other communities.

51. Source: Harris, B. (2008). The potential impact of super‐volcanic eruptions on the Earth's atmosphere. Weather, 63(8), 221-225. https://doi.org/10.1002/wea.263 [16]

    Probability: **75%** probability of a VEI 8 eruption occurring within the next 1 million years, and a **1%** probability of such an eruption occurring in the next 460-7200 years. (p. 222) This equates to an annual probability range of 1.5×10^{-6} to 2×10^{-5}

    Methodology: Frequentist probabilities are derived from observations about past eruptions in the historical and geological record.

52. Source: 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 Report 20100806

    Probability: A VEI 8 eruption occurs every 30,000 years.

    Methodology: The recurrence rates of explosive volcanism were estimated based on analysis of a global database of large magnitude explosive volcanic eruptions (LaMEVE) at the University of Bristol. The LaMEVE dataset consists of information on magnitudes and ages of explosive eruptions. Historically recorder data are largely confined to explosions that occured less than 500 years ago, whilst data from 500-40,000 years ago are derived from radiocarbon. Data for events more than 40,000 years ago are derived from other methods including from the geological record.

53. Source: Pamlin, D. & Armstrong, S. (2015). Global Challenges: 12 Risks that Threaten Human Civilisation, Global Challenges Foundation

    Probability: "Based on available assessments the best current estimate of a super-volcano in the next 100 years is: **0.002%** for infinite threshold [and] **0.00003%** for infinite impact" (p. 158).[17]

    Methodology: See source 17, under Nuclear War section above.

Ecological Catastrophe

16 Although not explicitly referenced, the probability figures from Harris (2008) can be found in Ben Mason et. al, 2004, "The Size and Frequency of the Largest Explosive Eruptions on Earth", *Bulletin of Volcanology*, 66, pp. 735-48, p. 745.

17 Stated sources include: Aspinall, W., et al.: Volcano hazard and exposure in GDRFF priority countries and risk mitigation measures-GFDRR Volcano Risk Study. Bristol: Bristol University Cabot Institute and NGI Norway for the World Bank: NGI Report (2011) and Mason, B. G., Pyle, D. M., & Oppenheimer, C. (2004): The size and frequency of the largest explosive eruptions on Earth. Bulletin of Volcanology 66.8 : 735-748.

This risk refers to the chance that ecological systems which sustain human life will collapse. This can have potentially disastrous impacts on human and animal populations. There is a close connection with catastrophic climate change.

54. Source: Pamlin, D. & Armstrong, S. (2015). Global Challenges: 12 Risks that Threaten Human Civilisation, Global Challenges Foundation

    Probability: "Based on available assessments the best current estimate of an ecological catastrophe in the next 100 years: **0.5%** for infinite threshold".

    Methodology: See source 17, under Nuclear War section above.

Synthetic Biology

Synthetic biology is the design and construction of biological devices for useful purposes. A potentially devastating impact of synthetic biology would be the release of an engineered pathogen that targeted humans or ecosystems. There is a great deal of uncertainty about these risks.

55. Source: Pamlin, D. & Armstrong, S. (2015). Global Challenges: 12 Risks that Threaten Human Civilisation, Global Challenges Foundation

    Probability: "Based on available assessments the best current estimate of an impact from synthetic biology in the next 100 years is: **1%** for infinite threshold [and] **0.01%** for infinite impact" (p. 160).[18]

    Methodology: See source 17, under Nuclear War section above.

Nanotechnology

Nanotechnology is the branch of technology that deals with applications on the atomic level. Potential examples that are relevant to existential and catastrophic risk include a nanotechnology arms race and rapid uranium extraction for the construction of nuclear bombs. There is a great deal of uncertainty about whether this technology will lead to catastrophic consequences.

56. Source: Sandberg, A. & Bostrom, N. (2008): "Global Catastrophic Risks Survey", Technical Report #2008-1, Future of Humanity Institute, Oxford University: pp. 1-5.

    Probability: **5%** chance of human extinction being caused by nanotech weapons and **0.05%** chance of it being caused by a nanotech accident.

    Methodology: See source 7, under Extinction and Collapse Regardless of Cause above. Note that for these predictions no time frame was given.

57. Source: Pamlin, D. & Armstrong, S. (2015). Global Challenges: 12 Risks that Threaten Human Civilisation, Global Challenges Foundation

    Probability: "Based on available assessments the best current estimate of an impact from nanotechnology in the next 100 years is: **0.8%** for infinite threshold [and] **0.01%** for infinite impact" (p. 160).[19]

    Methodology: See source 17, under Nuclear War section above.

18 Stated sources include: Bennett, Gaymon, et al (2009). "From synthetic biology to biohacking: are we prepared?" Nature Biotechnology 27.12: 1109-1111.; Mukunda, G., Oye, K.A., & Mohr, S.C. (2009). What rough beast? Synthetic biology, uncertainty, and the future of biosecurity. *Politics and the Life Sciences* 28.2: 2-26; Russ, Z. N. (2008). Synthetic biology: enormous possibility, exaggerated perils." *Journal of biological engineering* 2.7; Radosavljevic, V., & Belojevic, G. (2009). A new model of bioterrorism risk assessment. *Biosecurity and bioterrorism: biodefense strategy, practice, and science* 7.4: 443-451 and Tucker, J. B., & Zilinskas, R. A. (2006). The promise and perils of synthetic biology. *New Atlantis* 12.1): 25-45.

Space Weather

The following risks relate to our vulnerability to the radiation emitted by stars, either our own sun or more powerful emissions from distant stars. Solar flares, intense bursts of radiation from the Sun's surface due to the sudden release of magnetic energy, produce streams of highly energetic particles, and are often accompanied by coronal mass ejections (CMEs) of gas and magnetized plasma. When these hit the Earth's magnetosphere this can induce electrical currents in conductive material, causing significant damage to our global infrastructure.[20] The most extreme solar flare recorded was the 'Carrington event' in 1859[21], which caused telegraph systems across Europe and North America to fail. Gamma-ray bursts (GRBs) are jets of gamma-rays that are emitted by exploding massive stars or by the collisions of black holes or neutron stars. The threat from GRBs and Supernovae comes from their interaction with the global atmosphere, which could add to the number density of NO_2 (increasing the atmosphere's reflectiveness) whilst decreasing the number density of O_3 (reducing the atmosphere's ability to absorb harmful solar UV radiation). This would produce conditions similar to those of a nuclear winter.

58. Source: Riley, P. (2012). On the probability of occurrence of extreme space weather events. Space Weather, 10(2). https://doi.org/10.1029/2011SW000734

    Probability: The probability of an extreme solar flare event, of the same or greater magnitude as the 'Carrington event' of 1859, impacting Earth within the next decade is ~12%.

    Methodology: Historical data are used to calculate the power law relation between the size and occurrence of solar flares. The 'Disturbance-storm time' (Dst) index captures the main phases of a geomagnetic storm. For a holistic probability estimate, the variations in the Dst index over the past 45 years, the distribution of Coronal Mass Ejection speeds from 1996-2010, and the nitrate spikes in ice cores at the poles were all considered to give the final probability.

59. Source: 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

    Probability: "The most likely Poisson occurrence probability for another Carrington event in the next 10 years is 0.063"

    Methodology: To address the fact that his occurrence probabilities rely on observations of a small number of events, Love assumes that the time occurrence of events can be described by an idealised Poisson model. Both frequentist and Bayesian inference methods are used to obtain analytical estimates of long-term occurrence rates, occurrence probabilities, and associated confidence and credibility intervals for geomagnetic storms. The 68.3% confidence interval is 0.016 - 0.137; Love notes that the modal result of 0.063 is half that of Riley's (2012) prediction thus serving as an "example of the limited accuracy of statistical estimates".

60. Source: Homeier, N. et al. (2013). *Solar Storm Risk to the North American Electric Grid.* Lloyd's

    Probability: "The return period for a Carrington-level event is 150 years, with a reasonable range of 100 - 250 years"

19 Stated sources include: Altmann, J., & Gubrud, M.A. (2004).: Military, arms control, and security aspects of nanotechnology. Discovering the Nanoscale: 269 http://nanoinvesting.myplace.us/altmann-gubrud.pdf; Berube, D. M., et al. (2010). Communicating Risk in the 21st Century: The case of nanotechnology. *National Nanotechnology Coordination Office, Arlington* ; Turchin, A. (2008). Structure of the global catastrophe. Risks of human extinction in the XXI century. Lulu.com, 2008 and Williams, R. A., et al.(2010). Risk characterization for nanotechnology. *Risk Analysis* 30.11: 1671-1679. These authors argue that there is currently insufficient information to create a risk analysis for nanotechnology.

20 Bell, J. T., Gussenhoven, M. S., & Mullen, E. G. (1997). Super storms. Journal of Geophysical Research: Space Physics, 102(A7), 14189–14198.

21 Muller, C. (2014). *The Carrington Solar Flares of 1859: Consequences on Life*. Origins of Life and Evolution of Biospheres, 44(3), 185–195.

Methodology: The authors aggregated historical recordings of intense geomagnetic storms. There were 2 Carrington-scale storms over the period 1137-1648 A.D. in East Asia and 5 credible such sightings from Yemen, Iraq, Egypt, Syria, and Morocco over the period 817-1570 A.D. This sets a lower limit on the recurrence interval of Carrington-scale events at ~150 years.

61. Source: 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

    Probability: We "expect a dangerous [Gamma-ray burst] event rate of $6x10^{-9}yr^{-1}$".

    Methodology: The authors estimate that a typical GRB pointed at the Earth from 3 kiloparsecs or closer would constitute a serious threat to the biosphere by considering the relevant atmospheric physics of gamma ray photons dissociating ozone to a level which would render the planet uninhabitable. The observed GRB rate at low redshift (nearby in the universe) is $0.44Gpc^{-3}yr^{-1}$ and the number density of typical galaxies is $3x10^{-3}Mpc^{-3}$. This gives an event rate of $0.44x10^{-9}$ $3x10^{-3}$ = $1.5x10^{-7}yr^{-1}$ in our galaxy. Since 3 kiloparsecs is about one-fifth of the radius of our galaxy, and considering the Milky Way to be a 2D disk, the dangerous GRB event rate is $1.5x10^{-7}$X (1/5)2 = 6x10-9 yr-1.

62. Source: 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.

    Probability: The "rate of core-collapse supernovae occurring within 8 parsecs is 1.5 Gyr^{-1}."

    Methodology: The authors considered detailed atmospheric models to determine the extent of the reduction in ozone due to elevated levels of nitrogen induced by gamma-rays and cosmic rays produced in a local supernova. The authors found that a core-collapse supernova would need to be situated approximately 8 parsecs away to produce a combined ozone depletion from both gamma-rays and cosmic rays of 47%. This would roughly double the globally-averaged, biologically active UV reaching the ground, which would significantly endanger life on the planet. By considering the kinematics of the Sun in the galactic plane and supernova data, the rate of core-collapse supernovae occurring within 8 parsecs is 1.5 Gyr^{-1}.

Particle Physics Experiments:

Particle Physicists aim to discover the fundamental forces and elementary particles that make up nature. This can involve building bigger colliders to create ever-greater centre-of mass energies to search for new exotic phenomena. Because of the high energy densities reached in such colliders, concerns have been raised about their safety. There are three main causes of speculative concern[22] Firstly, microscopic Black holes, but the potential for these to produce any adverse consequences has been excluded by robust empirical evidence[23]. Secondly, a vacuum transition. Our universe's vacuum state might be metastable and so, via stimulation through a high energy event, it could decay to a lower energy, true vacuum.Thisnew vacuum's expansion would propagate outwards at the speed of light rendering our future light cone uninhabitable. Finally, Strangelets are hypothetical particles containing equal numbers of up, down and strange quarks that are more stable than ordinary nuclei. They could catalyse the conversion of the Earth into 'strange matter', which would inevitably be our demise..

63. Source: Dar, A., De Rújula, A., & Heinz, 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*

    Probability: The probability of a fast catastrophic destruction event due to collisions at the RHIC each year is less than **$2x10^{-9}$** and the probability for a 'slow destruction', that would be completed in the billion years before the Sun expands beyond Earth's orbit, is less than **$2x10^{-7}$**.

22 Hut, P. & Rees, M. J. (1983). *Nature* 302, 508-509

23 Ellis, J., Giudice, G. et al. (2008). Review of the safety of LHC collisions. *Journal of Physics G: Nuclear and Particle Physics, 35(11)*

Methodology: The probabilities are empirically derived from the observation that collisions between very high-energy particles occur in nature due to cosmic rays. Since strangelets would be slowed by galactic magnetic fields they would be absorbed during star formation. If they were dangerous, then the star would be unstable and initiate a supernova. The authors combined numbers for the production rate of strangelets in free space due to cosmic ray collisions with the observed supernova rate to estimate the potential dangers of stranglet formation.

64. Source: Jaffe, R. L., Busza, W., Wilczek, F., & Sandweiss, J. (2000). Review of speculative "disaster scenarios" at RHIC. Reviews of Modern Physics, 72(4), 1125.

Probability: The probability of vacuum transition yearly at RHIC as a result of iron-iron collisions is less than $\mathbf{2x10^{-37}}$ and as a result of gold-gold collisions is less than $\mathbf{2x10^{-27}}$. There are three estimates for the upper bound probability of strangelet production yearly at RHIC. These are $\mathbf{2x10^{-11}}$, $\mathbf{2x10^{-6}}$ or $\mathbf{2x10^{-5}}$.

Methodology: There have been 10^{47} iron-iron collisions due to cosmic rays in our past light cone and there will be $2x10^{11}$ iron-iron collisions over the lifetime of the RHIC – 10 years - giving $2x10^{-37}$ as the upper bound vacuum transition probability per year. A similar calculation is used to estimate the probability of a gold-gold vacuum transition. The empirical upper bound probabilities for the production of strangelets are derived by considering the heavy ion collisions due to cosmic rays hitting the moon's surface and recognising that the moon has survived for 4.5 billion years (the earth could not be used due to the fact that this would have left an 'anthropic shadow'). The 3 different probabilities assume different hypothetical strangelet production mechanisms. The authors point out that these probabilities are based on a total "worst case analysis" and that there are many theoretical reasons why the existence of dangerous strangelets is thought to break the laws of physics.

65. Source: Tegmark, M., & Bostrom, N. (2005). Is a doomsday catastrophe likely? *Nature, 438(7069), 754-754. https://doi.org/10.1038/438754a*

Probability: "The risk from present-day particle colliders is reassuringly small: say, less than $\mathbf{10^{-12}}$ per year."

Methodology: The authors account for observation selection effects, whereby observers are precluded from noting anything other than that their own species has survived up to the point when the observation is made. An upper bound on the cosmic catastrophe frequency free from such observation-selection bias, 1 per 1.1 Gyr at 99.9% confidence, is derived from planet-formation rates, the distribution of intelligent species birth dates, and our own temporal location. The authors assume from Jaffe, R. L., W. Busza, et al. (2000) that the rate of possible disasters is 10^3 less likely in particle colliders than nature. Therefore, the yearly risk of catastrophe is $10^{-3}x10^{-9} = 10^{-12}$.

66. Source: 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*

Probability: The probability of a catastrophic event at the LHC each year is less than $\mathbf{1x10^{-32}}$.

Methodology: The paper focuses on the Large Hadron Collider's (LHC) run 1 operation with a beam energy of 7 TeV. 10^{16} proton-proton collisions occured at this energy yearly in the experiment. An equivalent centre-of-mass collision energy is reached when a cosmic-ray proton with an energy of at least 10^8 GeV hits a fixed target like the Earth or the Sun. The flux of such cosmic-ray protons incident on the Earth or the Sun is $5x10^{-14}s^{-1}cm^{-2}$. The authors use the known surface areas of stars, there being 10^{11} stars in our galaxy, and there being 10^{11} galaxies in the observable Universe, to determine that nature has completed the equivalent of 10^{32} LHC yearly proton-proton collisions with no indication of "large-scale consequences". The resultant probability (less than $1x10^{-32}$) applies for "vacuum bubbles, magnetic monopoles [or] microscopic black holes" and the paper notes that "strangelet production at the LHC is less likely than at previous lower energy machines" such as the RHIC.

Unknown Consequences

The following estimate concerns risks associated with scenarios that at present seem either very unlikely or very unlikely to pose a significant risk, but where there is a possibility that these assessments represent a significant underestimate of the threat. This includes physics experiments, as described in the previous section, but also risks such as animal cognitive enhancement and the search for extraterrestrial intelligent life. Pamlin and Armstrong also discuss the existential threat posed by Global System Collapse and Future Bad Global Governance, but believe that the probability of even reaching infinite threshold from such risks cannot be quantified.

67. Source: Pamlin, D. & Armstrong, S. (2015). Global Challenges: 12 Risks that Threaten Human Civilisation, Global Challenges Foundation

    Probability: "Based on available assessments the best current estimate of an uncertain risk in the next 100 years is: **5%** for infinite threshold [and] **0.1%** for infinite impact" (p. 160).[24]

    Methodology: See source 17, under Nuclear War section above.

24 Stated sources include: Ellis, J., et al. (2008).: Review of the safety of LHC collisions. Journal of Physics G: Nuclear and Particle Physics 35.11): 115004; S Korean dies after games session, BBC News; Alex Knapp, Is It Ethical to Make Animals As Smart As People? Forbes and George Dvorsky, New Project to Message Aliens is Both Useless and Potentially Reckless, io9.

Appendix B - Our evaluation of methodologies

Our approach to the evaluation of methodologies centred around the definition of four evaluative categories and key criteria for assessing each of these. These are set out below:

Rigour

Ability to access a broad range of information and expertise from across multiple perspectives:

1 = only accepts info within strict parameters,
2 = information from only one discipline,
3 = information would require translation across disciplines,
4 = information flows freely.

Suitability of this method's means for turning this information into a final judgement:

1 = arbitrary and/or opaque,
2 = inflexible and/or unclear,
3 = flexible and transparent,
4 = insight generating.

Ease, or difficulty, of incorporating new information into this judgement using the same method, or of combining different judgements together:

1 = impossible,
2 = requires redoing entire exercise,
3 = requires substantial work,
4 = easy.

Uncertainty

To what extent does this approach provide opportunities to quantify the level of confidence, or uncertainty, in its estimates:

1 = no,
2 = with subjective judgement,
3 = with standard statistical tools,
4 = providing new insights.

Does the application of this method tend either to systematically ignore or compound sources of uncertainty in the process of forming a final judgement:

1 = yes,
2 = likely unless extreme care is taken,
3 = easily avoided,
4 = makes different sources of uncertainty clearer.

To what extent is this method vulnerable to biases:

1 = very prone to individual biases,
2 = provides opportunities for overcoming biases,
3 = helps identify and address biases in a proactive way,
4 = actively addresses individual biases.

Accessibility

Total amount of time required to implement this approach in a reasonable way (leaving aside all background reading a research):

1 = > 1 researcher year,
2 = <1 researcher year,
3 = <1 researcher month,
4 = <1 researcher week.

How hard would it be for a researcher to take a lead or principal role in implementing this method in a reasonable way.

1 = requires membership of a particular organization (such as a project or research group),
2 = requires the skills and expertise of a domain specialist (PhD equivalent) to implement,
3 = requires a non-specialist receive training to implement in a reasonable way,
4 = could be implemented by a non-specialist with no further training

What other barriers exist to implementing this method in a reasonable way. Score 1 point for any of the following:

a) requires access to a community of experts to provide advice or respond to questions,
b) requires significant computing power,
c) requires interdisciplinary or intercultural working.

Utility

Credibility, both with scientists and non-scientists e.g. due to previous track record:

1 = controversial or widely distrusted,
2 = generally held in a neutral regard,
3 = generally held in a positive regard,
4 = widely trusted and esteemed.

Ability to provide useful quantitative information:

1 = information is technical and hard to understand,
2 = information requires translation for policy use,
3 = information easy to understand,
4 = information tailored for policy / media use

Ability to provide further information and insights about the risk under study:

1 = only usable for probability estimates,
2 = reveals strength of different drivers of probability,
3 = reveals opportunities for risk management,
4 = makes drivers and management opportunities immediately clear and compelling.

For each of these criteria each method was scored (1 to 4) by the three authors of this paper, representing a cross disciplinary group of scholars. One works as an academic programme manager at an Existential Risk research organization, one works as an academic philosopher and the third has recently completed a degree in physics and will shortly be starting a PhD in Machine Learning. The scores for all three criteria of each category were summed to produce an overall score for that category and then the mean across the authors scores was taken as the final category score for that method. This was then interpreted using the following classification scheme for each of the categories
3 - 5.5 = Very Low
5.5 - 7,5 = Low
7.5 - 9.5 = Medium
9.5 - 12 = High

Where there was no more than a 1 point difference between any of the author's and the mean then this classification was reported as it is. Where there was a greater difference then either a split classification (such as Low/Medium or Medium/High) was supplied if this difference bridged two classes or an asterix was added to the final classification to denote a lower degree of certainty about how this method should be evaluated.

The scores of the three authors for each of the four categories and the mean scores of the three criteria are presented in the following two tables, while the final classification of each method is presented in section 6 of the main table.

Scores by author	Analytical approaches	Extrapolation and Toy Modelling	Fault Trees	Bayesian Networks	Adapting large-scale models	Individual Subjective Opinion	A O
Rigour	4.3	6.7	9	9.7	9.7	6.7	8
Simon Beard	4	7	9	10	9	8	9
Tom Rowe	5	6	10	11	10	6	8
James Fox	4	7	8	8	10	6	8

Uncertaity	5	7	6.7	9.7	10.7	4.3	6
Simon Beard	6	8	6	10	10	5	6
Tom Rowe	4	6	7	9	11	4	5
James Fox	5	7	7	10	11	4	8
Accessibility	11.3	10.3	8.3	7	3	11.7	9
Simon Beard	12	11	10	8	3	12	7
Tom Rowe	11	10	7	7	3	11	1
James Fox	11	10	8	6	3	12	1
Utility	3.7	7	8.3	9	10.7	7	6
Simon Beard	4	7	8	10	10	7	6
Tom Rowe	4	7	10	10	11	6	7
James Fox	3	7	7	7	11	8	6

Scores by criteria	Analytical approaches	Extrapolation and Toy Modelling	Fault Trees	Bayesian Networks	Adapting large-scale models	Individual Subjective Opinion
Rigour	4.3	6.7	9	9.7	9.7	6.7
Ability to access a broad range of information and expertise from across multiple perspectives	3	5	8	9	8	8
Suitability of means for turning this information into a final judgement	5	8	10	10	12	4
Ease of incorporating new information into this judgement using the same method or combining different judgements together	5	7	9	10	9	8
Uncertainty	5	7	6.7	9.7	10.7	4.3
Provision of opportunities to quantify the level of confidence, or uncertainty, in estimates	7	9	8	12	12	6
Tendency to systematically ignore or compound sources of uncertainty in process of forming a final judgement	3	7	6	10	11	4
vulnerability to biases	5	5	6	7	9	3
Accessibility	11.3	10.3	8.3	7	3	11.7
Amount of time required to implement in a reasonable way	12	10	8	7	3	12
Difficulties for a researcher to take a lead or principal role in implementing in a reasonable way	10	9	7	6	3	11
Other barriers exist to implementation	12	12	10	8	3	12
Utility	3.7	7	8.3	9	10.7	7
Credibility, both with scientists and non-scientists	3	8	8	9	12	6
Ability to provide useful quantitative information:	5	8	7	7	9	11
Ability to provide further information and insights about risk	3	5	10	11	11	4