Thomas Rowe and Simon Beard

Probabilities, methodologies and the evidence base in existential risk assessments

Working paper

Probabilities, Methodologies and the Evidence Base in Existential Risk Assessments

CSER Working Paper

Thomas Rowe[1] and Simon Beard[2]

This paper examines and evaluates a range of methodologies that have been proposed for making useful claims about the probability of phenomena that would contribute to existential risk. Section One provides a brief discussion of the nature of such claims, the contexts in which they tend to be made and the kinds of probability that they can contain. Section Two provides an overview of the methodologies that have been developed to arrive at these probabilities and assesses their advantages and disadvantages. Section Three contains four suggestions to improve best practice in existential risk assessment. These suggestions centre on the types of probabilities used in risk assessment, the role of methodology rankings including the ranking of probabilistic information, the extended use of expert elicitation, and the use of confidence measures to better communicate uncertainty in probability assessments. Finally, Section Four 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.

Introduction

"Theories that involve the end of the world are not amenable to experimental verification – or at least, not more than once" Carl Sagan[3]

It is very hard to make empirical claims about unprecedented events such as global catastrophes and existential risk. On the one hand, as Carl Sagan notes in the above quote, any predictions we make cannot be empirically verified, either in terms of the likelihood of an event that might precipitate some global catastrophe or in terms of how humanity will respond to it. On the other hand, the question of what sorts of existential and catastrophic risks we might, as a species, be facing is pressingly urgent, and being able to decide between competing priorities, even relating to what kinds of risks are most likely and most dangerous, is vital if we are to make any progress towards building a safer world.

Despite this difficulty, however, it seems clear to many working in the field of existential risk mitigation that priorities can be determined based on scientific facts and rational deliberation. In order to do this people have had to develop a variety of new methodologies

[1] Department of Philosophy (Virginia Tech) and Centre for the Study of Existential Risk (University of Cambridge)

[2] Centre for the Study of Existential Risk (University of Cambridge), Centre for Philosophy of the Natural and Social Sciences (London School of Economics) and Institute for Futures Studies

[3] Television Appearance. "The world After Nuclear War". October 31st 1983, Reported in "The Atomic Origins of Climate Science". The New Yorker. January 30th 2017.

for predicting the unpredictable and empirically studying the unverifiable. Such methodologies however tend to be used and developed in a piecemeal fashion and often remain highly specific to one particular domain or discipline. For instance, conservation biologists have made important innovations in the use of aggregated expert opinions for the purpose of horizon scanning for novel threats and opportunities. Analysts in the Intelligence Community have developed super-forecasting techniques to improve individual reasoning and estimation of probabilities, even when available information is very low. Physicists, economists and demographers have developed increasingly detailed models for understanding complex processes and exploring the likely impact of unprecedented events. Computer Scientists have worked out techniques to quantify probability estimates from large and varied groups of stakeholders in order to pool collective knowledge about the likelihood and probable outcome of long term expected trends in the development of new technologies. These emerging methodologies each come with their own limitations and benefits and it is time that they be set free from their individual domains. The aim of this article is to introduce different techniques for those who may not be familiar with each of them and to give an objective assessment of their capabilities, together with some recommendations for how they can be improved and implemented.

1. What types of claim? What types of context?

Before proceeding, it is worth noting the various purposes that probability estimates may serve. Three such ways in which probabilities can be used are for predicting catastrophic events, for prioritizing interventions and for coordinating efforts. The first of these aims at an accurate assessment of the probability of a particular risk in order to assist mitigation strategies, for instance by determining the expected costs and expected benefits of different interventions to mitigate it. The second does not require such accurate estimates, but merely calls for an ordinal assessment of the relative likelihood of a particular event or risk in order to help prioritize particular strategies. For instance, it might be sufficient for policy makers to know that one particular catastrophe is more probable than another, without aiming for an accurate characterization of how probable either risk is, whereas, it might be necessary to provide such a characterization for insurance purposes. The final use of probability claims is for coordination between different groups, by fostering provisional agreement, setting a placeholder value to allow for work on other aspects of existential risk mitigation to proceed more smoothly or for raising awareness amongst groups who demand quantitative estimates of probability whether they are appropriate or not. This often calls for even less accuracy in its estimates, merely requiring that a claim be the subject of a sufficient level of agreement to allow for its use in the coordination of diverse actors. For instance, if everyone can agree on the upper or lower bound probability for a particular risk, even if such bounds are quite extreme, this may well be sufficient to allow for a general agreement about its importance or significance without any need to establish its actual probability.

As well as being used for different purposes, claims about the probability of risks can be used “in context”, where they directly assist in assigning resources to mitigate risk, or “out of context”, where, for instance, they are used to set a time discount rate, raise awareness of the importance of an issue, justify further research or simply draw attention to an area of tractable uncertainty. It can be problematic to assume that all estimates related in some way to existential risk are being used “in context”. This is because different uses of claims demand different levels of accuracy and precisions, and claims that have been developed for use out of context are easy to misapply without reassessing their suitability for this new use. These claims may, therefore, be justified on grounds that do not strictly relate to existential risk itself, but to something else such as uncertainty, time discounting, discipline building or the development of public discourse and activism on global challenges.

1.1 What Type of Probabilistic statements should we make, If Any?

Having briefly considered the purposes of probability assessments, we will now examine different *types* of probabilistic statement that we might make in such claims.

Firstly, we often make *quantified* probability estimates, though they are only strictly required for the purposes of prediction and may not even be appropriate if one is only interested in using such estimates for prioritization or coordination. An example of a quantified probability estimate is the following: ‘there is a **1×10^{-7}** annual chance of an extinction-level asteroid impact’[4]. This preference for quantified estimates is given in, for example, the surveys of Artificial Intelligence expert opinion and estimates of the probability of nuclear war. Quantified probabilities are necessary for cost-benefit analysis, but are often unrealistic in the case of existential and global catastrophic risk, given the severe uncertainties involved in risk assessment and the possibility of error. Whilst it is, of course, possible to quantify uncertainty and this can be a very useful thing to do, quantification can easily lead to overconfidence and this has led some organizations that work in areas of high uncertainty, such as GiveWell, to prefer less quantified approaches to assessment.

Secondly, one can provide semi-quantified probability estimates that provide some degree of numerical precision about the nature of a particular risk, without implying the degree of precision associated with quantified probability estimates. A common example of such claims are probability intervals that give a range of potential probabilities within a specified degree of confidence or which state that the probability is most likely to fall within a particular range or to be above or below a particular threshold. An example of the first kind of semi-quantified probability is the following: ‘with 90% confidence that the annual probability of accidental nuclear war between the US and Russia is from 0.001% to 7%’.[5] An example of the second kind of semi-quantified probability is the IPCC’s claim that it is ‘very unlikely’ that the Climate Sensitivity to Carbon is greater than 6 degrees (implying that

[4] Derived from Chapman 2004, 11 - see Appendix

[5] Derived from Barrett et al. 2012, 118 - see Appendix

climate change is more likely to pose an existential threat), where this is calibrated to mean that the probability of this most likely falls into the range 0-10% probability. As can be seen, semi-quantified claims can be expressed either numerically or with calibrated non-numerical language (so called 'Words of Estimative Probability, such as likely, very unlikely and so on). Whilst quantification is generally preferred because it generally communicates more information and is easier to standardize the use of calibrated qualitative expressions can be valuable in that these can better communicate the degree of uncertainty in a claim without needing to introduce significant error bars, which can be off putting in a policy context. However, they are expressed, semi-quantified probability estimates are often heavily misreferred by certain communities, including many policy makers and the media who will tend to convert them into quantified estimates, for instance by taking the mean or median value or one of the bounds of the range being expressed as an actual prediction about the probability under consideration. One reason for doing this is that it is difficult to determine expected costs or benefits from probabilities that are expressed only imprecisely. Similarly, semi-quantified probabilities may lead people to believe that the assessor has insufficient knowledge about the risk if the interval is wide, even if these estimates actually reflects a wealth of available evidence, as for instance is the case with the IPCCs assessments of the risks associated with Climate Change. For example, when a probability range is very large, such as Barrett et al's estimate for the likelihood of a nuclear war between the USA and Russia, this may lead people to feel that since so little is known about the risk it can be ignored for policy purposes when deciding how to allocate resources. However, in truth the breadth of this range, and the fact that it's upper bound is so high, should instead be seen as a reason in itself for paying much more attention to the risk, relative to other risks that we know to be a lot lower (such as risks from conventional terrorism or public safety). In the event that the upper bound remains stubbornly high even with the addition of more evidence, such as with the possibility that the Climate Sensitivity to Carbon will exceed 6 degrees despite a century of research seeking to reduce our degree of uncertainty about it, this should be enough to suggest that this is a risk that should be taken very seriously indeed.

Thirdly, there is the option of purely qualitative assessments of risks. For instance, in an environmental scanning approach an assessor builds a simple multiple component causal model of human extinction and then scans human knowledge for "leads" that influence these components (Tonn and Stiefel, 2013: 1780).[6] This approach can be fruitful if one thinks that existential risks are the sort of thing that it is inappropriate to assign a quantified probability to, given the unique and unprecedented nature of such events. This is because it can still help to map out the contributory causes, and how events relate to one another. It can also be useful for policy-makers who are interested in how particular areas of interest will be affected by particular events. However, the lack of quantification has definite

[6] Tonn and Stiefel conduct an environmental scan of the various factors that contribute to the threatening of various species extinction in "The Race for Evolutionary Success", *Sustainability*, Vol. 4, 2012, pp. 1787-1805.

drawbacks, given that we are not only interested in the relationship between events, but also in using this information to make accurate assessments, for instance of their expected impact or relative importance. One might think that any quantitative probability judgment is better than none, since it gives us additional useable information about that event, and we can always seek to improve upon this at a later stage. However, this additional information does not come without risks, as an erroneous or overconfident estimate of probabilities may lead to a worse outcome than no assessment at all, especially where this simply reflects existing biases or may have wider systemic effects[7].

Another qualitative approach is to simply rank existential/catastrophic events by their anticipated worst outcome. Efforts to mitigate potential catastrophic outbreaks of a deadly disease may be ranked more important than efforts made to avert a catastrophic asteroid. Here the badness of outcomes solely determines the badness of the event, rather than being discounted by the improbability of its occurrence. This is an ordinal ranking of catastrophes that only takes into account the relative magnitude of outcomes. This view can help to prioritize the allocation of resources, but the obvious problem is that it is insensitive to the likelihood of events where this may make a big difference to whether one would intervene to mitigate it or not. In cases where there are multiple catastrophic risks, such as near-extinction from artificial intelligence, or from a natural pathogen with pandemic potential, one may think of the former as a worse state of affairs than the latter and wish to direct more resources to mitigating it. However, this view has problems with extinction events, since presumably any extinction event of the human species is equally bad.

As well as considering what kinds of statements we might make about the probability of certain risks or events we should also briefly discuss disagreements about the actual nature of probability, although this is an issue we will pay little attention to in this paper. Broadly, and very roughly, speaking there are two competing conceptions of probability in contemporary epistemology, statistics and decision making. The first of these is the frequentist, or objective, notion of probability. According to this approach probabilities represent objective facts about the frequency of certain events based on past observations. Probabilities can only be determined within precise parameters and cannot be applied outside of these parameters. However, new evidence cannot be used to simply 'update' our existing probabilities but require us to reassess probabilities from scratch. The second notion of probability is the Bayesian, or subjectivist, notion of probability. According to this approach, probabilities represent subjective judgements about the likelihood of some event conditional upon other information that we know about the world. Probabilities can only be determined by updating our previous beliefs in light of new evidence via Bayes' Theorem (or the Bays Rule) which allows us to determine how additional information affects the conditional probability of an event. However, in the absence of any existing information

[7] See also, Yudkowsky, E., 2015, "When (not) to use probabilities. In *Rationality from AI to Zombies*. Machine Intelligence Research Institute", pp. 1499-1505

about the probability of an event we have nothing to update, so must start with a probability derived via some other means.

Since frequentist notions of probability are only of limited use in making predictions and cannot be applied to cases in which there is not the right kind of evidentiary record, Bayesian probability dominates in the field of Existential Risk assessment and almost all scholars working in this field accept the validity of this approach to probability assessments. However, there are still fields that existential risk scholarship must interact with, most notably in the domain of public health, where Bayesian probabilities are less widely accepted and some policy makers continue to look down on this interpretation of probability, or at least to prefer frequentist notions of probability. Where available we shall therefore discuss the two notions of probability interchangeably with only a few passing remarks, and will instead focus on the quality of the methodologies that can be used to produce both kinds of probability in this paper.

1.2 Defining existential and catastrophic risk as our locus of concern

Whilst there is much common ground between scholars about what constitutes catastrophic and existential risks, these concepts are not universally and precisely defined. Clarifying the definition of existential and catastrophic risk is undoubtedly a crucial step in studying and mitigating them properly. However, since the aim of this paper is to understand the methodology for studying such risks, and since methodologies are at times being used to study differently conceptualized risks we shall, in this paper, characterize these risks roughly as those associated with human extinction, civilizational collapse and other catastrophes of an equally significant impact.[8]

In the next section, we evaluate a number of the core methodologies used for assessing catastrophic and existential risks.

2. Extrapolation from Data Sets

A common methodology for establishing probabilities of existential risk is through extrapolation from data sets. Data is collected from historical instances of a particular risk, and the assessor then uses this to establish a probability of a future such instance of the risk. This methodology is used frequently, either as the sole method for establishing the risk (such Torres 2016, who uses frequentist methods to establish the probability of super-volcanic eruptions and asteroid impacts) or as an ingredient in an overall probability

[8] It should also be noted that any assessment of the probability of such events cannot be given independently of human actions, and hence indeed of our assessment of their probability (if we overestimate the likelihood of a catastrophe this is likely to lead to greater efforts to mitigate it, thus rendering it less likely and so on). This is not strictly an issue for this paper, which is based upon assessing existing methodologies that all must handle this issue. Unless started explicitly we simply assume that the claims we are examining all assume a roughly 'business as usual' account of future activities where people will be roughly as committed to the mitigation of a particular risk as they were at the time when the probability of that risk was assessed.

assessment (such as Hellman 2008 who extrapolates from previous data to determine the underlying frequency of Cuban Missile Type Events or Lipstick and Inglesby 2015 who use a frequentist analysis of previous laboratory acquired infections in their risk assessment of Gain of Function Research on Pathogens with Pandemic Potential - see appendix for more details).

Some immediate benefits of this approach are as follows. On a frequentist interpretation of probability, this method allows for the establishment of objective probabilities.[9] For example an annualized probability of a catastrophic super-volcanic eruption can be gleaned from past instances of such eruptions (Torres, 2016). This gives us an objective, if not necessarily precise, probability of a specific event, which is something that is useful both for the public communication of a particular risk and also for use in public policy. Such probabilities make it easier to calculate the upper and lower bounds for the expected costs and benefits of intervening to lower the severity of a risk. A further benefit is that well established historical data can feature as one part of the evidence base in the assessment of probability. It is also possible to incorporate these objective probability assessments into more sophisticated subjective assessments. For example, Hellman (2008) uses historical frequencies of nuclear near-misses as a stepping stone to the arrival at a probability of a Cuban missile-type crisis. A difficulty with this approach, however, is that since there has not been a human extinction event, we are unable to establish a relative frequency of extinction. Human extinction events are, by their nature, unprecedented, and so even their relationship with other kinds of historical events, such as natural extinctions, super volcanic eruptions and nuclear near misses cannot be precisely determined and must be estimated with some degree of subjectivity. Further, the frequencies of catastrophes that lead to the non-existence of observers will be underestimated due to the anthropic shadow such events would leave. This is because the presence of any event that would have caused humanity to go extinct, or never to have come into existence, in our historical timeline would preclude the existence of present human observers who could then include such an event in their assessment of the probability of such events occurring. As such, there is an anthropic bias for the estimation of catastrophic risks (Cirkovic, Sanbderg, and Bostrom, 2010).

There are further issues with this methodology. First, the relative frequencies of historic catastrophic events may not be continuous. For example, there is no reason to think that there is an underlying mechanism such that an eruption of a super-volcano occurs like clockwork.[10] Fouchier (2015), for example, questions the use of historical data on pathogen

[9] A frequentist interpretation of probability states that the probability of an event is the limit of the event's frequency in a repeated number of trials.

[10] Gordon Woo (forthcoming: 14) considers this point with the example of a sequence of six earthquakes at Parkfield, California between 1857 and 1966. An earthquake occurred at approximately regular time intervals of 22 years. On these grounds it was predicted that the next earthquake would occur in 1988. In fact, the next earthquake occurred in 2004. Woo argues that the 1966 event was not inevitable in that year, and the same is true for the other events. They could have occurred on different years. If the randomness of each event had

outbreaks from laboratories (Lipsitch & Inglesby, 2014) as they fail to take into account contemporary safety standards. This limits the scope of this approach, as it can only be used when the conditions under which past data points have occurred still holds and thus may fail to give apply to future catastrophic events that take place under different conditions. This leads to a second issue. How the data set is carved up or selectively used can create probabilities of differing magnitudes. For instance, how the frequency of laboratory-acquired infections are recorded (whether infections per laboratory-year or worker-year) makes a difference to the calculated probability (Lipsitch & Inglesby). There may also be other reasons to "tailor" the frequency, by omitting particular cases in order to avoid "being alarmist" (Hellman, 2008) by creating a probability that appears too large. This points to the use of probabilities for goals other than prediction, and in particular for this use in coordination where causing alarm may be counterproductive, even if it merely resulted from trying to accurately representation truth as one finds it.

A related issue is that there are often many "near misses" of a particular risk. Do we include such cases in the frequency? Gordon Woo has argued that we should sometimes incorporate a "counterfactual analysis" of near miss events in order to more accurately model risks (forthcoming), because of the likelihood that very rare events will be underrepresented in any historical data sample. For instance, if the underlying probability of a catastrophic event occurring were 0.004 per year and we had 100 years of data that were relevant to the event occurring, then it is most probable that that event would not occur within this time period, leading us to underestimate it's probability. Woo cites numerous near-misses in air and sea travel, where the inclusion of such near misses would have given a more realistic depiction of the risks (Ibid: 15). In the nuclear case, there are a number of near misses that *could* be included in a category of "almost Cuban Missile Type crises", but if such cases are included then the probability of such events occurring *increases*. An issue with this approach is that the objective determination of a near miss is difficult. Is it a near miss when an airplane misses a building by 1 meter or by 2 meters, etc.? However, if near misses are excluded, this leads to the underestimation of risks. A dilemma ensues, whereby one either underestimates risks by excluding near misses, or incorporates too much subjectivity with respect to the selected relevant counterfactuals.

Although the approach is useful for arriving at relative frequencies of catastrophic events, it therefore has limited appeal for estimating probabilities of existential risks. However, the approach can still play a useful role, for instance by giving a foundational probability for risks that can be updated with new information via so-called 'objective-Bayesian' approaches.

3. Modelling

been incorporated in the risk assessment, seismologists would have been less confident about the probability of the next earthquake.

Another methodology that is sometimes used for making claims about the probability of existential and catastrophic risks is systems modelling. Models are useful for studying how complex systems might respond to dramatic external shocks, from mega disasters to global wars, and can also be used to identify what it would take to push a robust system, such as a nation or region's energy or food supply, into a failure state from which it could not easily recover.

The simplest form of model one might use in studying existential risks is the "influence diagram". These represent systems by breaking them down into component parts and the influence that they have on one another. Components may have a unidirectional influence (one component can affect the other, but not vice versa) a bidirectional influence (either component can affect the other) or no influence. Components can also be influence by, and can sometimes also influence, elements of 'the environment', things that fall outside of the system being analyzed and whose influences on each other are not assessed as part of the model. One example of an influence model that is highly relevant to the study of existential risk is the MIT Integrated Global Systems Model, see Sokolov et al. 2005. Such models allow for scientists to study complex systems and make Judgements about the them, and their potential failure states, such as whether the system contains positive or negative feedback loops, its degree of sensitivity to influences from the environment and the extent to which different parts of the system function independently or are closely correlated to one another. However, their use in making quantitative predictions is limited by the fact that the nature and strength of the influence components have upon one another is not included in these models.

Another modelling technique that combines influence diagrams with principles from decision theory is the use of fault trees. This is a technique whereby an undesirable event is modelled by linking the possible scenarios that could lead to that event occurring and the influence that decisions will have on these. Each node is a type of event, for example the introduction of some technology, or the implementation of a policy. Whilst the influence these events will have on one another is marked with the use of logic gates that precisely describe the nature of the dependency between them. By assigning probabilities to the nodes based on a variety of different methodologies it is possible to use fault trees to make predictions about the probability, or even the possibility, of events further down the fault tree. This technique has been used in nuclear war analysis (Barrett et. al., 2013 - see appendix), as well as the field of AI safety (Barrett & Baum 2016 - see appendix).

Barrett & Baum (2016: 6) provide one example of how a fault tree analysis can be used for modelling an artificial superintelligence catastrophe. This assess the probability of such a risk by considering the preconditions for a catastrophe to take place, and then considering the preconditions for these preconditions. These then form a set of 6 well defined conditions that must all be satisfied before an AI catastrophe can occure, and whose probability can be determined by other means.

Other models separate out the necessary and sufficient conditions for a fault to occure, defining each of these via a system of AND and OR gates. This helps to pinpoint particular causes, and clusters of causes, of events of special importance. It can then be used to identify and prioritize policy-relevant interventions to prevent or lessen the likelihood of particular events, such as existential catastrophe.

One difficulty with this methodology, however, is that it might not always be possible to comprehensively identify those factors that have an influence on particular events. We might also mistake whether a particular combination of events is necessary (or sufficient) to cause another event, rather than a sub-set of that combination. There is also the potential for a lengthy regress of contributory causes, since one can find causal influences on the events at the base of the diagram, and causal influences on these events. A solution to this problem would be to make a judgement about when salient causes with the potential for policy interventions stops. A benefit of this approach is that it gives a simplifying outline of the causal structure leading up to an existential risk/catastrophe, and is thereby relatively straightforward to communicate with policy-makers.

A final modelling methodology that has been used to study existential risk is the use of Bayesian networks. This also builds upon influence diagrams, but provides a more sophisticated means for assigning probabilities to events. A Bayesian network is a directed graph with a number of values and events. Each node in the tree, or network, is associated with a probability function, rather than simply a logic gate, that determines the probability of an event conditional on prior events occurring. Hence, rather than being assessed merely in terms of direction of influence these diagrams allow for the quantification of probability across the system as a whole. The assessor creates the tree and then assesses the conditional probability of each particular event or node depending upon the state of other nodes in the tree. The Bayesian network algorithm is then used to determine the final expected probability of each node in the network based on the assess prior probability of other nodes, and this can be calculated dynamically and updated to take account of additional information. The concept of using Bayesian Networks to assess catastrophic risks under uncertainty has been shown by Li et. al (2010), who use a Bayesian network to model catastrophic flooding in China, who obtain the prior probability of particular events either by from experts or from other models (e.g. Agent modelling). These authors note that Bayesian networks have been underused in the assessment of catastrophic risks, and, so far, they have not been extended to modelling existential, or even globally catastrophic, risks.

Modeling is a fruitful way of categorizing existential risks and their contribution to extinction events. Problems remain with the specification of events and their relationship to one another, as well as with the selection of probability values for particular events. It is not always made clear how precise probability values are arrived at for particular events. This may not be a large problem if one’s aim is the identification of contributory causes as well as a focus on the prioritization of interventions to mitigate risks. In this latter case it may be most useful to know what the biggest contributory cause is, or the cheapest cause that one can mitigate.

4. Individual Expert Elicitation

As well as using datasets and models it is also possible to derive probability claims about existential risk from subjective expert opinion, an approach to dealing with uncertainty that has a long and successful history of use across a variety of scientific fields (Aspinall 2010, Morgan 2014). Given the limitations of other approaches this has tended to be the most widely utilized technique, in its various forms. However, in many cases expert opinion is used in its simplest form, with individual experts offering a 'best guess' of what they take the probability of various risks to be.

One established technique for arriving at such individual judgements is the "holistic probability assessment". Tonn and Stiefel define this approach as follows: "the individual probability assessor estimates the holistic extinction risk through informed reflection and contemplation" (2013). This can be done literally as a guess, or as a reasoned estimate given the available evidence where this already exists. For example, Martin Rees states that "I think the odds are no better than fifty-fifty that our present civilization on earth will survive until the end of the present century" (2003: 8). Similarly, Millet and Snyder-Beattie state in a discussion of their model for determining the probability of bio-warfare that "for the purposes of this model, we assume that for any global pandemic arising from [laboratory research into potentially pandemic pathogens], each [global pandemic] has only a one in ten thousand chance of causing an existential risk" (2017: 5). However, in this case the context of this judgement is that they are attempting to assess a realistic lower bound for the damage function of a potential pandemic for the purposes of arguing that research into preventing such pandemics through improved biosafety is a neglected area of study, a classic example of an estimate being used for the purposes of coordination rather than prediction.

This approach helps provide a clear and precise figure for the likelihood of existential risks that is easy to use and communicate. However, there are obvious flaws with this approach. There is a high level of subjectivity and it is plausible to think that different experts will arrive at different probability assessments for the same event. Even if individual assessments tend to converge this subjectivity may mask, conscious or unconscious, biases that are shared by different experts and so cannot necessarily be taken to reflect a convergence on the truth. The approach is also less trusted by policy-makers (although this depends on the perceived authority of the individual who makes it). As we have seen, experts are also sometimes make use of the opportunity of making their own subjective judgement to express figures that they believe to be inaccurate, for instance, Hellman (2010) deliberately understates the probability of a nuclear war because he believed that this was likely to increase the number of people who would take this estimate seriously for political purposes. Similarly, Nicolas Stern (2007) produced one of the most widely cited estimates of long term human extinction simply in order to give the highest pure discount rate that he believed might be justified (although it appears that his judgement was strongly influenced by similar statements made by Martin Rees). In this case, neither expert's opinion should be viewed as truth tracking and great care should be taken that their claims are not used out of context.

5. Group Expert Elicitation

Many methods for improving upon the subjective judgements of individual experts involve combining multiple such judgements into a revised estimate. Broadly speaking, there are two reasons to think that combining the subjective opinions of multiple experts will produce better probability estimates than any of them taken individually. The first of these is more purely epistemic and appeals to the Condorcet Jury Theorem. The second is more psychological and appeals to the elimination of bias and error.

5.1 The Condorcet Jury Theorem

The Condorcet Jury Theorem is an epistemic justification for the aggregation of large numbers of individual opinions as a means of determining the truth of some proposition. It is assumed that individuals receive some kind of signal pointing to the truth or falsity of that proposition, and that as a result they are slightly more likely to judge correctly concerning its truth than to judge incorrectly. Then, so long as individuals are judging independently of one another adding more such individuals will increase the probability that the group will be correct, as it will effectively cancel out the noise that leads some individuals to judge incorrectly and thus increase the likelihood that the groups judgements will correctly reflect what the signal indicates. It is therefore conceivable that enough competent experts can be assembled to establish a more robust probability assessments than a single expert. However, this will only be the case if we assume that there is some signal that each individual expert is using to inform their judgement so that they are more likely to judge correctly than incorrectly.

5.2 The elimination of bias and individual error

We can expect that whilst individual judgements will be affected by multiple biases that not only serve to make these judgements less truth tracking, but can also render them actively harmful by systematically ignoring important phenomena or prejudicing them again specific perspectives and points of view. When combined, however, individual biases might cancel each other out.

However, this can be counterproductive. Biases are often shared by multiple experts, and the notion that a belief is widely held can easily create the illusion that it must therefore be free from bias when this may not be so. Indeed one source of bias is the pressure to conform with others, so that even knowing that one is participating in a social process of collective estimation may lead to the reinforcement, rather than the elimination, of biases.

5.3 Group Expert Elicitation via the simple aggregation of subjective opinions

The simplest approach to combining multiple subjective opinions into a more robust estimate is to mathematically *aggregate* these opinions, for instance by identifying the mean, median or modal estimate. Examples of this methodology for calculating existential risk are from Sandberg & Bostrom (2008), Muller & Bostrom (2014) and Katja Grace et. al. (2017). The most prominent area where this methodology has been used is making predictions about the future of Artificial Intelligence. Part of the explanation why this is so is that there is a lack of a historical record on which to establish a relative frequency of AI-related catastrophic events, and a relatively small expert community on which to draw. There is also a great deal of uncertainty regarding future impacts of AI.

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 – well as old-fashioned optimism and pessimism" (pp. 1-2). There are further issues regarding the independence of participants. Since these surveys are often completed at conferences, it is difficult to ensure that individuals will present their own judgment without influence from others, potentially violating the independence condition of the Condorcet Jury Theorem. Some potential ways of rectifying this are having participants participate remotely, such as via an anonymous online platform, although given how close knit many academic and technical communities are this still may not ensure independence of judgements on its own.

Although there are obvious shortcomings with these surveys, they remain prominent and widely used. For example, the Global Challenges Report listed Sandberg & Bostrom's (2008) survey as the sole source for establishing a probability of unknown risks in the future. In areas where we have very little information at all, such surveys can serve as a first step towards achieving the end of predicting and prioritizing the relevant importance of particular risks.

The shortcomings of the preceding simple accounts of group expert elicitation lead us to the next group of methodologies: structured expert elicitation.

6 Structured Expert Elicitation

Structured expert elicitation methods aim to elicit probabilities from experts through a specially designed procedure. A variety of such methods have been developed by scientists and others, although most of these have so far been little used in the literature on existential risk.

6.1 The Classical Approach

One such approach is Roger Cooke's (1992) so-called "classical" approach to expert elicitation. With this methodology, experts' performance on a series of calibration questions creates weights reflecting the accuracy of the experts' predictions. An expert who more often than not gets closer to the truth has a larger weight in the overall aggregation of probability judgments. This approach to aggregating expert opinion has been shown to clearly outperform that of merely taking the median or mean opinion (Colson and Cooke, 2018)

However, this method is not well suited to the needs of existential risk research because it would clearly be difficult to calibrate experts' competency at predicting catastrophic and existential risks. Perhaps, it might be possible to test experts' putative accuracy through their success at predicting nearby future events, such as using prediction markets for calibrating assessors and testing accuracy at prediction. However, although this methodology has been used for risk analysis and forecasting, there has been little if no use of this method for existential risk. This is probably because of the difficulty for calibrating individual's accuracy at making predictions in the short term against their accuracy at, or even shear ability to, make predictions in the long term. There isn't even any clear method for assessing how accuracy in short term prediction making relates to accuracy in long term prediction making, and it is not apparent that there should even be a positive correlation between these two things. It would nevertheless be of interest to see if the use of this methodology could strengthen the probability estimates that reports such as the *Global Challenges Report* produce, especially as some of these do relate to relatively short-term trends, or at least trends that might be tested within the lifetime of those making the predictions.

6.2 The Delphi Method

Another structured expert elicitation method is the Delphi method. This has been developed as a tool for horizon scanning using a panel of experts. There are two or more rounds where the experts are asked to provide their responses to questionnaires. After each round, a facilitator provides an anonymised summary of the results along with the reasons that were provided. Experts can then revise their answers in light of the anonymised answers given by others. Extreme outliers are asked to provide reasons for their positions. Since experts revise their answers in light of the estimates given by others as well as the reasons that were given, the experts will then be able to converge towards the "correct" answer.

This approach has more structure than the classical approach, and has a rigorous method for providing a consensus result. It also helps maintain anonymity, which gives it advantages over the aggregated expert opinion and structured expert elicitation discussed above. Whilst there has been little use of the Delphi method for estimating existential risks, and many regard it as more useful for qualitative research like horizon scanning, it has been applied to other areas of risk analysis. For example, Cousien et. al. (2014) examine the use of the Delphi method for arriving at probabilities of death and transition between different health stages of Hepatitis C virus. They found that the Delphi method was unsuccessful at establishing precise probabilities because there was a large discrepancy between the answers from experts, and no consensus could be reached. However, whilst this may have been seen as a problem within the field of public health, it should be viewed in a very different light within the field of existential risk, where confidence in predictions is often overstated. Another recent use of a structured expert elicitation with similarities to the Delphi Method was to assess the probability and extent of extreme ice sheet melting to better evaluate the risk of future sea level rise due to climate change (Bamber and Aspinall 2013).

Since the Delphi method aims at producing a consensus among experts it tends to reduce the independence of their individual judgements, despite allowing for the anonymous collection and processing of expert opinions. Furthermore, the presence of opinion outliers, and the fact that people vary in their willingness and ability to shift their views in light of new evidence, mean that convergence on a consensus may substantially differ to convergence upon the "correct" answer, or even the answer that makes best use of the group's collective knowledge and expertise. Nevertheless, reaching a consensus may be important for the purposes of establishing a probability claim that will be effective for coordination purposes, or if one merely wishes to create maximally authoritative claims about the relative importance of certain risks.

Another issue is that experts may feel compelled by the process to give answers even if they are completely uncertain, and unsure about how to quantify the uncertainty in a meaningful way. On the other hand, another off cited objection to the methodology is that participants may be seen to be "punished" for giving extreme or outlying results by having more work to do justifying their response.

6.3 Prediction Markets

Prediction markets function by providing a platform on which people can make trades whose value depends on their assessment of the likelihood of particular outcome or events, such as catastrophe or extinction event. The price at which people are willing to make these trades will then adjust dynamically to take account of people's predictions about these probabilities *and* their level of credence in these predictions. The more accurate any person is, the higher their pay-off. This thereby incentivizes individuals to be as accurate as possible, and allows for aggregation to take place over a potentially unlimited number of participants.

Prediction markets can use "play money" or real money. An example of the former is the website Metaculus.[11] This aggregates user's predictions on a number of different questions. For example, "What is the probability that SpaceX will land people on moon before 2030?"

[11] https://www.metaculus.com/questions/. Another example is Foresight Exchange: http://www.ideosphere.com/fx-bin/ListClaims

and “What is the probability that 20 more languages will go extinct by 2021”? Users’ predictions are also weighted by domain experience and past prediction track record. Those who tend to be more accurate have their prediction count for more compared to an individual with a less accurate track record. Users win “points” in virtue of their predictions accuracy once the question has been settled.

However, in order for markets to function well, predictions must generally be both specific and quantified, which can set a high bar for domain expertise and statistical ability before people can effectively participate. For example, it would be difficult for the lay person to assess the quantifiable properties of a mega disaster (e.g. total economic loss) given this this requires them to assess not only the size of the disaster but also the vulnerability of affected communities and the wider economic impacts of their loss. On the other hand, a benefit of this methodology is that it pools the “wisdom of the crowd” in a way that takes account of both beliefs and the strength of their convictions in those beliefs, and has been shown to greatly improve people’s naturally tendency to over commit to a belief since they know that something of value is on the line if they get things wrong. Therefore, whilst they place quite high epistemic demands on their participants, prediction markets can help democratize the assessment of probabilities. Pennock et al. (2001) note the success of prediction markets in domains as far apart as the probability of whether CERN will locate the Higgs boson and the probability that particular actors will win awards.

Another issue is the fact that individuals are able to see others’ predictions and therefore be open to influence to what has already been predicted, once more this undermines the independence of individual’s assessments. For instance, people who have outlying beliefs about the probability of such an event may find themselves unwilling to engage in a prediction market when they see how different their own price is to that set by the market, even if they have good reason to go with their own beliefs.

A further worry is that it makes little sense to use prediction markets for truly existential risks, because the bet would not pay out. Toby Ord has suggested this, along with the claim that it might make sense to use prediction markets for near miss cases, since a payout in such cases is possible. We can then use data about near misses to assist with arriving at a probability of the extinction event itself. This technique is likely to be most useful when there is a clear distribution of different events, such as the distribution of asteroid paths near to the earth or of different strengths of volcanic eruptions throughout history, but at less useful for novel cases where no such pattern can be easily deduced.

Although prediction markets have not been used systematically for the prediction of catastrophic events and extinction events,[12] there is scope for its greater use in probability assessment, particularly as a “supplement” to pre-existing data and expert opinion.

[12] A question on Metaculus (Will humans go extinct by 2100?) has the median probability of extinction at 1%, with 96 participants as of 18th February 2018: https://www.metaculus.com/questions/578/human-extinction-by-2100/

Furthermore, such probability assessments may prove fruitful for getting a rough estimate for the likelihood of particular events for the purposes of drawing attention to risks and prioritizing resources.

6.4 Super-forecasting

A final form of structured expert elicitation that has gained prominence in recent years is the use of "Super-forecasters". Super-forecasters are individuals who are remarkably successful at predicting future events. The idea of super-forecasting arose out of The Good Judgment Project. Over 2,000 individuals were selected by the group and tasked with assessing the likelihood of various world events. The project leader, Philip Tetlock, found that the most successful predictions were made by a concentrated group of skilled super-forecasters. These individuals had particular psychological traits that lead them to make more accurate predictions about a wide variety of global events than particular intelligence services. These traits included 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). Whilst some individuals already possessed all of these traits and were 'natural' super-forecasters, others were able to learn and develop these traits over time and greatly improve the accuracy of their predictions. Super-forecasters were also trained with elements of statistics and psychology to improve their performance. A possible means of establishing figures for catastrophic and existential risks is through the use of these super-forecasters.

Barbara Mellers et. al (2014) argue that probability training, teaming, and tracking lead to the improvement of forecasts, even from existing super-forecasters. 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 placed those who performed best in teams together. Although it appears to be possible to identify and select individuals with special traits and abilities, and train them in statistics, there is still a difficulty with predicting catastrophic and existential risks. Many such risks span over decades and centuries, and super-forecasters' reliability typically spans up to twelve months ahead. This is obviously problematic, given the wide horizons that existential risks refer to. It is also not obvious that super-forecasting will be effective for scenarios where there is severe uncertainty about the likelihood of particular events, such as human-level machine intelligence.

Super-forecasting limits individual bias and helps consolidate people's "natural" forecasting ability through training and institutional design, with the most successful forecasters group together to share their justifications. There is scope for greater use of this method, especially given the possibility that more formal methods of aggregation and elicitation could complement these "natural" forecasting abilities.

7. Discussion

In this section, we present a number of brief suggestions for improvements to current practice, as well as some tools that will be of use for thinking about the ways that probabilities are understood in existential risk studies.

7.1 Communicating the aim and context of claims about risk

As discussed at the very start of this paper, estimates about the probability of catastrophic risks can be produced for different purposes, and these purposes place different requirements for justifying such claims. This difference is important, yet it is not often explicitly made clear by those who write on existential risk. A suggestion for best practice is the explicit caveating of probabilities used in reports on the likelihood of existential risks.

Furthermore, given the diversity of probabilistic claims and the justifications for them, great care should be given to the suitability of qualitative probability estimates at all as these can easily be misinterpreted when read out of context. Perhaps it would be more appropriate if quantified probability estimates should be viewed merely as a general guide for identifying risks that may be orders of magnitude different from one another, but never treated is if they were precisely determined. The precise probabilities tends to be made in the literature on risk prediction and mitigation, rather than the literature on risk identification and prioritization. We therefore suggest that the intended *purpose* of probability assessments be made explicit.

7.2 Comparing Methodologies

Given the large number of available methodologies for assessing existential risks, we need to consider the question of whether some methodologies ought to be preferred to others. The answer to this question depends on what purpose one has in mind. If one is concerned with prediction then one will surely want a methodology (or set of methodologies) that gets as close as possible to the "actual" probability of an event. Whereas if one is concerned with coordinating the work of scholars then it will be more important that one produces an estimate about which there can be maximum agreement, which may mean preferring methods that are more transparent or that produce results that are easier to communicate.

For this reason, we have summarized some of our findings regarding the various methodologies currently used to study existential risk in the following table. The first column considers a methodologies rigor, or how arbitrary the probability assessment is, and to what extent techniques are employed to produce the most accurate and reliable assessments using scientific techniques. The second column considers how well the method operates under uncertainty, both in terms of being able to produce useful results even under conditions where uncertainty is high and in terms of being able to identify and quantify sources and levels of uncertainty. The third column considers how complex the methodology is. For example, the methodology of an individual subjective opinion can be incredibly simple, whereas the use of modelling or a multi stage structured expert elicitation is more complex, and involves more time and resources. The final column considers how easy it is to communicate probability assessments, by which we mean not only the probability figure itself, but also to its potential role in decision-making by practitioners.

Table 1.

Methodology	**Rigor of probability assessments**	**Ability to handle uncertainty**	**Simplicity of the methodology**	**Ease of communicating and using results**
1. Extrapolation from existing data	High	Low	High	Medium
2. Modelling	High	Medium	Medium	Medium
3. Individual subjective opinion	Low	High	High	High
4. Unstructured aggregation	Medium	Low	High	High
6. Classical approach	High	High	Medium	Medium
7. Delphi method	High	High	Low	High
8. Prediction markets	Medium	Medium	Medium	Medium
9. Super-forecasting	Medium	Medium	Low	Medium

The most notable previous attempt to compare and rank methodologies is that of Bruce Tonn and Dorian Stiefel (2013). The authors considered a number of methodologies for arriving at probabilities of existential risk, including: holistic probability assessments (where the assessor estimates an extinction risk through introspection and contemplation), a Whole Bayesian approach (where an assessor uses Bayes' Theorem to aggregate a set of mutually exclusive risks), Bayesian networks, environmental scanning, extinction scenarios, and possible worlds modelling (where risks of extinction in different paths leading to extinction scenarios are aggregated by a computer).

The authors evaluated these methodologies along the following criteria: (1) Level of effort required by the probability assessors; (2) Level of effort needed to implement the approach, including updating; (3) Ability of each method to model the human extinction event; (4) Ability to incorporate scientific estimates of contributory events; (5) Transparency of the inputs and outputs; (6) Acceptability to the academic community (e.g., with respect to intellectual soundness, familiarity, verisimilitude); (7) Credibility and utility of the outputs of the method to the policy community; (8) Ease of communicating the approach's processes and output to nonexperts; and (9) Accuracy in other contexts.

As with our own analysis, each methodology was found to have its strengths and weaknesses, and an overall ranking was not provided. Tonn and Stiefel suggested using all methodologies together, by giving "results of all methods to a panel of experts to reflect upon before they are asked for holistic assessments" (2013, p. 1784). They argue that the methodologies that map out the contributory factors to existential risks, such as environmental scanning and extinction scenarios can be synthesised with those that aim to provide and aggregate probabilities for these events. Causal models of existential risks can be supplemented with the Bayesian network method to provide a methodologically diverse assessment of existential risks.

However, this strikes us as leaving out the important issue of context in selecting the appropriate method for studying existential risks. Whilst each methodology has its own strengths and weaknesses we feel that some are clearly more appropriate for different tasks than others. For instance when existential risks stem from natural phenomena whose frequency is relatively easy to determine then the frequentist analysis of the previous historical record is extremely useful, however when the phenomena being studied is anthropogenic and/or involves a very high degree of uncertainty this approach may be worse than usefulness. Similarly, for purposes of raising awareness and providing the basis for consensus building and coordination we believe that consensus driven processes such as the Delphi Method show distinct promise. However, where they can be produced, even relatively simple models, such as Halliwell's Fault Tree analysis of the probability of nuclear war, have the potential to produce more analytically rigorous insight into the fundamental nature of a risk, even if they are also easier to dismiss.

Perhaps a panel of experts would we well placed to navigate these issues, however that is by no means certain, and the risk of putting all methodologies together is that we risk confusing the context within each of them is most appropriate and illegitimately transferring results from one context to another.

Rather than suggest that *all* of the methods that either we or Tonn and Steifel have considered should be aggregated together, our suggestion is that one be sensitive to the demands and purposes of each domain. Bayesian networks may be appropriate for some domains, where probabilities for some events have been assessed but inappropriate for others. We submit that the most appropriate way to study existential risks is to *cluster* methodologies that (a) have been demonstrated to be efficacious within the domain (such as extrapolation from data sets for potential pandemic pathogens, or expert elicitation in artificial intelligence) and that are (b) appropriate to the context within which one is operating

7.3 Structured Expert Elicitation

Nevertheless, we find that the use of Structured Expert Elicitation appears to be especially underdeveloped within the field of Existential Risk, and that this deserves more attention. We submit that something like the Delphi Method be recommended for arriving at probability assessments. Given its general lack of application to predictions on existential risk, there is rich scope for this method to be tested in different domains. There has been much written on the viability of the Delphi method in the areas of conservation (Mukherjee et al, 2015; Sutherland, 2017) and bioengineering (Wintle et al. 2017), but not existential risks in particular. For instance, given that there was a lack of a transparent methodology for establishing probabilities in the *Global Challenges Report*, the Delphi approach may be an appropriate addition to their existing methodologies for future updates.

We also make the following two suggestions for improving this method for use in the study of existential risk. Firstly, expert opinion, when it is used, ought to be made more explicit. Most authors are upfront about the shortcomings of holistic probability assessments, and do admit that some probability assessments are assumptions or best-guesses, but are less forthcoming about exactly how they produced the estimate that they did or what factors appeared most salient to them in producing it. A second suggestion is for greater structure to expert elicitation. Often in the case of holistic probability assessments it is the assessor themselves giving a figure based on the evidence available to them. In the surveys that aggregate expert opinion there is little structure to ensure that appropriate experts are recruited who are able to review the evidence available, or that the experts are informed about well-known biases in human judgment.

7.4 Ranking the Quality of Probabilistic Evidence

Despite the need for sensitivity and context dependency in the choice of method and communication of results, it is still an undoubted fact that one can, and should, distinguish between better and worse probability estimates. For example, an expert in the domain of artificial intelligence may consider a limited amount of evidence and then reflect on what they take to be the probability of an existential threat within the next century, and another might use a form of structured expert elicitation like the Delphi method that compares their own judgement to that of other experts. We submit that, for most purposes, the latter ought to be preferred to the former. It is similarly preferable to employ structured methods for eliciting probabilities rather than the best intuitive judgements of a single expert.

This is basic science, and common sense. However, it is arguable that within the field of existential risk 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, but it arguably also holds back the development of the field by failing to stimulate scholars to improve the quality of probability claims that they make. In order to move towards an approach that explains these judgements, we outline a “hierarchy of evidence” for probabilities. There is a large literature on the ranking of evidence in medicine (Evans, 2003), but there is very little on the ranking of probabilistic information.

To begin with probabilistic information should be ranked with respect to arbitrariness. For instance, objective probabilities, derived from the relative frequency of events, are often considered preferable for a variety of contexts than subjective probabilities derived from expert opinions, in part because they are seen as less arbitrary and in part because their arbitrariness, for instance the arbitrariness of sample selection, is seen as less prone to bias.

Similarly, subjective judgements that are based on a more substantial evidence base, and that present a clear, if not necessarily perfect, method by which they were produced should be preferred to those that do not for the same reason. The aggregation of expert opinion also helps arrive at an estimate that smooths out some of the outlying estimates, although where this aggregation involves additional deliberation between experts this smoothing does not always lead to a reversion to the mean but can in fact lead individuals all to shift towards the most extreme views under discussion.[13] The ranking then moves to those probability estimates that arise from a single expert, and then to those that arise out of an intuitive guess.

This list provides a rough outline of the types of processes that give a probability estimate. This is determined by the level of arbitrariness involved in specifying the probability. For instance, how much evidence was used as a basis for the probability judgment, and how was expert opinion treated? Was it simply aggregated, or were there measures taken to calibrate the experts and weight their contributions due to previous competence? Although we have so far only sketched this approach, it is an area for further research that I think will be fruitful for discussions on the use of probabilities in existential risk. Particular domains of existential risk may only have access to some parts of the hierarchy, for instance AI and subjective probabilities, but this means that where possible one should aim for probabilistic information that it "higher up" the hierarchy.

7.5 Confidence

Finally however, given, as we have shown, that there are multiple ways that one can both *establish* and *use* probabilities in existential risk assessments, we argue that more should be done to communicate the context and purpose of these estimates.

One suggested way of improving the communication of probabilities to a wider audience is make use of what is called a "confidence measure". Probabilities are often presented without much context. For example, that "there is a 0.0001 chance of extinction this century through an asteroid strike". There is rarely a communication of the amount of *confidence* that an assessor has in the probability assessment. A larger body of evidence is likely to be reflected in a higher level of confidence. Holding the amount of evidence fixed, greater agreement in the expert judgement based on it engenders greater confidence.

This concept has been most readily used with respect to climate change, most notably with the Intergovernmental Panel on Climate Change (IPCC) reports. In these reports, there is a use of an uncertainty framework. This combines probability judgments with confidence judgements. For example, the following statement from the IPCC:

[13] Sunstein, C. R., 2000, Deliberative trouble? Why groups go to extremes. *The Yale Law Journal*, *110*(1), 71-119.

> "In the Northern Hemisphere, 1983-2012 was *likely* the warmest 30-year period of the last 1400 years (*medium confidence*)" (IPCC, 2013: 27, my emphasis).[14]

Here, "likely" refers to a probability judgment. In the uncertainty framework, a variety of terms are used to refer to probability intervals: virtually certain (probability of the occurrence of the outcome is 99-100%), very likely (90-100%), likely (66-100%), about as likely as not (33-66%), unlikely (0-33%), very unlikely (0-10%), and exceptionally unlikely (0-1%) (Mastrandrea et al., 2010, 679). These probabilities are then couched by a confidence measure which "integrates the evaluation of evidence and agreement in one metric". 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, 679).

Confidence judgments can help communicate probabilities of existential risk. For example, a probability assessor might be a lot more confident in the probability estimates of catastrophic asteroid impacts, but a lot less confident in judgments about deadly AI, reflecting both the scarcity of evidence that can be used to arrive at and support such judgements and the difficulty in turning what evidence we have into robust and accurate predictions. A potential strength of this approach is that is can be sensitive to particular limitations within a domain. For example, if it is the case that for estimations of deadly AI the only available evidence base is expert opinion, then more robust methods of eliciting expert opinion will engender a *higher* confidence in probability assessments than a method that merely considered the intuitive judgements of experts and aggregated them. I submit that the communication of the uncertainties present in existential risk will be benefitted by including a reflection on the amount of confidence that assessors have in the available evidence and agreement between experts. In some domains this may be difficult, or "confidence" may be low, but including a measure of confidence will help incentivize the pursuit of greater agreement between experts.

Conclusion

This paper has reviewed a number of methodologies that have been used to establish probabilities of existential risk events, and provided three suggestions for improvements to the use of probabilistic information in this field. As we have argued, the appropriate methodologies for assessing existential risks will turn on the *use* of these probabilities. However, given the fact that probabilities are often assumed to be used for prediction, it is worth highlighting the fact that often it is often the other two uses of probabilities that are at play in existential risk claims.

[14] This quote was sourced from Wuethrich (2017).

We also make three suggestions for improvement are the following. Firstly, we proposed that structured expert elicitation deserves more widespread use amongst scholars of existential risk, and that a modified version of the Delphi method may be the most appropriate currently available for this. Secondly, we proposed the greater care should be taken in communicating the context for which probability claims were intended when they were produced alongside the claims themselves. Thirdly, we proposed that the existential risk community should be more discerning in its use of evidence and should attempt to produce probability claims using only the best available methodologies. Finally, we considered the potential benefits of using a confidence measure alongside regular probability assessments. Using a confidence measure will assist in the communication of probabilities, by signaling the amount of evidence and level of agreement when arriving at a probability assessment.[15]

References (excluding references cited only in the appended literature review)

Aspinall, W., 2010," A route to more tractable expert advice", *Nature*, 463(7279), 294.

Barrett, Anthony M. and Seth D. Baum, 2016, "A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis", *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 29 (2), pp. 1-18

Bamber, J.L. and Aspinall, W., 2013, "An expert judgement assessment of future sea level rise from the ice sheets", *Nature Climate Change*, 3(4), p.424.

Cirkovic, Milan M., & Anders Sandberg, & Nick Bostrom, 2010, "Anthropic Shadow: Observation Selection Effects and Human Extinction Risks", *Risk Analysis*, Vol. 30, No. 10, pp. 1495 – 1506.

Clauset, Aaron, Maxwell Young, & Kristian Skrede Gleditsch, 2007, "On the Frequency of Severe Terrorist Events", *Journal of Conflict Resolution*, 51, pp. 58-87.

Colson, A.R. and Cooke, R.M., 2018, "Expert elicitation: using the classical model to validate experts' judgments", *Review of Environmental Economics and Policy*, 12(1), pp.113-132.

Cooke, Roger, 1992, *Experts in Uncertainty*: *Opinion and Subjective Probability in Science*, Oxford University Press: Oxford.

Cousien, Anthony, et. al., 2014, "Is Expert Opinion Reliable when Estimating Transition Probabilities? The Case of HCV-related Cirrhosis in Egypt", *BMC Medical Research Methodology*, 14:39, pp. 1-9.

[15] This research was supported by a grant from the Templeton World Charity Foundation on Managing Extreme Technological Risks. The views expressed are those of the authors and do not necessarily reflect that of the foundation. The research of Simon Beard was also supported by a short-term visit to the Centre for the Philosophy of the Natural and Society Sciences at the London School of Economics. The authors are grateful for feedback from Adrian Currie, Phil Torres, researchers at the Centre for the Study of Existential Risk and the audience at the Cambridge Conference for Catastrophic Risk 2018.

Evans, David, 2003, “Hierarchy of Evidence: a Framework for Ranking Evidence Evaluating Healthcare Interventions”, *Journal of Clinical Nursing*, Vol. 12, Issue 1, pp. 77-84.

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.

Inglesby, Thomas & David A Relman, 2016, “How likely is it that biological agents will be used deliberately to cause widespread harm?”, *EMBO Report*, 17(2), pp. 127-130.

IPCC, 2013, *Climate change 2013: The physical science basis. Contribution of Working Group I to the fifth assessment report of Intergovernmental Panel on Climate Change*, ed. Thomas F. Stocker, Dahe Qin, Gian-Kasper Plattner, Melinda M. B. Tignor, Simon K. Allen, Judith Boschung, Alexander Nauels, Yu Xia, Vincent Bex, and Pauline M. Midgley. Cambridge, UK/New York: Cambridge University Press.

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*, Vol. 30, pp. 1157-75.

Mastrandrea, Michael D, et. al, 2011, “The IPCC AR5 guidance note on consistent treatment of uncertainties: a common approach across the working groups”, *Climatic Change*, Vol. 108, pp. 675-691.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., . . . Tetlock, P. E., 2014, Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*, 1106–1115. http://dx.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, pp. 7176-7184.

Pennock, David M.; Lawrence, Steve; Giles, C. Lee; Årup Nielsen, Finn, 2001, "The real power of artificial markets", *Science*. 291 (5506): 987–88. doi:10.1126/science.291.5506.987

Rees, Martin, 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.

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).

Sunstein, C. R., 2000, “Deliberative trouble? Why groups go to extremes”, *The Yale Law Journal*, Vol. 110(1), pp. 71-119.

Tetlock, P.E., Mellers, B.A. and Scoblic, J.P., 2017, “Bringing probability judgments into policy debates via forecasting tournaments”, *Science*, 355(6324), pp.481-483.

Tonn, B, and Stiefel, D, 2013, “Evaluating Methods for Estimating Existential Risks”, *Risk Analysis*, Vol. 10, pp. 1772-87.

Tonn, B, and Stiefel, D, 2012, “The Race for Evolutionary Success”, *Sustainability*, Vol. 4, pp. 1787-1805.

Wintle, B.C., Boehm, C.R., Rhodes, C., Molloy, J.C., Millett, P., Adam, L., Breitling, R., Carlson, R., Casagrande, R., Dando, M. and Doubleday, R., 2017. Point of View: A transatlantic perspective on 20 emerging issues in biological engineering. Elife, 6, p.e30247.

Woo, Gordon, forthcoming, “Counterfactual Disaster Risk Analysis”, *Variance*, Accepted February 26, 2016, pp. 1-30.

Wuethrich, Nicolas, 2017, “Conceptualizing Uncertainty: An Assessment of the Uncertainty Framework of the Intergovernmental Panel on Climate Change”, *EPSA15 Selected Papers*, pp. 95-107.

Yudkowsky, E., 2015, “When (not) to use probabilities. In *Rationality from AI to Zombies*”, Machine Intelligence Research Institute, pp. 1499-1505.

Appendix: Review of claims relating to the likelihood of existential threats and global catastrophes

This literature search was 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. The third was an early version of the Doomsday Database, a semi-automated system for finding publications that are relevant to x-risks hosted on x-risk.net. Further papers where identified by searching through works cited and via conversations with colleagues at the Centre for the Study of Existential Risk and elsewhere.

Nuclear War

The predictions listed below are of the chance of nuclear war. Most of the existential risk stems from the likelihood of a nuclear winter, and only a subset of all possible nuclear wars would cause a nuclear winter.

1. Source: Martin Hellman, “Risk Analysis of Nuclear Deterrence”, 2008, pp. 20-21.

   Probability: Annualised probability of a Cuban Missile Type Crisis (CMTC) resulting in World War III is given as:

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

   λ_{IE}: Probability of an “initiating event” (a potential first cause of CMTC) is **0.06**. There were three such events in the last 50 years. P_1: Because only 1 of the 3 initiating events actually was the Cuban missile crisis, the probability of an initiating event resulting in a CMTC is **1/3**. P_2: Then a conditional probability of a CMTC leading to the use of a nuclear weapon is arrived at: **0.1 to 0.5**. P_3: A fourth probability, that of the use of a nuclear weapon results in full scale nuclear war is given: **0.1 to 0.5**.

   Methodology: This paper analyses the possibility of a “Cuban Missile Type Crisis” (CMTC).
   The paper begins by noting that 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 deployment of American missiles in Eastern Europe.
   Taking the average rate of these possible initiating events (3 in 50 years) it is possible to arrive at the annualized probability of an initiating event as being 0.06 (which is 3 divided by 50). Already there are potential problems in how to define an “initiating event”. 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 Yom Kippur War of 1973.

From here, a probability is assigned (conditional on an initiating event) as being 1/3 that a CMTC will actually occur. This is based on the fact that only one of these three initiating *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.

A third conditional probability is calculated for the use of a nuclear weapons in the event of a CMTC occurring. Since this hasn't happened before, the authors rely on *subjective probability estimates* from those involved in the Cuban missile crisis (ranging from 0.01 to 0.5). This is a large range. The authors update the lower bound of the probability estimate to 0.1 to accommodate the fact that the participants stated their estimates before the Russian battlefield nuclear weapons were known in the West, which gives an updated range of 0.1 to 0.5).

Finally, the author calculated the conditional probability that the use of nuclear weapons would result in full-scale nuclear war. The authors use Kennedy and McNamara's personal estimates to arrive at the same probability bound of 0.1 to 0.5 for this

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.

2. Source: Hellman, Martin, 2011, "How Risky is Nuclear Optimism?", *Bulletin of the Atomic Scientists*, pp. 47-56.

   Probability: Risk of **10%** that there will be a nuclear war in the next century.

   Methodology: Hellman draws in this paper on the lower bound estimate from his 2008 paper of the probability of a Cuban Missile Type Crisis leading to a nuclear war. In this paper, Hellman reflects on the fact that this preliminary risk analysis was "endorsed by a number of prominent individuals", and produced a statement that "urgently called on the international community to undertake in-depth risk analyses of nuclear deterrence". This figure is therefore an out of context rendering of his previous findings, based upon the assumption that it will be easiest to coordinate international action around a lower bound estimate of just one potential cause of nuclear war then upon something closer to the true probability for such a war, but that may be the subject of less agreement.

3. Source: Antony Barrett et. al., "Analysing and Reducing the Risks of Inadvertent Nuclear War Between the United States and Russia", *Science and Global Security*, 21.

No. 2, 2013, pp. 106-33.

Probability: 90% confidence that the annual probability of accidental nuclear war between the US and Russia is **from 0.001% to 7%**.
"The assumed annual probability of occurrence of a U.S.–Russia crisis is given by the triangular (**0, 0.02, 0.06**) distribution. The lower bound is **0** if U.S.–Russia crises are no longer possible. The most likely value of **0.02** is based on the Hellman best estimate [discussed above] of one crisis in 50 years; that corresponds to counting the Cuban Missile Crisis as the only historical crisis. The upper bound value of **0.06** is based on the Hellman estimate of three possible events in 50 years" (p. 118).

Methodology: Mathematical modelling and fault tree analysis to estimate 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).

4. Source: Lundgren, Carl, 2013, "What are the Odds? Assessing the Probability of a Nuclear War", *Nonproliferation Review*, Vol. 20, No. 2, pp. 361-374.

   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. Lundgren notes that due to the absence of a previous nuclear war, there is little data on its probability. For this reason, Lundgren opts for Bayesian statistical reasoning, based on "an especially applicable mathematical method of calculating probabilities where only limited data are available and assured knowledge is not possible" (p. 362).

5. Source: Pamlin, Dennis, & Stuart Armstrong, 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).[16] "Infinite impact" refers to the state where civilization collapses to the state of great suffering and does not recover, or a situation where all human life ends, and "infinite impact threshold" refers to an impact that can trigger a chain of events that could result first in a civilization collapse and then later result in an infinite impact (Dennis & Armstrong, 2015: 11).

[16] The sources given for this probability estimate are the following two risk analysis websites: http://metabiota.com/ and http://instedd.org/.

Methodology: Probability estimates from expert opinion. This constituted an "expert review" of the relevant literature and relevant risks. "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." (p. 12).

Pathogen with Pandemic Potential

Here the catastrophic risk stems from the possibility of an escape of a pathogen with pandemic potential from a laboratory. 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.

1. Source: Troy Day, Jean-Baptiste Andre & Andrew Park, 2006, "The Evolutionary Emergence of Pandemic Influenza", *Proceedings of the Royal Society – Biological Sciences*, 273, pp. 2945-2953.

   Probability**:** Probability of a pandemic occurring in any given year is **4%** (p. 2945). The probability (and source) is used by Tonn and Stiefel (2013) to arrive at the probability of a pandemic of **4 x 10^{-2}**.

   Methodology: This probability is derived from historical data on pandemics: approximately 10 influenza pandemics in the past 250 years.[17] This is described by Day et al. (2006) as "anecdotal evidence" (p. 2945).

2. Source: Marc Lipsitch & Thomas V. Inglesby, "Moratorium on research intended to create novel potential pandemic pathogens", *MBio*, 5, 2014, pp. 1-6.

   Probability: **0.2%** chance of a laboratory-acquired infection per BSL3 laboratory year (this probability is also used by Inglesby & Relman, 2016), and **1%** chance of a laboratory-acquired infection per full-time worker year. Probability that an infection would lead to a global spread: **5% to 60%.** This leads to a probability of a transmissible form of influenza virus creating a pandemic of between **0.01% and 0.1%** per laboratory-year of creating a pandemic, using the select agent data, or between **0.05% and 0.6%** per full-time worker-year using the NIAID data.

[17] The authors back up this claim with evidence from the following sources: Robert G. Webster, 1998, "Influenza: An Emerging Disease", *Emerging Infectious Diseases*, Vol. 4, No. 3, pp. 436-441, p. 437, and Ann H. Reid, Jeffery K. Taubenberger & Thomas G. Fanning, 2004, "Evidence of an Absence: the Genetic Origins of the 1918 Pandemic Influenza Virus", *Nature Reviews Microbiology*, 2, pp. 909-914.

Methodology: The starting point for quantifying the risk is the "record of laboratory incidents and accidental infections in biosafety level 3 laboratories" (p. 2). Concentrating on the generation of transmissible variants of avian influenza, the authors provide a calculation of the sort that would be performed in greater detail in a risk analysis. Using data on past laboratory infections in the US, a probability of future infections is extrapolated. 4 infections have been observed over <2,044 laboratory-years of observation, "indicating at least a 0.2% chance of a laboratory acquired infection per BSL3 laboratory year".

Another method mentioned was the use of data from BSL3 labs at the National Institutes of Allergy and Infectious Diseases, which reported 3 accidental infections in 634,500 person-hours of work between 1982 and 2003, or about 1 accidental infection for every 100 full-time person-years (2000hrs of work). This yielded the probability of an accidental infection per full-time worker-year of **1%**. This is larger than the probability calculated based on infections per lab year (**0.2%**). "A simulation model of an accidental infection of a laboratory worker with a transmissible influenza virus strain estimated about a 10 to 20% risk that such an infection would escape control and spread widely". Apparently, alternative estimates from simple models range from **5%** to **60%** that an infection would escape control and spread widely.

These values (of escape and spread) are then multiplied by the probability of an accidental laboratory infection, yielding the risk of between 0.01% and 0.1% per laboratory year of creating a pandemic, or between 0.05% and 0.6% per full-time worker-year.

3. Source: Ron Fouchier, 2015, "Studies on Influenza Virus Transmission between Ferrets: the Public Health Risks Revisited", *MBio*, Vol. 6, No. 1, pp. 1-4.

   Probability: This paper revises down the probability of a global pandemic in Lipsitch and Inglesby (2014) and says that "the published estimates were based on historical data and did not take into account the numerous risk reduction measures that are in place in the laboratories where the research is conducted" (p. 1). This changes the risk from serious to negligible.

   From an analysis of how particular safety measures impact upon the probability of a laboratory-acquired infection (LAI), Fouchier states that the probability of an LAI will be below **0.2 x (5 x 10^{-7}), or <1 x 10^{-7} per person-year**. This, Fouchier states, leads to 1 LAI occurring less frequently than once every 1 million years.

   After considering how various safety measures will reduce the risk of onward transmission (if an infection took place), Fouchier arrives at the probability of transmission of an LAI as being **<2.5 x 10^{-6} to 3 x 10^{-5}**.

   By multiplying the probability of occurrence of an LAI by the probability of onward transmission, the estimated probability of an LAI resulting in an onward transmission ranges between **$(1 \times 10^{-7}) \times (2.5 \times 10^{-6})$ (or 2.5×10^{-13}) and $(1 \times 10^{-7}) \times (3 \times 10^{-5})$ (or 3×10^{-12})**.

As such, Fouchier states, "1 LAI with onward transmission would be expected to occur far less frequently than **once every 33 billion years** (p. 3).

Methodology: In discussing the probability of a LAI, Fouchier revises down the probability of an infection in light of particular safety measures that are in place in the relevant laboratories. He says that "although the magnitude of this increase in safety is not known, I assume that it is at least a factor of 10" (p. 2). He then uses "the risk analysis done by others and the assumption of reduced risk" to arrive at the probability. The probability is revised down further by considering the "general accepted efficacy of influenza vaccine of around 65%", which leads to a probability of an LAI leading to viral escape being .35. So, Fouchier uses the analysis of Lipsitch and Inglesby as a starting point, and then uses pre-existing data regarding the expected safety of particular measures to modify these probabilities.

4. Source: Pamlin, Dennis, & Stuart Armstrong, 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).[18]

   Methodology: Probability estimates from expert opinion. See source 5 in Nuclear War section above.

[18] Sources for the estimates include: Bagus, Ghalid (2008): Pandemic Risk Modeling http://www.chicagoactuarialassociation.org/CAA_PandemicRiskModelingBagus_Jun08.pdf, Broekhoven, Henk van, Hellman, Anni (2006): Actuarial reflections on pandemic risk and its consequences http://actuary.eu/documents/pandemics_web.pdf, Brockmann, Dirk and Helbing, Dirk (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 Outbreaks. 2013 Jul 17. Edition 1. doi: 10.1371/currents. outbreaks.a7b70897ac2fa4f79b59f90d24c860b8. 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,
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

Bioweapons

This section considers the catastrophic and existential risks associated with biological warfare. Out of all predicted wars, only a subset will involve biological weapons. From this, only an extremely small number will lead to conflicts with the risk of causing extinction.

1. Source: Piers Millett and Andrew Snyder-Beattie, 2017, "Existential Risk and Cost-Effective Biosecurity", *Health Security*, Vol. 15, No. 4, pp. 1-11.

   Probability: The paper examines three different approaches to approximate the risk of extinction from bioweapons: (1) utilization of surveys of experts, (2) previous major risk assessments, and (3) simple toy models.

   (1) Use of Sandberg and Bostrom (2008). Participants in their study had a median risk estimate of 0.05% that a natural pandemic would cause human extinction by 2100, and a median risk estimate of 2% that an "engineered" pandemic would lead to extinction by 2100 (Millet and Snyder-Beattie, p. 5). They acknowledge the weaknesses of this study.

   (2) *Existential risk arising from deliberately created potentially pandemic pathogens*.
   Citing Gryphon Scientific (2015), the authors 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%**.[19]
   Following this, the authors write that: "The Gryphon report also concluded that 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**" (p. 6).[20]
   The authors then estimate the probability that a pandemic will cause an existential risk: "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 existential risk**" (Ibid.).[21]
   Now these two sets of probabilities are combined to arrive at the probability of an existential risk arising from a global pandemic:

[19] There is no explicit reference to these particular probabilities in the original report.

[20] As a note, this is a very wide probability interval, wider still than the original one presented at the beginning of the paragraph.

[21] 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.

"Multiplying the probability of an outbreak with the probability of an existential risk gives us an annual risk probability between **1.6 x 10^{-8} and 8 x 10^{-7}**". (Ibid.).

Existential risk arising from bioweapons.
Annual existential risk from terrorist attacks using biological and chemical weapons is **0.0000014 (or 1.4 x 10^{-6})**.[22]

Existential risk from bioweapons.
"Assuming that 10% of biowarfare escalations resulting in >5 billion deaths eventually leads to extinction we get an annual **existential risk from biowarfare of 0.0000005 (or 5x10^{-7})**.
For this probability, the evidence base is the historical record on the number of wars (97 wars between 1820-1997). The parameter given for warfare is 0.41. They assume that wars will occur again with the same frequency as 1820-1997.

Methodology: This is the methodology in the authors' own words: "In order to produce **reasoned estimates** of the likelihood of different categories of biothreats, we bring together **relevant data and theory** and produce some **first-guess estimates of the likelihood** of different categories of biothreat" (p. 3). There is a use of extrapolation from existing data-sets on past biological warfare.

Climate Change

Climate change constitutes an existential and catastrophic risk due to the chance that global average temperatures increase far enough to cause uninhabitable conditions and the degradation of the global economy. Many of the predictions below utilize the modelling of the IPCC report to generate estimates of catastrophic climate change.

1. Source: Stern, Nicholas, 2006, Stern Review on The Economics of Climate Change, H.M. Treasury, London [retrieved 12th July 2017].

   Probability: the probability of human extinction by the end of the century is **10%**.

[22] This probability is calculated by extrapolating from a past study from Clauset et al (2007) which states that casualty numbers from terrorism and warfare follow a power law distribution. This means that "the ratio in likelihood between events that cause the deaths of 10 people and 10,000 people will be the same as that between 10,000 and 10,000,000 people" (Millet and Snyder-Beattie, p. 6). Clauset et al (2007) estimate the ratio for terrorism using biological and chemical weapons to be about 0.5 for one order of magnitude. Millet and Snyder-Beattie extrapolate from this to find the probability that an attack will kill over 5 billion. In arriving at the final probability, they also assume that there would be one attack per year, and that 10% of such attacks that lead to over 5 billion people killed will eventually lead to extinction.

Methodology: Subjective probability estimate. However, this probability of extinction is not related to climate change, but occurs within the context of a discussion of climate change when discussing an appropriate discount rate. The figure is selected to be one that his audience will not see as too low, given that most would prefer to see the discount rate higher. This is a case of an estimate being made for the purposes of *coordination* rather than prediction.

2. Source: Gernot Wagner and Martin Weitzman, 2015, *Climate Shock*, pp. 53-56. (inferred estimates of probabilities from IPCC figures)
   Probability: On a low-medium emissions scenario, there is **at least a 3% chance** of eventual 6 degrees warming (with significant uncertainty). On the medium-high emissions scenario, the chance could be **around 10%.**

   Methodology: Based on assessments of known levels of uncertainty (as classified by the IPCC) in current predictions of climate change and the likely distribution of different future scenarios within these.

3. Source: Pamlin, Dennis, & Stuart Armstrong, 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).[23]

   Methodology: Probability estimates from expert opinion. See source 5 in Nuclear War section above.

[23] Stated sources for the estimates include: Fekete, Hanna; Vieweg, Marion; Rocha, Marcia; Braun, Nadine; Lindberg, Marie; Gütschow, Johannes; Jefferey, Louise; Höhne, Niklas ; Hare, Bill; Schaeffer, Michiel; Macey, Kirsten and Larkin, Julia (2013): Analysis of current greenhouse gas emission trends http://climateactiontracker.org/assets/publications/publications/CAT_Trend_Report.pdf, New, Mark G.; Liverman, Diana M.; Betts, Richard A.; Anderson, Kevin L. and West, Chris C. (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., and Co-authors, 2009: Probabilistic Forecast for Twenty-First-Century Climate Based on Uncertainties in Emissions (Without Policy) and Climate Parameters. J. Climate, 22, 5175–5204. http://journals.ametsoc.org/doi/abs/10.1175/2009JCLI2863.1, 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.pd

4. Source: Raftery, Adrian E., 2017, “Less than 2 degrees Celsius warming by 2100 unlikely”, *Nature Climate Change*, published online. https://www.nature.com/nclimate/journal/vaop/ncurrent/pdf/nclimate3352.pdf

   Probability: “**5%** chance that [global temperature increase] will be less than 2 degrees Celsius [by 2100]” (p. 1).[24]

   Methodology: Statistical modelling. “We build a joint Bayesian hierarchical statistical model for GDP per capita and carbon intensity in most countries, and combine it with the UN probabilistic population projections to produce a predictive distribution of quantities of interest to 2100. We develop a probabilistic forecast of global temperature increase by combining them with the relationship between cumulative CO2 emissions and temperature used by the IPCC” (p. 1). “Our approach is fully statistical” (p. 4).

5. Source: Ian Dunlop & David Spratt, 2017, *Disaster Alley*: *Climate Change Conflict and Risk*, Breakthrough – National Centre for Climate Restoration

   Probability: **50% chance** of catastrophic climate change (global temperature increase above 4 degrees Celsius)

   Methodology: Their methodology is based on IPCC modelling. The authors correct what they take to be a flaw in the original analysis. This is the omission of “longer-term” carbon cycle feedbacks, such as permafrost thaw and terrestrial carbon sinks, which they say are now becoming relevant.

6. Source: Yangyang Xu and Veerabhadran Ramanathan, 2017, “Well below 2 degrees Celsius: Mitigation Strategies for Avoiding Dangerous to Catastrophic Climate Changes”, *Proceedings of the National Academy of Sciences*, Vol. 114, no. 39, pp. 10315-10323

   Probability: “Long Term (>2050): 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: Amendments are made to IPCC report model. Amendments are made to better capture the uncertainties in emissions scenarios in the original IPCC data.

[24] 2 degrees Celsius has often been regarded as the threshold for so-called ‘dangerous’ climate change.

7. Source: Halstead John, 2018, "Stratospheric Aerosol Injection Research and Existential Risk", *Futures,* online first
   Probability: "the unconditional probability of existential catastrophe-level warming is ~3.5%."

   Methodology: Individual best guess based upon reasonable assumptions about two key factors 1) the unconditional 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.

Asteroid Impact

The predictions below consider the possibility that of catastrophic asteroid impacts. Of potential asteroid impacts, only a very small subset are considered to be existential risks.

1. Source: NASA's Near-Earth Object Program Office, https://www.nasa.gov/mission_pages/asteroids/news/asteroid20131017.html

   Probability: **1 in 63,000** chance (**0.000016**) that asteroid 2013 TV135 will hit earth. The (expected) impact is believed to have the kinetic energy of 3,200 megatons of TNT, approximately 60 times that of the most powerful nuclear bomb ever detonated. This would also be 16 times the 1883 eruption of Krakatoa, which was 200 megatons.

   Methodology: It is not clear what precise methodology is used to calculate the probability of an impact. Some remarks are given in the following: https://web.archive.org/web/20131031093724/http://neo.jpl.nasa.gov/risk/2013tv135.html.

2. Source: Clark Chapman, 2004, "The Hazard of Near-Earth Asteroid Impacts on Earth", *Earth and Planetary Science Letters*, 222, pp. 1-15.

   Probability: **1 x 10^{-7}** annual chance of an extinction-level asteroid impact ("1 in 100,000 chance", p. 11).

   Methodology: Probability estimate from expert opinion.

3. Source: Nick Bostrom, 2006, “Dinosaurs, Dodos, Humans?”, *Global Agenda*, pp. 2-3.

   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).

   Methodology: Based on historical average of meteorite strikes on the earth. Quote did not come with a source.

4. Source: Pamlin, Dennis, & Stuart Armstrong, 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).[25]

   Methodology: Probability estimates from expert opinion. See source 5 in 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.

1. 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%** probability of human extinction from superintelligent AI before

[25] Stated sources for the estimates include: Jablonski, David, and W. G. Chaloner.: 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, Gareth S., H. Jay Melosh, and Robert A. Marcus.: 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., Monthly Notices of the Royal Astronomical Society 327.1 (2001): 126-132, and Board, Space Studies.: 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, Clark R., and David Morrison. (1994): Impacts on the Earth by asteroids and comets: assessing the hazard. Nature 367.6458, Chapman, Clark R., Daniel D. Durda, and Robert E. Gold. (2001): The comet/asteroid impact hazard: a systems approach. Office of Space Studies, Southwest Research Institute, Boulder CO 80302.

2100.[26]

Methodology: Informal survey asking participants at the Global Catastrophic Risk Conference in Oxford (July, 2008) for their best guesses at the chances there will be disasters of different types.

2. Source: Mueller, Vincent C., and Bostrom, Nick, 2014, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion", in Vincent C. Mueller (ed.), *Fundamental Issues of Artificial Intelligence*, Synthese Library; Berlin, Springer.

   Probability: Mean probability for the claim that human level machine intelligence would lead to extinction is **18%**.

   Methodology: Survey of 550 experts with different backgrounds in AI. 170 responded. The percentage results are means and not medians. 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 & "Impacts and Risks of Artificial General Intelligence" 2012, both at Oxford; (3) 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)".

3. Source: Pamlin, Dennis, & Stuart Armstrong, 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

[26] This study also gives extinction risk estimates in other areas such as nuclear wars (1%) and molecular nanotech weapons (5%) by 2100.

impact" (p. 158).[27]

Methodology: Probability estimates from expert opinion. See source 5 in Nuclear War section above.

4. Source: Grace, Katja et. al., 2017, "When Will AI Exceed Human Performance? Evidence from AI Experts", online, https://arxiv.org/pdf/1705.08807.pdf

   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: Survey of experts. 352 researchers (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.

Super-volcano eruption

These risks refer to the possibility of catastrophic eruptions of super-volcanoes. Volcanic ash clouds and tsunamis can have a disastrous global impact, by altering weather patterns and destroying coastal populations.

1. Source: Bethan Harris, 2008, "The Potential Impact of Super-Volcanic Eruptions on the Earth's Atmosphere", *Royal Meteorological Society*, Vol. 63, Issue 8.[28]

   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).
   Tonn and Stiefel (2013) state the *annualised* probability of such an eruption occurring as 2 x 10^{-5} using Harris (2008) as the source.

   Methodology: The methodology is made explicit. Calculations of probability are derived from frequencies of past eruptions. To this extent, the methodology is extrapolation from pre-existing data-sets.

[27] Sources used include: Bostrom, Nick and Sandberg, Anders (2008): Global catastrophic risks survey civil wars 98.3 http://www.fhi.ox.ac.uk/gcr-report.pdf, Kruel Alexander (2013): Probability of unfriendly and friendly AI, including comments to the blog post http://kruel.co/2013/09/23/probability-of-unfriendly-andfriendly-ai/#sthash.xRhFOGHW.dpbs

[28] Although not explicitly referenced as such, 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.

2. Source: Pamlin, Dennis, & Stuart Armstrong, 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).[29]

   Methodology: Probability estimates from expert opinion. See source 5 in Nuclear War section above.

3. Source: Torres, Phil, 2016, “How likely is an existential catastrophe?”, *Bulletin of the Atomic Scientists*, online article. http://thebulletin.org/how-likely-existential-catastrophe9866 [Accessed 11/08/17]

   Probability: “A supervolcanic eruption capable of inducing a “volcanic winter” happens about **once every 50,000 years**”.

   Methodology: Geological record: “Similarly, the geological record tells us that …”. No source given.

Ecological Catastrophe

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.

1. Source: Pamlin, Dennis, & Stuart Armstrong, 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: Probability estimates from expert opinion. See source 5 in 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

[29] Sources not included.

pathogen that targeted humans or ecosystems. There is a great deal of uncertainty about these risks.

1. Source: Pamlin, Dennis, & Stuart Armstrong, 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).[30]

   Methodology: Probability estimates from expert opinion. See source 5 in 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.

1. Source: Pamlin, Dennis, & Stuart Armstrong, 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).[31]

   Methodology: Probability estimates from expert opinion. See source 5 in Nuclear War section above.

[30] Stated sources include: Bennett, Gaymon, et al. "From synthetic biology to biohacking: are we prepared?" Nature Biotechnology 27.12 (2009): 1109-1111., Bostrom, Nick and Sandberg, Anders (2008): "Globalcatastrophic risks survey: civil wars" 98.3 http://www.fhi.ox.ac.uk/gcr-report.pdf, Mukunda, Gautam, Kenneth A. Oye, and Scott C. Mohr. "What rough beast? Synthetic biology, uncertainty, and the future of biosecurity." *Politics and the Life Sciences* 28.2 (2009): 2-26., Russ, Zachary N. "Synthetic biology: enormous possibility, exaggerated perils." *Journal of biological engineering* 2.7 (2008)., Radosavljevic, Vladan, and Goran Belojevic. "A new model of bioterrorism risk assessment." *Biosecurity and bioterrorism: biodefense strategy, practice, and science* 7.4 (2009): 443-451., Tucker, Jonathan B., and Raymond A. Zilinskas. "The promise and perils of synthetic biology." *New Atlantis* 12.1 (2006): 25-45.

[31] Stated sources include Altmann, Jürgen, and Mark A. Gubrud.: Military, arms control, and security aspects of nanotechnology. Discovering the Nanoscale (2004): 269 http://nanoinvesting.myplace.us/altmann-gubrud.pdf, Berube, David M., et al. "Communicating Risk in the 21st Century: The case of nanotechnology." National Nanotechnology Coordination Office, Arlington (2010)., Bostrom, Nick and Sandberg, Anders (2008): Global catastrophic risks survey civil wars 98.3 http://www.fhi.ox.ac.uk/gcr-report.pdf, Turchin, Alexey.: Structure of the global catastrophe. Risks of human extinction in the XXI century. Lulu. com,2008., Williams, Richard A., et al. "Risk characterization for nanotechnology." *Risk Analysis* 30.11 (2010): 1671-1679. These authors argue that there is currently insufficient information to create a risk analysis for nanotechnology.

Unknown Consequences

1. Source: Pamlin, Dennis, & Stuart Armstrong, 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).[32]

   Methodology: Probability estimates from expert opinion. See source 5 in Nuclear War section above.

[32] Source used was: Bostrom, Nick and Sandberg, Anders (2008): Global catastrophic risks survey civil wars 8.3 http://www.fhi.ox.ac.uk/gcr-report.pdf