CONFERENCE REPORT

THE CAMBRIDGE CONFERENCE ON CATASTROPHIC RISK 2018 (CCCR2018)

17th and 18th April 2018, Cambridge

ACKNOWLEDGEMENTS

The Cambridge Conference on Catastrophic Risk 2018 contributed to the activities of the Centre for the Study of Existential Risk's Managing Extreme Technological Risks research project, made possible through the support of a grant from the Templeton World Charity Foundation. Supplementary funding for workshop activities was received from the Hauser-Raspe Foundation.

The opinions expressed in this report are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation or of the Hauser-Raspe Foundation.

We are grateful to all of the speakers and participants who took part in this conference, each of whom brought their own perspectives on the challenges we face, both as a research community and a global civilization.

This report was written by Simon Beard and Catherine Rhodes with additional contributions by Luke Kemp. Lalitha Sundaram, Lauren Holt, Shahar Avin, Seán Ó hÉigeartaigh and Julius Weitzdörfer

CONFERENCE REPORT

THE CAMBRIDGE CONFERENCE ON CATASTROPHIC RISK 2018 (CCCR2018)

The second of the Centre for the Study of Existential Risk's international conferences provided a timely opportunity for the Centre, along with the wide communities working on existential and global catastrophic risks and in related fields, to reflect on our work so far and to deepen and broaden our learning from other disciplines. This allowed us to both focus on some of the practical challenges of the task that we've set ourselves and identify what we are doing well. Hence, the Conference served not only to address important issues facing the existential risk research community, but also to establish and maintain the connections with other communities that have important contributions to make to the developing 'science' of existential risk research.

Introduction ... 3

Theme 1: Black Elephants and Unsexy risks ... 5

Theme 2: What are we here for, what are we doing ... 7

Theme 3: Roulette Wheels, Fruit Machines and Risk Ladders ... 9

Theme 4: People care about different things, this is a problem that is not easy to solve ... 10

Theme 5: The Risk of Studying Existential Risk ... 12

Summaries of each Panel ... 14

- Panel 1: Challenges of Evaluation ... 14
  - Diane Coyle – What to watch out for: an economists perspective ... 14
  - Matthew Rendall – Discounting, Aggregation and the Ecological Fallacy ... 14
  - Silja Voeneky - The Public International Law Perspective on Evaluating Existential Risks ... 16
- Panel 2: Challenges of Evidence ... 16
  - Seth Baum - Making the Most of Limited Evidence in Global Catastrophic Risk Analysis ... 16
  - Simon Beard - Probabilities, methodologies and the evidence base in existential risk assessments ... 17
  - Tamsin Edwards - How soon will the ice apocalypse come? ... 17
- Panel 3: Challenges of Scope ... 18
  - Karin Kuhlemann – Sexy Vs Unsexy Catastrophic Risks ... 18
  - Andrew Maynard - Risk Innovation: The roles of creativity, imagination and rigour in exploring existential risk ... 19
- Panel 4: Challenges of Communication ... 20
  - Kristel Fourie - Understanding Risk for effective communication ... 20
  - Alex Freeman - Communicating (Existential) Risks Responsibly ... 20
  - Peter Ho – Black Elephants ... 21

Introduction

The Centre for the Study of Existential Risk is an interdisciplinary research centre within the University of Cambridge dedicated to the study and mitigation of risks that could lead to human extinction or civilizational collapse. We work to identify, manage and mitigate global risks associated with emerging technological advances and human activity, focusing on high-impact risks that might result in a global catastrophe or threaten human extinction, even if only with very low probability.

In particular, we are concerned with risks that are plausible and tractable but currently understudied and poorly characterised, and we contribute high quality research to increase understanding of these risks and the options for their management and mitigation.

In our Managing Extreme Technological Risks project, we explore how to approach the establishment of a new interdisciplinary sub-field that can tackle the challenges of understanding and managing extreme technological risks in a systematic way.

In this task we have implemented an initial model of the structure of a 'science' for the study and management of extreme risks, which consists of:

- Theoretical work within and across several sub-projects, done by an interdisciplinary group of postdoctoral researchers, bringing in and bridging with their 'home' disciplines;
- Practical work in engagement with academic, policy and technology communities and the public.

To foster this emerging science, we have adopted a methodology of implement, test, reflect and refine whereby we constantly strive to construct and improve new approaches to the study of high impact risks.

This methodology has helped us to identify key challenges that arise from the model that we're using and the subject matter we're addressing, although in many cases these challenges are the things that make work in this area particularly interesting.

It is from some of these challenges that the themes for this conference emerged. These were:

- Challenges of Evaluation and Impact;
- Challenges of Evidence for Existential Risk Research;
- Challenges of Scope and Focus: What Might We Be Missing / Neglecting; and
- Challenges in Communication and Responsible Engagement.

Although none of these challenges is unique to the study of global catastrophic and existential risk, they may have different aspects or implications in different fields.

The conference thus provided an additional opportunity to build links and gain insights from other areas represented among speakers and participants, and how those fields are approaching such challenges.

In reviewing different approaches to these challenges, we discovered that they shared many features in common, and more importantly that the strategies for overcoming them were very similar. Rather than being structured by these four challenges; this report is, therefore, based these on shared themes that emerged from the panel and workshop sessions, which showcase important connections between solutions to each challenge and highlight concrete actions that we and the wider existential risk research community can take to address them.

One of these is the need to give greater consideration to what reaches the agenda of the research community, and the processes through which issues get defined as relevant to the study of existential and global catastrophic risks, or not. Closely connected to this is the need to pay attention to how the community can foster diversity - in the disciplines it includes, the approaches it adopts and the people whom it engages with.

As an emerging field of study, this will also need to be balanced with some form of boundary setting, that maintains coherence and focus. A third related challenge is the need to choose what and how we will measure, evaluate and collate evidence when assessing existential and global catastrophic risks and their management and mitigation. This in turn, ought to be connected to a greater understanding of where we wish to see changes being made and practical actions being taken - knowing which audiences we want to reach and the best ways of communicating with them should be shaping how we approach our work from its early stages.

We should thus work towards improved attention to all these elements in our work, in a systematic way, and continue to facilitate discussion and communication on these issues both within CSER and across the broader research community. This process will also benefit greatly from our continued engagement with policy and technology communities and aligned academic disciplines.

Videos of our keynote lectures can be found at www.cser.ac.uk

Theme 1: Black Elephants and Unsexy risks

When an otherwise predictable event takes people by surprise because it is unprecedented or rare this is known as a 'black swan'. When something obvious is being deliberately ignored this is known as an 'elephant in the room'. In his presentation Peter Ho suggests combining these metaphors so that when something that is otherwise predictable takes people by surprise because it is being deliberately ignored this should be called a 'black elephant'.

These Black Elephants often coincide with another group of risks, abstract, long-term, uncertain, or inactionable problems outside the current zeitgeist, that we choose to live with rather than to mitigate. Karin Kuhlemann described these risks as 'unsexy' and pointed to several clear examples including overpopulation, climate change and long term, second order, effects of natural disasters and terrorist attacks.

However, availability bias and cognitive heuristics tend to focus risk-management on what is immediate, dramatic, and emotionally charged (often fear-driven) rather than what is truly most important. Therefore, in contrast, 'sexy' problems are more likely to attract attention, funding, and capture the imagination of both the public and those in power. These are often in the category of spectacular events that are low probability and binary in nature (e.g. asteroid strikes), have featured in a Hollywood movie, or are popular because of they offer tractable, feel-good and actionable solutions. Karin Kuhlemann suggests that sexy risks that are global, external or non-contingent on human foibles are more often portrayed and more amenable to technical solutions. In contrast unsexy risks are human-based and tackling them requires a combination of decreasing people's standard of living and sacrificing thir liberties (such as unrestricted reproduction). On the other hand, Andrew Maynard observes that we focus on what we can understand and solve, and either ignore the rest (like the possibility of making General AIs that are neither bodybuilders nor sexy fembots) or try to bring it into the small realm of what is comprehended and manageable.

What can be done to draw attention to these Black Elephants and other unsexy risks, and therefore attract more funding and effort to mitigating them? As illustrated by Andrew Maynard, what gives these risks the sense of being unsexy is often how they are framed or the political, practical or religious barriers to mitigating them. For instance, he pointed out that whilst the USA appears incapable of instigating gun control, in part because it is a too vast and entrenched problem, the risk of choking on Kinder Eggs is effectively and disproportionately regulated - although kinder eggs do not appear any 'sexier' in Karin Kuhlemann's terms, than guns. Andrew Maynard also pointed out that since we exist within certain systems that shape our thinking, and since some existential risks (especially from AI) come from outside of this system, we may need to use purely speculative techniques such as science fiction or philosophical analysis to shift our conceptual paradigm before we can engage with these risks. Diane Coyle further pointed out that it is not enough for some people to spot or engage with these risks if they are ignored by others and that Policy makers prefer to work with the illusion of certainty than the reality of doubt.

Seth Baum drew attention to a different issue with certain kinds of unsexy risk, that the only evidence for them may come from near misses, events where we appeared to move closer to a catastrophe but with little or no observed effect. These represent blind-spots in our thinking and often relate to systematic risks and second order effects.

Whatever we do to bring greater attention to these unsexy risks we now need to engage with the fact that we are operating in an environment of fake news and internet anonymity in which speculative techniques are already being widely misused to undermine established science and trust in scientists currently stands at only 43%. Conversely, however, social influencers can have a positive effect and help create decisive action.

We will have to find some way to see the Black Elephants, near misses and intractable problems as the sexy ones.

Theme 2: What are we here for, what are we doing

Existential Risk research has a clear and easy to understand purpose. However, the scope of this purpose is extremely broad and the challenges in achieving it are, as we demonstrate in this report, significant. Given these facts it can be very hard for individual researchers to easily understand how their work fits into the wider project of mitigating global catastrophic risk and to effectively collaborate in achieving this aim.

For instance, Karin Kuhlemann pointed out that existential risk research often focuses on things that are easier to imagine and form part of the social zeitgeist, because these fall into an area of relative simplicity where a single brilliant researcher can make a significant difference. Complex global issues on the other hand demand a more collective and resource intensive approach. As a global community this is probably something that those studying existential risk can achieve, however that will only work if we can both collaborate and maximise the size of the group who work with us.

Achieving both of these ends will require a focus on diversity and on equity amongst researchers. It is important that the resources of elite centres of academic learning are being used to genuinely advance the interests of all, and not simply to entrench existing privileges. This also matters, because as Peter Ho pointed out, the very notion of risk is more of a social construct than a mathematical concept, and when something gets interpreted within the language of risk then it can have profound effects on ordinary people and policy professionals alike.

Kristel Fourie pointed out that we can't take the position that we know what is right and best and we just need to get other people to agree with us, this just doesn't work. Participation can be a far more effective means of generating change. Alex Freeman drew very much the same conclusion, arguing that communication strategies cannot be formed in a vacuum, they must be produced with a clear understanding of who you want to communicate with, why you want to communicate with them and how you want to do this communication. Conceptualizing something as a risk means paying a risk premium to manage it, and researches should not be dismissive about the costs of imposing this onto others.

A final existential question for the field of existential risk research is how we conceptualize what we are doing in scientific terms. Due to the epistemic challenges of studying global catastrophic risk there is a need for considerably scientific creativity in doing so. This in turn requires speculation and pushing the boundary. Seth Baum asked participants whether we are here to give quantifications of existential risk or more general ideas and narratives about it? Andrew Maynard went on to make this point more forcefully by pointing out that if our research is to have value then it needs to produce concrete products that people can actually use, and that in many cases cultural change may be more useful than empirical results. In his presentation, Simon Beard articulated the range of methodological approaches that can be used to contribute to either of these goals. However, he went on, unfortunately at present these methods are often poorly

implemented by existential risk researchers and most don't estimate their uncertainty or make clear their reasoning. It may be that in order to increase its credibility as a young science existential risk research needs to focus more on the accountability of its methods than the objectivity of its results.

Theme 3: Roulette Wheels, Fruit Machines and Risk Ladders

This leads onto another key theme for existential risk researchers, how to make our research truly relevant to the scale of the problems that we face. If moral theories about the badness of human extinction such as those discussed by Matthew Rendall are to be believed then the work being undertaken by researchers who may be able to mitigate global catastrophic risk is extremely valuable and important. However, we often fail to live up to this when dealing with the promotion and implementation of our research findings.

For instance, it is not uncommon to find researchers holding back from intervening in policy discussions because the precise level of risk in their area is yet to be quantified, or potentially even unquantifiable at this point. However, Seth Baum pointed out that while some policy makers require quantification of risks, not all do. When policymakers do require quantification then rather than caving into their desires it may be more effective to argue that they are mistaken and should change their views, because this can interfere with other desiderata and quantification is hard and methods for achieving it are often abused, such as not inviting the appropriate 'experts' to a structured expert elicitation exercise. Diana Coyle gave a particularly compelling case for this approach, arguing that measures of GDP don't really capture what people care about, but nevertheless persist because they match what policy makers are looking for.

Other approaches to making research outputs relevant to a wider audience were suggested by other presenters. Kristel Fourie pointed out that the history of disaster risk management shows that just giving people information is not effective at altering behaviour, and that what is needed is participatory risk communication. She had helped to develop interactive, compelling and informative techniques for risk communication like a 'disaster fruit machine'. Alex Freeman made similar suggestions including 'the spinning ring of doom' where concentric wheels are used to illustrate how conditions can combine to produce higher levels of risk than any individual threat on its own. This allows for a mechanism of communication that can both be calibrated according to scientific models and engaged with in a qualitative manner by a very broad audience.

Theme 4: People care about different things, this is a problem that is not easy to solve

During its nascent stages of development, the Global Catastrophic Risk community has been relatively homogeneous and this can lead to a false sense of certainty about the appropriateness of certain tools and perspectives. For instance, consequentialist ethics, Bayesian epistemology and transhumanism (or at least techno-optimism) are far more prevalent amongst the community as a whole than amongst the general public, or even other research communities with which we engage. This can lead us to ignore many potentially important differences and disagreements, limiting our effectiveness at engaging with wider communities and making it harder for us to access the information and perspectives that they contain, and which could help us to improve our work.

In the field of evaluation, this issue was brought out across the competing perspectives of philosophy, economics and law. For instance, Matthew Rendall highlighted in his talk how our view about the far future is shaped by whether we take a consequentialist (outcome based), a contractualist (agreement based) or a deontological (duty based) perspective to ethics, and suggested ways in which this difference might be addressed by further fundamental research into moral philosophy. The talks by Diana Coyle and Silja Voeneky then highlighted how these differences can be further compounded when our evaluation of the future is more than merely theoretical, but is bounded by the need for practical actions or the need to operate within existing institutions.

Opinion surveys suggest that the general public is more or less evenly split between the three major traditions of contemporary moral philosophy, in so far as they engage with them at all, and that if anything it is often consequentialism that is the least popular. Efforts to shift this balance are likely to be at best long term hopes. A failure to consider non-consequentialist and practice based evaluation tools could therefore significantly undermine the force of our arguments for prioritizing these risks.

This clearly connects with another key theme of this conference, the challenge of effective communication. As Seth Baum stressed in his presentation, it is important to know the audience you are communicating with, and to know the audience that you want to communicate with. Apart from evaluative differences, Seth pointed to differences in people's desire to see risks being quantified. Some audiences require that risks are quantified because this makes it easier for them to apply existing tools, such as integrated assessment models and cost benefits analysis to them. Others, however, find it hard to engage with numbers and prefer a more qualitative approach to risk communication. For instance, using narratives and scenarios to make risks, and the path to mitigating them, more vivid. This implies that we may need to apply not only different tools of risk communication, but also different tools of risk analysis.

Alex Freeman made a similar point when he pointed out that the media convey catastrophic risks by using capital letters and stressing that a catastrophe could

happen tomorrow; insurers emphasise the level of uncertainty, whilst trying to quantify this very precisely; and health professionals describe all risks as simply 'acceptable', 'unacceptable' or 'intolerable'. Alex argued that we should differentiate between getting people to care about things that they do not currently care about and providing information to people who already care about these things. Finding a research methodology that can inform all of these practices in a way that is coherent and rigorous will be very hard and Simon Beard explored a wide range of approaches that have thus far been taken by the community, from toy modelling to structured expert elicitation, although many of these have remained confined to a relatively small area of study. It is not unreasonable to invoke different standards of evidence for these two goals. However, perhaps we should only do so if we can be confident that there will not be epistemic leakage of information that has been produced for one context into another context, for which it is not suitable.

Finally, it is clear that the scope of risks and threats that we consider will also be determined, in part, by the audience that we are speaking to, their level of awareness, their normative beliefs and their degree of commitment to mitigating global catastrophic risk. The broader we want to cast the net of our research, the more flexibility we will need to show regarding all of these things. To this end, Andrew Maynard pointed out that Innovation, in any field, is about providing people with something that they are willing to pay for (in terms of time and attention, even if not financially). In order to be innovative in our field, and to expand our scope of research, he therefore stressed the importance of finding some new product, tool or practice that demonstrably protects or grows value for people, and that they will thus want.

Theme 5: The Risk of Studying Existential Risk

Studying and communicating existential risk is a risky business. While our intentions are good, there are also unintended, and even perverse, consequences to this work. We need to recognize these and ensure that we do more good than harm. Some potential dangers from communicating existential risks are well known. These include: information hazards, true information that may cause harm or enable others to cause harm; confusing or distracting people from more important work; creating a false sense of certainty, or uncertainty; and releasing potentially inaccurate information into the world that may nevertheless be hard to counteract in future once it has become well established.

As was mentioned previously, communicating existential risk to a sensationalist media is problematic. Tamsin Edwards observed that different media outlets have used competing studies on sea-level rise to fit their political agenda. Predictions below the IPCC projected sea-level maximum in 2100 were picked up by conservative branches to underline the message that there is no scientific consensus. On the other hand, progressive media latched onto studies that forecast above the maximum sea-level rise estimate to suggest that the IPCC was underplaying the true menace of ice melt. Most didn't understand or note the important differences in evidence and methods used. Communicating with media can be troublesome, particularly when motives differ. Being aware of how research can be misconstrued to match different political purposes, and transparent about uncertainty and the approach used, can help to ensure that our ideas are portrayed accurately and fairly.

The framing of risk can also directly expose some to harm. This can happen even when bypassing the media and speaking directly to the public. Alex Freeman highlighted the case of the 2009 L'Aquila earthquake in which a group of seismologists were convicted by the Italian government on the basis of poor risk communication. The experts provided the right absolute numbers (0.1% risk) to the citizens of L'Aquila. However, they failed to convey the relative risk: that the chance of a catastrophic earthquake was 100 times the normal background rate. This small slight in framing may have contributed to how people reacted to this warning, and hence to the number of people killed by the ensuing earthquake. The lessons for scholars of existential risk are clear: be transparent about the level of relative risk and uncertainty and frame risks in a way that will encourage appropriate action, rather than hazardous behaviour.

The tools we use to evaluate existential risk also run the peril of locking-in poor practices with unintended consequences. Diane Coyle critiqued the measurement of financial services to the economy. In the 1990s the metric of "Financial Intermediation Services Indirectly Measured" was adopted by many countries. While seemingly innocuous, it has served to systematically mask volatility in the economy and give a false illusion of consistent growth. Metrics and measurements are neither neutral nor without impact on the outside world. Having forethought about how the tools we use could lead to suboptimal or reckless practices helps to ensure that the field of existential risks doesn't fall into the same mistakes as that of the dismal science: economics. Unfortunately, as Simon Beard noted in his

presentation, there is some evidence that it already has with both bad practices, and poor quality results, proliferating around the community as a result of being adopted by researchers early on in the fields development or becoming associated with some of its biggest names.

It is our hope that the emerging community of existential risk scholarship will find ways of doing better in future, and that we will find ways to overcome these challenges.

Summaries of each Panel

Panel 1: Challenges of Evaluation

Diane Coyle – What to watch out for: an economists perspective

It can be very hard for people to predict catastrophic risks when they observe very small changes in their lives. For instance, prior to the 2008 financial crisis, the economic story was generally positive. There were some noticeable economic ups and downs, but generally everything looked very steady and stable in the early 2000s and the trends that indicated an impending catastrophe were hard to notice.

One of the problems was that the measurement of the economic value of the financial sector was problematic, the contribution of the financial sector to the economy as a whole was being systematically overstated and the level of volatility of this value was understated. But, how do you know what you should be looking at?

Measuring this is difficult. Adam Smith contended that it shouldn't be counted at all. Eventually during the 1990s many countries adopted the metric of "Financial Intermediation Services Indirectly Measured". The metric gave the illusion that the economy was consistently growing and that economic volatility had been eliminated.

This raises the key question of what should you look at and measure? What is measured is usually acted upon. The use of metrics in economics is rarely logical. The use of yield curves can be telling as a measurement of economic stability but is rarely used. Metrics, and the choice of measurement tools, are even more contentious in the context of information overload and if you are not an expert in the area.

Economists want to know what implications something has for your life and your decisions in the here and now. This can lead to both self-fulfilling and self-averting risks, and so can make it very difficult to evaluate the data you have available to you. For instance, recessions and growth can be self-fulfilling or self-averting prophecies. Measurement tools have a clear, discernible impact on the world they are trying to measure. This can make it difficult to evaluate data and the effect of interventions.

While the 'millennium bug' didn't end in catastrophe, it was likely due to months of work by computer scientists and IT specialists. Yet, public figures outside of the field tend to dominate the narrative of how Y2K was a false risk. What systems of representation can best be used for scientific data and publications that would most directly relate this data to the things that people should be most concerned about?

We tend to deceive ourselves because we are reluctant to abandon our conventional lives and ways of thinking.

Matthew Rendall – Discounting, Aggregation and the Ecological Fallacy

If we don't destroy the world, then a whole lot of people (and hence a whole lot of utility) could exist in the future. This should be seen as a very good thing.

However, economists objected to the Stern review on climate change on the grounds that Stern overestimated the value of this future wellbeing by rejecting a pure time welfare preference and hence giving too much weight to things that will happen in the future.

William Nordhouse suggested that if we saw the world in the way that Stern suggests then a very small reduction in perpetual GDP in the far future (when people will be much better off than we are) would be worth sacrificing more than half of our current consumption levels. This, he

holds, is a reduction-ad-absurdum of failing to apply a pure time discount rate.

Martin Weitzman argued that we cannot rule out extreme climate change and that, far from being a wrinkle in future consumption (because it is so unlikely) this could be utterly catastrophic (because it's impact would be so great). If we don't give these future consequences such weight, because we discount them away due simply to when they occur, then we seem to get an equally ridiculous result.

If discount rates are set too low, then the interests of the present are swamped by the interest of the future, if we set it too high then the interests of the future get no say whatsoever. This seems to be a specific example of a generalized set of cases that were of great concern to philosophers like John Rawls and Tim Scanlon. In these cases, huge sacrifices to a small number are set against small benefits to a very large number.

Tim Scanlon has attempted to resolve such cases by appealing to the principle that even a minority can always reasonably reject some great cost, if it would be sufficiently large, no matter how much larger the total collective value to the majority might be of receiving less benefits.

However, there are cases in which our society clearly does focus on smaller benefits to larger numbers instead of giving much greater benefits to many fewer people. For instance, we will provide relatively trivial pain relief to people with headaches and muscle pains even though these may take resources away from research into deadly, but neglected, diseases.

Perhaps there are two kinds of criteria that we can apply. Parfit distinguished between Scanlonian and Kantian contractualism. Kant was concerned about what we should rationally accept and Scanlon was concerned about what people could reasonable reject. He argued that we could rationally accept a principle that was bad for us, so long as it was collectively the best thing to do. Because we can all rationally accept such a principle then we can establish collective institutions (such as the NHS) which embody these principles. However, we can also reasonably reject principles that are very costly to us regardless of their collective benefits, this is what is owed to us, and is a better way to conceive of personal morality, such as charitable giving. However, we can only reasonably reject principles under certain conditions, for instance when the people effected are better off than we are, but not when they might be worse off.

Rational acceptance implies accepting that we have reasons to accept a principle even if it may not benefit us because we must accept that other people could provide us with reasons to accept this principle. Reasonably reject is the opposite, we can present others with reasons not to accept a principle because of its effect on us, even though they would benefit from its acceptance. Rationality is about accepting what we ourselves have reason to do, reasonability is about offering reasons to others that they might accept as a justification for accepting a principle.

The ecological fallacy 'a statistic that describes and aggregate group should not be treated as if it applied to its individual members'. Thomas Schelling discussed the importance of inequality when determining the social discount rate on the grounds that its justification implied just such a fallacy. He argued that whilst people in the future may be on average richer than we are, some people will be poorer. This should also apply across different possible worlds (worlds in which there was catastrophic climate change will be worse off, and hence morally more important, than those in which there was less severe climate change).

The morally acceptable method of discounting would be to disaggregate different possible future scenarios. Benefits that accrue in scenarios, where people are a lot better off than we, are can simply be struck out of our moral considerations as irrelevant. However, those that accrue to people who are on a par with us or who are worse off than we are should be handled separately according to their strength whilst also taking into account the subjective probability of their likelihood.

The sheer size of these costs and benefits is likely to be so large that it will mandate very

high levels of precaution for catastrophic risks, even for very low probability risks.

Silja Voeneky - The Public International Law Perspective on Evaluating Existential Risks

Extreme Technological Risks are usually transnational risks whose potential damage cannot be limited to any one state, including the state in which they originated. It would therefore be desirable if there was a coherent body of international law to manage these risks.

Yet there is currently no such body of law, with the possible exception of the UN charter itself. Different areas of international law (environmental law, laws of war, human rights, trade law etc.) are tailored to deal with individual problems and with particular technological risks, but there is little attention to how these might turn into existential risks, or to the problem of global catastrophic risk in general.

However, there are different normative tools available in international law for dealing with how risks can be evaluated and diminished. These were shaped by the contexts in which each area of law developed. Sadly, these contexts were often those of a post catastrophe desire to avoid the mistakes of the past. These include provisions for the legal use of force within the UN charter and the various environmental treaties that have been negotiated to mitigate risks from climate change and similar risks.

One recent example of this process in action are the ongoing negotiations amongst 82 states on the limitation of Lethal Autonomous Weapons, although this only extends to conventional weapons and only 26 states so far in favour of whole ban of their use. Another is France's initiative for an international global pact for the environment (though soft-law, very broad, heavily criticised and not promising)

We need to approach this problem with a healthy scepticism about the possibility of achieving international agreement.

An example of where the law might develop in the future is in the regulation of gene drives - CBD and Cartagena Protocol on living modified organisms applicable to gene drives, espousing an international version of the precautionary principle to prevent transboundary movement, but no ban, and relevant state actors such as US not signed or ratified; same applies to Nagoya Protocol and Kuala Lumpur Protocol. BWC also applicable, but does not cover peaceful use and no implementation regime.

The strengths and weaknesses of international law as an evaluative mechanism are that, whilst they are not comprehensive, there are strong and binding principles that can be used to establish a risk assessment process. Sometimes these occur in a global enforceable regime. They can also, sometimes, take account of non-anthropocentric goods and values. There is a well-established principle of precaution against catastrophic risk and customary law regarding the attribution of responsibility for risk imposition. Finally, it is possible to negotiate new norms today in order to govern emerging risks in the future, such as AI, with the development of soft law declarations being a good place to start with this.

Panel 2: Challenges of Evidence

Seth Baum - Making the Most of Limited Evidence in Global Catastrophic Risk Analysis

There are two camps when it comes to risk quantification: those who want it and those who don't. Some believe that risks can and should be quantified, and that decision-makers require such information. Other think that you don't need to quantify them and the risks are often too uncertain and complex.

The problem of near misses. For example, in the case of nuclear weapons we have only two cases of nuclear weapons being used in an attack. Both were during wartime. However, we have a multitude of incidences of 'near misses', such as the Cuban missile crisis, of varying severity.

We can make risk estimates even with scarce historical data e.g. nuclear war. However, we need to be appropriately sceptical of these numbers and communicate this scepticism.

In the case of scant evidence, we can instead turn to best estimates from experts. Granger Morgan said: "Some may find it tempting to view expert elicitation as a low-cost, low-effort alternative to doing serious research and analysis. It is neither."

However, sometimes the expertise we want doesn't exist e.g. forecasting future technologies like AI. Computer scientists may have some ideas, but they not map onto the actual development of the technology.

In other cases, we may be doing too much quantification, and often decision-makers don't need numbers. Many decision-makers don't follow an ideal, rational approach. Does the chance and severity of Nuclear Winter change the approach of decision-makers? It more likely simply underlines the rationale of nuclear deterrence.

Similarly, in climate change, accounting for fat tails would lead to a much higher carbon price, but most countries don't even have a carbon price, let alone one that reflects the social cost of carbon. The underlying issues are ones of political implementation, not expected value.

One approach to dealing with this is to map out decision-makers on key risks and their perceptions and risks. This can help align our research and make it more useful in being practical, listened to and actually reducing risk.

Simon Beard - Probabilities, methodologies and the evidence base in existential risk assessments

There are a surprisingly diverse range of methodologies currently being used to assess the level of existential risk, both in general terms and in relation to specific threats and hazard. This paper drew on a literature review of these methods and highlighted how they attempted to overcome the challenges inherent in this field.

Apart from purely analytical approaches such as the 'doomsday argument', methods can be placed into one of four categories, use of modelling and historical data, individual subjective opinion, group surveys and structured expert elicitation.

The paper uses an informal evaluative framework to consider the relative merits of these methodologies in terms of their rigor, ability to handle uncertainty, accessibility for researchers with limited resources and utility for communication and policy purposes.

Whilst they are popular and carry a strong scientific track record, traditional modelling techniques and the use of historical data can often be misleading, or at least give a false sense of security, when applied to unprecedented phenomena such as existential risk.

Subjective opinions are probably overused amongst existential risk scholars, however they can be greatly by applying rigorous techniques for Bayesian reasoning, although these are often ignored.

Group opinion surveys can both widen the scope of information and perspectives that are included and also hide instances of faulty reasoning and bad epistemic practice.

Structured Expert Elicitation holds opportunities for combining all the benefits of other methods, however it has so far been little used, potentially because it is costly and hard to implement.

Whilst there is no unique best way to estimate existential risk, different methods have their own merits and challenges, suggesting that some may be more suited to particular research purposes than others. More importantly however, we conclude that, in many cases, estimates based on poor implementations of a method exist alongside far better ones, but it is often the case that the community still frequently invokes them.

Tamsin Edwards - How soon will the ice apocalypse come?

Different stories have arisen around sea-level rise and the chance of tipping points such as the collapse of the Greenland Ice Sheet. Some of these are of profound importance. For instance, the collapse of the West Antarctic Ice Shelf (WAIS) could produce a runaway affect (Marine Ice sheet Instability).

James Hansen suggested that we could experience 4-5 metres of sea-level rise by the end of the century. This was based on local evidence from the Global Pulse 1A event. However, this occurred at a time when there was more ice in the world and we were coming out of the glacial age. Problematically, these past periods didn't just have different amounts of ice, but also different amounts of radiative forcing. 4-5 metres of sea-level rise was then updated to 2.5 metres (UK government) while others suggest 1.1 metres.

Some scientists then turned to the more recent past for extrapolation. Expert elicitation was then used by the IPCC to help determine sea-level rises based on the plethora of evidence sources. A classical approach to structured expert elicitation (the Cook method) was used, which drew on sea-level rise experts rather than ice sheet experts, who estimated a more dire level of sea rise (1.5 metres rather than 1).

New models using Bayesian calibration based on the recent past and ensembles came to a more conservative estimate. DeConto and Pollard then threw a spanner in the works by proposing a new hypothesis of Marine Ice Cliff Instability (a domino effect whereby a sheet collapse leaves a cliff which collapses more quickly). They came to 1m sea level rise by 2100 just from Antarctica (2m or more overall), and potentially 16m by 2500. This was then redone using a skewed rather than Gaussian distribution and came to a more conservative estimate again.

Evidence for fat tails is problematic even in more developed fields (climatology and sea-level rise) due to the range of methods and sources that can be used, even for structured expert elicitation. In each case, different estimates of sea-level rise were seized by different factions of the media to either stress the danger, or lack of consensus, on sea-level rise.

Panel 3: Challenges of Scope

Karin Kuhlemann – Sexy Vs Unsexy Catastrophic Risks

Some existential risks are sexy and some are unsexy. However, the challenges that face us from these two kinds of risks are very similar.

Sexy risks include asteroids, nuclear war, pandemics and AI. These are characterised by their epistemic neatness (easy to define) crystalize suddenly (these threats are either 'on' or 'off') and involvement of technological risk (either as a cause or a solution) which is itself appealing.

Unsexy risks include climate change, soil erosion, biodiversity loss and social collapse like mass unemployment. Even though there are many people working on these risks and they are starting to have a significant impact on our day to day lives, there is a lot of denial and unwillingness to engage with what is going on. These are all epistemically messy in that they are complex and hard to study and require significant expertise to understand. They are also slow moving and emerge over long periods of time (over time the size of fish caught by sports fisherman has declined, but their sense of satisfaction with the fish they are catching has remained the same because the decline is too slow for them to be aware of it. These risks are also not amenable to technological fixes because they relate to the interactions between individual behaviours brought about by the interconnectivity of our global systems. In fact, a lot of these risks are related to one another and they are all clustered around the problems of overpopulation.

Another feature of unsexy risks that makes them so pernicious is that, due to their complexity, it is easy for people who don't appreciate global systems as a whole (which nobody does yet) to do things that are actively harmful. For instance, the 1994 Cairo Consensus decreed that people should not talk about

population growth in ethical terms, but only about sexual and reproductive health rights. This was based upon the good intentions of circumventing political barriers and achieving consensus between different political, religious and ethical groups. However, by ignoring the issue it effectively allowed a narrative to develop that lead people to talk about consumption as being unrelated to the number of people consuming. This is clearly not true. For instance, one can roll down consumption during an emergency, but not population. Population change is a very long term change. It also misses out the fact that many people who are consuming are doing so in level of poverty that are unacceptably low and there is clearly an ethical imperative to help these people.

Overpopulation is not an end point (a mad max world where people literally cannot move) it is a trajectory of increasing unsexy risks like climate change, and social collapse. Even if we still have resources left to extract, a failure to take account of these unsexy risks, and especially the risk of overpopulation, means committing yourself to a long term trajectory that is inevitably driving you towards an end result that is morally unacceptable to you. There are good reasons to think that we are already on such a trajectory and that the timescale until which we reach this unacceptable end point is within the lifetime of many of those who are currently living, which should be well within most people's moral time horizon, i.e. the period over which we cannot ignore the consequences of our actions.

Humanity has an apocalyptic blindness. We do not want to believe that our own story will come to a bad end. Thus, when we see clearly unsustainable trajectories in our relationship with the earth system then, even though we would conclude that such trends would inevitably lead to disaster for plants and animals we chose to overlook the fact that they may lead to disaster for our own species as well.

Andrew Maynard - Risk Innovation: The roles of creativity, imagination and rigour in exploring existential risk

We need to get out of the rut of thinking conventionally about risks, in ways that were broadly developed to think about chemicals and radiation, i.e. Hazard x exposure x dose response = probability of harm.

This works well for short lived and acute toxins for example, bust starts to fall apart when we start thinking about even long-lasting or environmentally damaging chemicals, and especially when we start applying that mindset to things such as AI or genetics.

Two examples of this are 1) A sort of Alexa / google dot PA that has been installed in all dorm rooms at a university (a great idea that gets complicated when we think about confidential things being discussed/ recorded) and 2) driverless cars.

These don't fit in with our conventional ideas about risk and so our frameworks don't work. Risks have typically been identified that a) we can conceive of and b) that we can cope with. Anything that falls out of this narrow band is ignored, or we try to force it into the narrow band somehow. This gets worse with high severity, high uncertainty, low probability risks like the ones that could bring about a global catastrophe.

We should deal with this problem by addressing risk with innovation. Innovation is defined as "translating an idea into a good or service that creates value for which customers will PAY for". We can move this idea to risk as follows "a new way of approaching risk that leads to new knowledge, understanding and capabilities and translates these into products tools or practices that protect and grow societal, environmental, economic or other value." The message is that it needs to create tangible, valuable outputs and we need to think of risk in this way, with a focus on creativity.

We can also use sci-fi as a tool to inform and think differently, for instance Andrew uses ex-Machina and Plato's cave, his point being that AI would have a totally different conception of reality (like philosopher in the cave analogy). The question is though, is AI going to add value to the humans in the cave, or is it going to manipulate and enslave them?

For example, in ex-Machina, - Caleb thinks he's the philosopher conversing with Eva in her cave - Eva starts manipulation of Caleb - She knows what buttons to press and levers to pull to get him to do what she wants and she leaves the 'cave' as an enlightened being.

So, sci-fi is a great way to get an idea about how to think about the future, and particularly things like AI that could affect humans, who either would be unable to perceive the manipulation, or would be unable to resist it. I.e. the machines are able to understand, but are not subject to the heuristics of humans.

Previously, we have anticipated that if we design machines to be 'like' us then they will want the same objectives as us, but it is impossible to program something to experience the world the same way as we do. This makes machines essentially unknowable and they may be seeing the world in a way that is completely inaccessible to us. This isn't about intelligence, but more about how the machines are outside of our reference.

Panel 4: Challenges of Communication

Kristel Fourie - Understanding Risk for effective communication

The typical idea of disaster risk is conceptualised as: hazard x vulnerability / capacity. Hence the aim is to mitigate hazards, reduce vulnerability and build capacity. Viewing risk in this way implies that there are no natural disasters. Vulnerability and capacity are within our control and therefore all disasters (at least in terms of their severity) are in some way caused by humans.

Making risk newsworthy, the transdisciplinary nature of disaster risk research, and the incentives of research (incentives and disincentives) are all major barriers to disaster risk reduction. Achieving the aim of communicating disaster risk and reducing it needs to be kept in mind. Unfortunately, we have not been successful in this endeavour. We have struggled in actually catalysing action from major actors.

There are two models of the problem we face. The first is the knowledge-deficit model, according to which if you give people information, make sure to repeat it until they understand it then voila, the problem will be solved. According the second, transmission of information model, if people have information they will change their behaviour and decisions accordingly.

Participatory development communication has been found to be the preferred approach across both disaster risk and communications research. Experts need to understand the communities they are trying to help and tailor their messages accordingly. We need to 1. Understand how people perceive risk by using participation; and 2. To understand the dimensions of risk.

Understanding the risk perception of different individuals and communities is key. For example, those that are present and past focused are less likely to see climate change as a major risk. Disaster risk reduction tends to focus on a timescale of 10-15 years, climate change over decades and existential over centuries of millennia.

One suggestion for improved communication would be via the medium of TV: a Netflix for risk communication aimed at an international audience of NGOs, policy-makers, academics and students.

Alex Freeman - Communicating (Existential) Risks Responsibly

The media has been working on existential risk for ages. They convey the scale of a risk by using capitals and saying that the end is tomorrow. Insurance has also been working on risks for ages and they communicate in a much more opaque way. It tends to be back ended in long reports in a risk matrix which stresses model uncertainty but provides exact estimates down to multiple decimal places. They also tend to underplay risk by framing fat-tail events as 1/100 years. Finally, health professionals tend to communicate risk as simply being

'acceptable', 'unacceptable', 'intolerable' and so on.

So there are different ways of communicating risk: it can be all about the emotion (making you care), or all about the numbers (if you already care and just want the information). Both of these approaches can work depending on the audience and purpose. In the case of daily risks (car crashes) risks are salient with a clear cause and effect. That is not the case with existential risks. So we need to be aware of both the benefits and the costs associated with mitigating risks.

An example of where this went terribly wrong was the 2009 L'Aquila earthquake where seismologists were convicted of criminal offences based on their poor risk communication. The absolute numbers they communicated were right (0.1% risk), but not the relative context (the earthquake risk was 100 times the normal background rate). Another example of bad practice are increasingly prevalent health warnings on products that imply a miniscule risk of death or harm but give no indication of their relative safety (e.g. coffee in California). Using risk ladders, employed by anaesthetists can help to explain low probabilities by ranking different risks.

25 1-in-500-year storms have occurred the US since 2010. The problem is the communication of probabilities. Winning the lottery may be in a 1-in-40-million yet someone wins it every year. Another approach to overcoming this is the 'Spinning wheel of doom' as a tangible way to communicate a model on storm surge for Greater Yarmouth. Concentric wheels representing different risk factors are marked red (high risk) (yellow (medium risk) and green (low risk). People are then asked to spin the wheel to see if a disaster might occur or not. This can illustrate the impact of different conditions by changing the length of the lcoloured lines on each wheel and people see how multiple failures can easily combine to lead to what is intuitively an unlikely scenario., but ultimately it comes down to spin of the wheel and a good dose of randomness.

Peter Ho – Black Elephants

In 1859 a large solar flare (the Carrington Event) created a significant geomagnetic storm. IT was uneventful then, causing dazzling auroras, but little else. Now, it could be catastrophic. It would wipe out GPS, power grids and most other electrical systems. The first-order effects would be devastating, but the second and third order effects could be far worse.

The availability heuristic is pertinent for risk management. After 9/11, air traffic decreased while road traffic increased, despite the latter being far more dangerous. Professor Gerd Gigerenzer estimated that an extra 1,595 Americans died in road accidents in the year after the attack. They were indirect victims of 9/11; victims of the availability heuristic.

Similarly, hubris in the financial world led to the Global Financial Crisis. Our reluctance to deal with fat-tail risks often stems from cognitive dissonance. An inability to acknowledge uncertainty and accept a radically different future possibilities, particularly when they run contrary to the interests of current powers. This leads to procrastination until a crisis occurs and emergency management is required.

The concept of a Black Elephant is a cross between a Black Swan and the Elephant in the Room. A well-known issue which we would prefer to avoid. Once it evolves into a crisis we then feign surprise. Catastrophic floods in Thailand were an example of this.

Big risks are being amplified by interconnectivity e.g. Thai floods leading to the disruption of global supply chains and knock-on economic effects. These risks are not the province of actuaries. They are a wider social construct, requiring agreement from multiple segments of society to pay a social risk premium if we want to address them. The Netherlands and their use of dykes to prevent flooding is an example of this.

Building consensus around risks requires public trust in communication. Unfortunately, this is at an all-time low!

Centre for the Study of Existential Risk Workshop Report Series

Reports can be accessed online through www.cser.ac.uk

Suggested citation for this report: Simon Beard, Catherine Rhodes, et al., 2019, Conference Report: the Cambridge conference on catastrophic risk 2018, Centre for the Study of Existential Risk (Cambridge)