# Defence in Depth Against Human Extinction: Prevention, Response, Resilience, and Why They All Matter

Owen Cotton-Barratt, Max Daniel and Anders Sandberg
*University of Oxford*

Research Article

## Abstract

We look at classifying extinction risks in three different ways, which affect how we can intervene to reduce risk. First, how does it start causing damage? Second, how does it reach the scale of a global catastrophe? Third, how does it reach everyone? In all of these three phases there is a defence layer that blocks most risks: First, we can prevent catastrophes from occurring. Second, we can respond to catastrophes before they reach a global scale. Third, humanity is resilient against extinction even in the face of global catastrophes. The largest probability of extinction is posed when all of these defences are weak, that is, by risks we are unlikely to prevent, unlikely to successfully respond to, and unlikely to be resilient against. We find that it's usually best to invest significantly into strengthening all three defence layers. We also suggest ways to do so tailored to the classes of risk we identify. Lastly, we discuss the importance of underlying risk factors – events or structural conditions that may weaken the defence layers even without posing a risk of immediate extinction themselves.

## Policy implications

- We can usually best reduce extinction risk by splitting our budget between all defence layers.
- We should include measures that reduce whole classes of risks, such as research uncovering currently unseen risk. We should also address risk factors that would not cause extinction themselves but weaken our defences, for example, bad global governance.
- Future research should identify synergies between reducing extinction and other risks. For example, research on climate change adaptation and mitigation should assess how we can best preserve our ability to prevent, respond to, and be resilient against extinction risks.

## Our framework for discussing extinction risks

Human extinction would be a tragedy. For many moral views it would be far worse than merely the deaths entailed, because it would curtail our potential by wiping out all future generations and all value they could have produced (Bostrom, 2013; Parfit, 1984; Rees, 2003, 2018).

Human extinction is also possible, even this century. Both the total risk of extinction by 2100 and the probabilities of specific potential causes have been estimated using a variety of methods including trend extrapolation, mathematical modelling, and expert elicitation; see Rowe and Beard (2018) for a review, as well as Tonn and Stiefel (2013) for methodological recommendations. For example, Pamlin and Armstrong (2015) give probabilities between 0.00003% and 5% for different scenarios that could eventually cause irreversible civilisational collapse.

To guide research and policymaking in these areas, it may be important to understand what kind of processes could lead to our premature extinction. People have considered and studied possibilities such as asteroid impacts (Matheny, 2007), nuclear war (Turco et al., 1983), and engineered pandemics (Millett and Snyder-Beattie, 2017). In this article we will consider three different ways of classifying such risks.

The motivating question behind the classifications we present is 'How might this affect policy towards these risks?' We proceed by identifying three phases in an extinction process at which people may intervene. For each phase, we ask how people could stop the process, because the different failure modes may be best addressed in different ways. For this reason we do not try to classify risks by the kind of natural process they represent, or which life support system they undermine (unlike e.g. Avin et al., 2018).

### Three broad defence layers against human extinction

An event causing human extinction would be unprecedented, so is likely to have some feature or combination of features that is without precedent in human history. Now, we see events with *some* unprecedented property all of the time – whether they are natural, accidental, or deliberate – and many of these will be bad for people. However, a large majority of those pose essentially zero risk of causing our extinction.

Why is it that some damaging processes pose risks of extinction, but many do not? By understanding the key differences we may be better placed to identify new risks and to form risk management strategies that attack their causes as well as other factors behind their destructive potential.

We suggest that much of the difference can usefully be explained by three broad defence layers (Figure 1):

1. First layer: prevention. Processes – natural or human – which help people are liable to be recognised and scaled up (barring defeaters such as coordination problems). In contrast processes which harm people tend to be avoided and dissuaded. In order to be bad for significant numbers of people, a process must either require minimal assistance from people, or otherwise bypass this avoidance mechanism.
2. Second layer: response.[1] If a process is recognised to be causing great harm (and perhaps pose a risk of extinction), people may cooperate to reduce or mitigate its impact. In order to cause large global damage, it must impede this response, or have enough momentum that there is nothing people can do.
3. Third layer: resilience. People are scattered widely over the planet. Some are isolated from external contact for months at a time, or have several years' worth of stored food. Even if a process manages to kill most of humanity, a surviving few might be able to rebuild. In order to cause human extinction, a catastrophe must kill everybody, or prevent a long-term recovery.

The boundaries between these different types of risk-reducing activity aren't crisp, and one activity may help at multiple stages. But it seems that often activities will help primarily at one stage. We characterise *prevention* as reducing the likelihood that catastrophe strikes at all; it is necessarily done in advance. We characterise *response* as reducing the likelihood that a catastrophe becomes a severe global catastrophe (at the level which might threaten the future of civilisation). This includes reducing the impact of the catastrophe after it is causing obvious and significant damage, but the response layer might also be bolstered by mitigation work which is done in advance. Finally, we characterise *resilience* as reducing the likelihood that a severe global catastrophe eventually causes human extinction.[2]

Successfully avoiding extinction could happen at each of these defence layers. In the rest of the article we explore two consequences of this.

First, we can classify damaging processes by the way in which we could stop them at the defence layers. In section 2, we'll look at a classification of risks by their origin: understanding different ways in which we could succeed at the prevention layer. In section 3, we'll look at the features which may allow us to block them at the response layer. In section 4, we'll classify risks by the way in which we could stop them from finishing everybody. We conclude each section by policy implications.

Each risk will thus belong to three classes – one per defence layer. For example, consider a terrorist group releasing an engineered virus that grows into a pandemic and eventually kills everyone. In our classification, we'll call this prospect a *malicious risk* with respect to its origin; a *cascading risk* with respect to its scaling mechanism of becoming a global catastrophe; and a *vector risk* in the last phase we've called endgame. We'll present more examples at the end of section 4 and in Table 1.

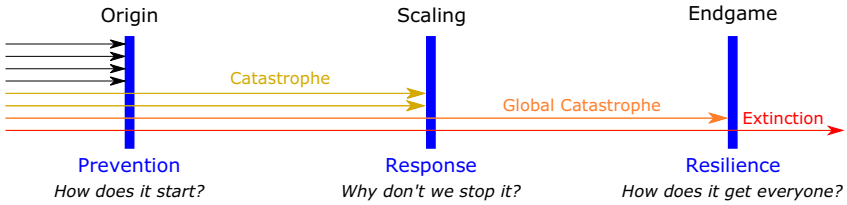
Second, we present implications of our framework distinguishing three layers. In section 5, we discuss how to allocate resources between the three defence layers, concluding that in most cases all of prevention, response, and resilience should receive substantial funding and attention. In section 6, we highlight that risk management, in addition to monitoring specific hazards, must protect its defence layers by fostering favourable structural conditions such as good global governance.

## Related work

Avin et al. (2018) have recently presented a classification of risks to the lives of a significant proportion of the human population. They classify such risks based on 'critical systems affected, global spread mechanism, and prevention and mitigation failure'. Our framework differs from theirs in two major ways. First, with extinction risks we focus on a more narrow type of risk. This allows us, in section 4, to discuss what might stop global catastrophes from causing extinction, a question specific to extinction risks. Second, even where the classifications cover the same temporal phase of a global catastrophe, they are motivated by different questions. Avin et al. attempt a comprehensive survey of the natural, technological, and social systems that may be affected by a disaster, for example listing 45 critical systems in their second section. By contrast, we ask why a risk might break through a defence layer, and look for answers that abstract away from the specific system affected. For instance, in section 2, we'll distinguish between unforeseen, expected but unintended, and intended harms.

We believe the two classifications complement each other well. Avin and colleagues' (2018) discussion of prevention and response failures is congenial to our section 6 on underlying risk factors. Their extensive catalogues of critical systems, spread mechanisms and prevention failures highlight the wide range of relevant scientific disciplines and stakeholders, and can help identify fault points relevant to particularly many risks. Conversely, we hope that our coarser typology can guide the search for additional critical systems and spread mechanisms. We believe that our classification also usefully highlights different ways of protecting the same systems. For example, the risks from natural and engineered pandemics might best be reduced by different policy levers even if both affected the same critical systems and spread by the same mechanisms. Lastly, our classification can help identify risk management strategies that would reduce whole clusters of risks. For example, restricting access to dangerous information may prevent many risks from malicious groups, irrespective of the critical system that would be targeted.

**Figure 1.** Three broad defence layers.




Our classification also overlaps with the one by Liu et al. (2018), for example when they distinguish intended from other vulnerabilities or emphasise the importance of resilience. While the classifications otherwise differ, we believe ours contributes to their goal to dig 'beyond hazards' and surface a variety of intervention points.

Both the risks discussed by Avin et al. (2018) and extinction risks by definition involve risks of a massive loss of lives. This sets them apart from other risks where the adverse outcome would also have global scale but could be limited to less severe damage such as economic losses. Such risks are being studied by a growing literature on 'global systemic risk' (Centeno et al., 2015). Rather than reviewing that literature here, we'll point out throughout the article where we believe it contains useful lessons for the study of extinction risks.

Finally, it's worth keeping in mind that extinction is not the only outcome that would permanently curtail humanity's potential; see Bostrom (2013) for other ways in which this could happen. A classification of these other *existential risks* is beyond the scope of this article, as is a more comprehensive survey of the large literature on global risks (e.g. Baum and Barrett, 2018; Baum and Handoh, 2014; Bostrom and Ćirković 2008; Posner, 2004).

## Classification by origin: types of prevention failures

Avoiding catastrophe altogether is the most desirable outcome. The origin of a risk determines how it passes through the prevention layer, and hence the kind of steps society can take to strengthen prevention (Figure 2).

### Natural risks

The simplest explanation for a risk to bypass our background prevention of harm-creating activities is if the origin is outside of human control: a *natural risk*. Examples include a large enough asteroid striking the earth, or a naturally occurring but particularly deadly pandemic.

We sometimes can take steps to avoid natural risks. For example, we may be able to develop methods for deflecting asteroids. Preventing natural risks generally requires proactive understanding and perhaps detection, for instance scanning for asteroids on earth-intersecting orbits. Such risks share important properties with anthropogenic risks, as any explanation for how they might materialise must include an explanation of why the human-controlled prevention layer failed.

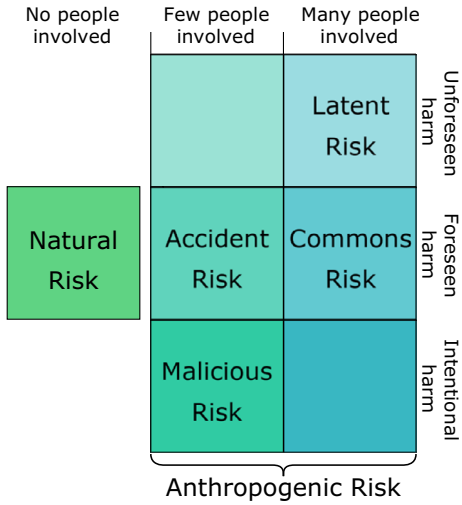
### Anthropogenic risks

All non-natural risks are in some sense *anthropogenic*, but we can classify them further. Some may have a localised origin, needing relatively small numbers of people to trigger them. Others require large-scale and widespread activity. In each case there are at least a couple of ways that it could get through the prevention layer.

Note that there is a spectrum in terms of the number of people who are needed to produce different risks, so the division between 'few people' and 'many people' is not crisp. We might think of the boundary as being around one hundred thousand or one million people, and things close to this boundary will have properties of both classes. However, it appears to us that for many of the plausible risks the number required is either much smaller (e.g., an individual or a cohesive group of people such as a company or military unit) or much larger than this (e.g., the population of a major power or even the whole world), so the qualitative distinction between 'few people' and 'many people' (and the different implications of these for responding) seems to us a useful one.

Also potentially relevant are the knowledge and intentions of the people conducting the risky activity. They may

**Figure 2.** Classification of risks by origin.



be ignorant of or aware of the possible harm; if the latter, they may or may not intend it.[3]

### Anthropogenic risks from small groups

The case of a risk where relatively few people are involved in triggering and they are unaware of the potential harm is an *unseen risk*.[4] This is likely to involve a new kind of activity; it is most plausible with the development of unprecedented technologies (GPP, 2015), such as perhaps advanced artificial intelligence (Bostrom, 2014), nanotechnology (Auplat, 2012, 2013; Umbrello and Baum, 2018), or high-energy physics experiments (Ord et al., 2010).

The case of a localised unintentional trigger which was foreseen as a possibility (and the dynamics somewhat understood) is an *accident risk*. This could include a nuclear war starting because of a fault in a system or human error, or the escape of an engineered pathogen from an experiment despite safety precautions.

If the harm was known and intended, we have a *malicious risk*. This is a scenario where a small group of people wants to do widespread damage;[5] see Torres (2016, 2018b) for a typology and examples. Malicious risks tend to be extreme forms of terrorism, where there is a threat which could cause global damage.

### Anthropogenic risks from large groups

Turning to scenarios where many people are involved, we ask why so many would pursue an activity which causes global damage. Perhaps they do not know about the damage. This is a *latent risk*. For them to remain ignorant for long enough, it is likely that the damage is caused in an indirect or delayed manner. We have seen latent risks realised before, but not ones that threatened extinction. For example, asbestos was used in a widespread manner before it was realised that it caused health problems. And it was many decades after we scaled up the burning of fossil fuels that we realised this contributed to climate change. If our climate turns out to be more sensitive than expected (Nordhaus, 2011; Wagner and Weitzman, 2015; Weitzman, 2009), and continued fossil fuel use triggers a truly catastrophic shift in climate, then this could be a latent risk today.

In some cases people may be aware of the damage and engage in the activity anyway. This failure to internalise negative externalities is typified by 'tragedy of the commons' scenarios, so we can call this a *commons risk*. For example, failure to act together to tackle global warming may be a commons risk (but lack of understanding of the dynamics causes a blur with latent risk). In general, commons risks require some coordination failure. They are therefore more likely if features of the risk inhibit coordination; see for example Barrett (2016) and Sandler (2016) for a game-theoretic analysis of such features.

Finally, there are cases where a large number of people engage in an activity to cause deliberate harm: *conflict risk*. This could include wars and genocides. Wars share some features with commons risk: there are solutions which are better for everybody but are not reached. In most conflicts, actors are intentionally causing harm, but only as an instrumental goal.

### Risk creators and risk reducers

In the above we classify risks according to who creates the risk and their state of knowledge. We have done this because if we want to prevent risk it will often be most effective to go to the source. But we could also ask who is in a position to take actions to avoid the risk. In many cases those creating it have most leverage, but in principle almost any actor could take steps to reduce the occurrence rate. If risk prevention is underprovided, this is likely to be a tragedy of the commons scenario, and share characteristics with commons risk.

From a moral and legal standpoint intentionality often matters. The possibility of being found culpable is an important incentive for avoiding risk-causing activities and part of risk management in most societies. If creating or hiding potential catastrophic risks is made more blameworthy, prevention will likely be more effective. Unfortunately it also often motivates concealment that can create or aggravate risk; see Chernov and Sornette (2015) for case studies of how this misincentive can weaken prevention and response. This shows the importance of making accountability effectively enforceable.

### Policy implications for preventing extinction risk

- To be able to prevent *natural risks,* we need research aimed at identifying potential hazards, understanding their dynamics, and eventually develop ways to reduce their rate of occurrence.
- To avoid *unseen* and *latent risks,* we can promote norms such as appropriate risk management principles at institutions that engage in plausibly risky activities; note that there is an extensive literature on rivalling risk management principles (e.g. Foster et al., 2000; O'Riordan and Cameron, 1994; Sandin, 1999; Sunstein, 2005; Wiener, 2011), especially in the face of catastrophic risks (Baum, 2015; Bostrom, 2013; Buchholz and Schymura, 2012; Sunstein, 2007, 2009; Tonn, 2009; Tonn and Stiefel, 2014) – advocating for any particular principle is beyond the scope of this article. See also Jebari (2015) for a discussion of how heuristics from engineering safety may help prevent unseen, latent, and accident risks. Regular horizon scanning may identify previously unknown risks, enabling us to develop targeted prevention measures. Organisations must be set up in such a way that warnings of newly discovered risks reach decision-makers (see Clarke and Eddy, 2017, for case studies where this failed).
- *Accidents* may be prevented by general safety norms that also help reduce unseen risk. In addition, building on our understanding of specific accident scenarios, we can design failsafe systems or follow operational routines that minimise accident risk. In some cases, we may want to

eschew an accident-prone technology altogether in favour of safer alternatives. Accident prevention may benefit from research on high reliability organisations (Roberts and Bea, 2001) and lessons learnt from historical accidents. Where effective prevention measures have been identified, it may be beneficial to codify them through norms and law at the national and international levels. Alternatively, if we can internalise the expected damages of accidents through mechanisms such as insurance, we can leverage market incentives.[6]

- Solving the coordination problems at the heart of *commons* and *conflict risks* is sometimes possible by fostering national or international cooperation, be it through building dedicated institutions or through establishing beneficial customs.[7] One idea is to give a stronger political voice to future generations (Jones et al., 2018; Tonn, 1991, 2018).
- Lastly, we can prevent *malicious risks* by combating extremism. Technical (Trask, 2017) as well as institutional (Lewis, 2018) innovations may help with governance challenges in this area, a survey of which is beyond the scope of this article.
- Note that our classification by origin is aimed at identifying policies that would – if successfully implemented – reduce a broad class of risks. Developing policy solutions is, however, just one step toward effective prevention. We must then also actually implement them – which may not happen due to, for example, free-riding incentives. Our classification does not speak to this implementation step. Avin et al. (2018) congenially address just this challenge in their classification of prevention and mitigation failures.

## Classification by scaling mechanism: types of response failure

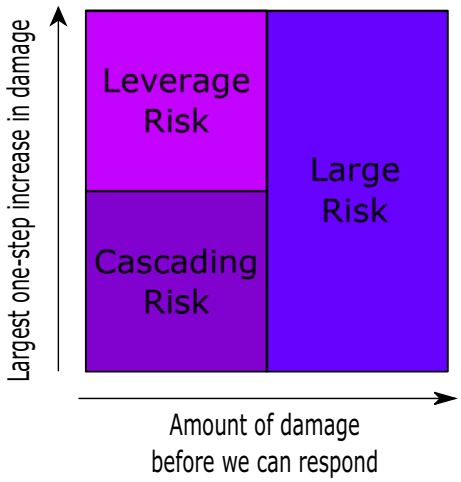
For a catastrophe to become a global catastrophe, it must eventually have large effects despite our response aimed at stopping it. To understand how this can happen, it's useful to look at the time when we could first react. Effects must then either already be large or scale up by a large factor afterwards (Figure 3).

If the initial effects are large, we will simply say that the risk is *large*. If not, we can look at the scaling process. If massive scaling happens in a small number of steps, we say there is *leverage* in play. If scaling in all steps is moderate, there must be quite a lot of such steps – in this case we say that the risk is *cascading*.

### Large risks

Paradigm examples of catastrophes of an immediately global scale are large sudden-onset natural disasters such as asteroid strikes. Since we cannot respond to them at a smaller-scale stage, mitigation measures we can take in advance (part of the second defence layer as they would reduce damage after it has started) and the other defence layers of prevention and resilience are particularly important to reduce such risks. Prevention and mitigation may benefit from detecting a threat – say, an asteroid – early, but in our classification this is different from responding after there has been some actual small-scale damage.

**Figure 3.** Classification of risks by scaling mechanism.




### Leverage risks

Leverage points for rapid one-step scaling can be located in natural systems, for example if the extinction of a key species caused an ecosystem to collapse. However, it seems to us that leverage points are more common in technological or social systems that were designed to concentrate power or control.

Risks of both natural and anthropogenic origin may interact with such systems. For instance, a tsunami triggered the 2011 disaster at the Fukushima Daiichi nuclear power plant. Anthropogenic examples include nuclear war (possible to trigger by a few individuals linked to a larger chain of command and control) or attacks on weak points in key global infrastructure.

Responding to leverage risks is challenging because there are only few opportunities to intervene. On the other hand, blocking even one step of leveraged growth would be highly impactful. This suggests that response measures may be worthwhile if they can be targeted at the leverage points.

### Cascading risks

With the major exception of escalating conflicts, cascading risks normally cascade in a way which does not rely on humans deciding to further the effects. A typical example is the self-propagating growth of an epidemic. As automation becomes more widespread, there will be larger systems without humans in the loop, and thus perhaps more opportunities for different kinds of cascading risk.

Since cascading risks are those which have a substantial amount of growing effects after we're able to interact with

them, it seems likely that they will typically give us more opportunities to respond, and that response will therefore be an important component of risk reduction. For risks which cascade exponentially (such as epidemics), an earlier response may be much more effective than a later one. Reducing the rate of propagation is also effective if there exist other interventions that can eventually stop or revert the damage.

However, there are a few secondary risk-enabling properties that can weaken the response layer and therefore help damage cascade to a global catastrophe which we could have stopped. For example, a cascading risk may:

- Impede cooperation: by preventing a coordinated response, the likelihood of a global catastrophe is increased. Cooperation is harder when communication is limited, when it is hard to observe defection, or when there is decreased trust.
- Not obviously present a risk: the longer a cascading risk is under-recognised, the more it can develop before any real response. For example, long-incubation pathogens can spread further before their hazard becomes apparent.
- Be on extreme timescales: if the risk presents and cascades very fast, there is little opportunity for any response. Johnson et al. (2012) analyse such 'ultrafast' events, using rapid changes in stock prices driven by trading algorithms as an example (Braun et al., 2018, however find that most of these 'mini flash crashes' are dominated by a single large order rather than being the result of a cascade). Note, however, that which timescales count as relevantly 'fast' depends on our response capabilities – technological and institutional progress may result in faster-cascading threats but also in opportunities to respond faster. On the other hand people may be bad at addressing problems that won't manifest for generations, as is the case for some impacts of global warming.

### Policy implications for responding to extinction risk

- By their nature, we cannot respond to *large* risks before they become a global catastrophe. Of particular importance for such risks are therefore: mitigation that can be done in advance, and the defence layers of prevention and resilience.
- *Leverage* risks provide us with the opportunity of a leveraged response: we can identify leverage points in advance and target our responses at them.
- While the details of responses to *cascading* risks must be tailored to each specific case, we can highlight three general recommendations. First, detect damage early, when a catastrophe is still easy to contain. Second, reduce the time lag between detection and response, for example, by continuously maintaining response capabilities and having rapidly executable contingency plans in place. Third, ensure that planned responses won't be stymied by the cascading process itself – for example, don't store contingency plans for how to respond to a power outage on computers.[8]

## Classification by endgame: types of resilience failure

For a global catastrophe to cause human extinction, it must in the end stop the continued survival of the species. This could be *direct*: killing everyone;[9] or *indirect*: removing our ability to continue flourishing over a longer period (Figure 4).

### Direct risks

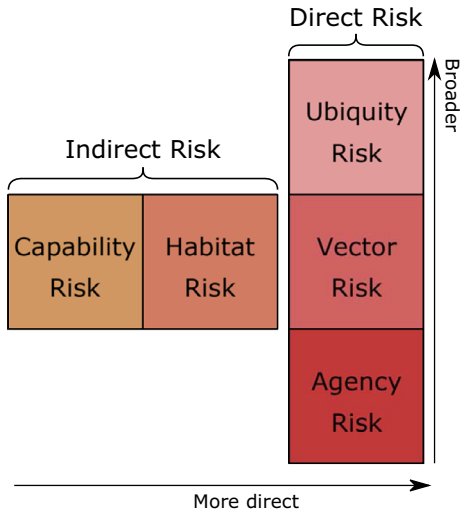
In order to kill everyone, the catastrophe must reach everyone. We can further classify direct risks by how they reach everyone.

The simplest way this could happen is if it is everywhere that people are or could plausibly be: a *ubiquity risk*. If the entire planet is struck by a deadly gamma ray burst, or enough of a deadly toxin is dispersed through the atmosphere, this could plausibly kill everyone.

If it doesn't reach everywhere people might be, a direct risk must at least reach everywhere that people in fact are. This might occur when people have carried it along with them: a *vector risk*. This includes risk from pandemics (if they are sufficiently deadly and have a long enough incubation period that it is spread everywhere) or perhaps risks which are spread by memes (Dawkins, 1976), or which come from some technological artefacts which we carry everywhere. Note that to directly cause extinction, a vector would need to impact hard-to-reach populations including 'disaster shelters, people working on submarines, and isolated peoples' (Beckstead, 2015a, p. 36).

If not ubiquitous and not carried with the people, we would have to be extraordinarily unlucky for it to reach everyone by chance. Setting this aside as too unlikely, we are left with *agency risk*: deliberate actors trying to reach everybody. The actors could be humans or nonhuman

**Figure 4.** Classification of risks by endgame.



intelligence (perhaps machine intelligence or even aliens). Agency risk probably means someone deliberately trying to ensure nobody survives, which may make it easier to get through the resilience layer by allowing anticipation of and response to possible survival plans. In principle agency risk includes cases where someone is deliberately trying to reach everyone, and only by accident does so in a way that kills them.

### Indirect risks

If the risk threatens extinction without killing everyone, it must reduce our long-term ability to survive as a species. This could include a very broad range of effects, but we can break them up according to the kind of ability it impedes.

*Habitat risks* make long-term survival impossible by altering or destroying the environment we live in so that it cannot easily support human life. For example a large enough asteroid impact might throw up dust which could prevent us from growing food for many years – if this was long enough, it could lead to human extinction. Alternatively an environmental change which lowered the average number of viable offspring to below replacement rates could pose a habitat risk.

*Capability risks* knock us back in a way that permanently remove an important societal capability, leading in the long run to extinction. One example might be moving to a social structure which precluded the ability to adapt to new circumstances.

We are gesturing towards a distinction between habitat risks and capability risks, rather than drawing a sharp line. Habitat risks work through damage to an external environment, where capability risks work through damage to more internal social systems (or even biological or psychological factors). Capability risks are also even less direct than habitat risks, perhaps taking hundreds or thousands of years to lead to extinction. Indeed there is not a clear line between capability risks and events which damage our capabilities but are not extinction risks (cf. section 6). Nonetheless when considering risks of human extinction it may be important to account for events which could cause the loss of fragile but important capabilities.

An important type of capability risk may be civilisational collapse. It is possible that killing enough people and destroying enough infrastructure could lead to a collapse of civilisation without causing immediate extinction. If this happens, it is then plausible that it might never recover, or recover in a less robust form, and be wiped out by some subsequent risk. It is an open and important question how likely this permanent loss of capability is (Beckstead, 2015b). If it is likely, the resilience layer may therefore be particularly important to reinforce, perhaps along the lines proposed by Maher and Baum (2013). On the other hand, if even large amounts of destruction have only small effects on the chances of eventual extinction, it becomes more important to focus on risks which can otherwise get past the resilience layer.

**Table 1.** Applying our classification to five examples. Note that each risk belongs to three classes, one for each defence layer

| Classification by<br>Associated defence layer | Origin<br>Prevention | Scaling<br>Response | Endgame<br>Resilience |
|---|---|---|---|
| Terrorists releasing engineered pandemic | Malicious | Cascading | Vector |
| Asteroid strike causing impact winter | Natural | Large | Habitat |
| False alarm triggering nuclear war with ensuing nuclear winter | Accident | Leverage | Habitat |
| Conventional proxy war escalating to nuclear war causing irreversible civilisational collapse | Conflict | Leverage | Capability |
| Unforeseen rapid learning producing an AI agent that kills humans to preempt interference with its objectives | Unseen | Leverage | Agency |

### Classifying example risks by each of origin, scaling, and endgame

We finally illustrate our completed classification scheme by applying it to examples, which we summarise in Table 1.

Throughout the text, we've repeatedly referred to an asteroid strike that might cause extinction due to an ensuing impact winter. We've called this a *natural risk* regarding its origin; a *large* risk regarding scale, with no opportunity to intervene between the asteroid impact and its damage affecting the whole globe; and, if we assume that humanity dies out because climatic changes remove the ability to grow crops, a *habitat risk* in the endgame phase.

Our next pair of examples illustrates that risks with the same salient central mechanism – in this case nuclear war – may well differ during other phases. Consider first a nuclear war precipitated by a malfunctioning early warning system – that is, a nuclear power launching what turns out to be a first strike because it falsely believed that its nuclear destruction was imminent. Suppose further that this causes a nuclear winter, leading to human extinction. This would be an *accident* that scales via *leverage,* and finally manifests as a *habitat risk*. Contrast this with the intentional use of nuclear weapons in an escalating conventional war, and assume further that this either doesn't cause a nuclear winter or that some humans are able to survive despite adverse climatic conditions. Instead, humanity never recovers from widespread destruction, and is eventually wiped out by some other catastrophe that could have easily been avoided by a technologically advanced civilisation. This second scenario would be a *conflict* that again scaled via the *leverage* associated with nuclear weapons, but then finished off humanity by removing a crucial *capability* rather than via damage to its habitat.

We close by applying our classification to a more speculative risk we might face this century. Some scholars (e.g. Bostrom, 2014) have warned that progress in artificial intelligence (AI) could at some point allow unforeseen rapid self-improvement in some AI system, perhaps one that uses machine learning and can autonomously acquire additional training data via sensors or simulation. The concern is that this could result in a powerful AI agent that deliberately wipes out humanity to pre-empt interference with its objectives (see Omohundro, 2008, for an argument why such preemption might be plausible). To the extent that we currently don't know of any machine learning algorithms that could exhibit such behaviour, this would be an *unseen risk;* the scaling would be via *leverage* if we assume a discrete algorithmic improvement as trigger, or alternatively the risk could be rapidly *cascading;* in the endgame, this scenario would present an *agency risk*.

### Policy implications for resilience against extinction

- To guard against what today would be *ubiquity risks,* we may in the future be able to establish human settlements on other planets (Armstrong and Sandberg, 2013).[10]
- *Vector risks* may not reach people in isolated and self-sufficient communities. Establishing disaster shelters may hence be an attractive option. Self-sufficient shelters can also reduce *habitat risk*. Jebari (2015) discusses how to maximise the resilience benefits from shelters, while Beckstead (2015a) has argued that their marginal effect would be limited due to the presence of isolated peoples, submarine crews, and existing shelters.
- Resilience against *vector* and *agency risks* may be increased by late-stage response measures that work even in the event of widespread damage to infrastructure and the breakdown of social structure. An example might be the 'isolated, self-sufficient, and continuously manned underground refuges' suggested by Jebari (2015, p. 541).

## Allocating resources between defence layers

In this section we will use our guiding idea of three defence layers to present a way of calculating the extinction probability posed by a given risk. We'll draw three high-level conclusions: first, the most severe risks are those which have a high probability of breaking through all three defence layers. Second, when allocating resources between the defence layers, rather than comparing absolute changes in these probabilities we should assess how often we can halve the probability of a risk getting through each layer. Third, it's best to distribute a sufficiently large budget across all three defence layers.

We are interested in the probability $p$ that a given risk $R$ will cause human extinction in a specific timeframe, say by 2100. Whichever three classes $R$ belongs to, in order to cause extinction it needs to get past all three defence layers; its associated extinction probability $p$ is therefore equal to the product of three factors:

1. The probability $c$ for $R$ getting past the first barrier and causing a catastrophe;
2. The conditional probability $g$ that $R$ gets past the second barrier to cause a global catastrophe, *given* that it has passed the first barrier; and
3. The conditional probability $e$ that $R$ gets past the third barrier to cause human extinction, *given* that it has passed the second barrier.

In short: $p = c{\cdot}g{\cdot}e$.

Each of $c$, $g$, and $e$ can get extremely small for some risks. But the extinction probability $p$ will be highest when all three terms are non-negligible. Hence we get our (somewhat obvious) first conclusion that the most concerning risks are those which can plausibly get past all three defence layers.

However, most concerning doesn't necessarily translate into the most valuable to act on. Suppose we'd like to invest additional resources into reducing risk $R$. We could use them to strengthen either of the three defences, which would make it less likely that $R$ passes that defence. We should then compare *relative* rather than absolute changes to these probabilities, which is our second conclusion. That is, to minimise the extinction probability $p$ we should ask which of $c$, $g$, and $e$ we can halve most often. This is because the same relative change of each probability will have the same effect on the extinction probability $p$ – halving either of $c$, $g$, or $e$ will halve $p$. By contrast, the effect of the same absolute change will vary depending on the other two probabilities; for instance, reducing $c$ by 0.1 reduces $p$ by $0.1{\cdot}g{\cdot}e$. In particular, a given absolute change will be more valuable if the other two probabilities are large.

When one of $c$, $g$, or $e$ is close to 100%, it may be much harder to reduce it to 50% than it would be to halve a smaller probability. The principle of comparing how often we can halve $c$, $g$, and $e$ then implies that we're better off reducing probabilities not close to 100%. For example, consider a large asteroid striking the Earth. We could take steps to avoid it (for example by scanning and deflecting), and we could take steps to increase our resilience (for example by securing food production). But if a large asteroid does cause a catastrophe, it seems very likely to cause a global catastrophe, and it is unclear that there is much to be done in reducing the risk at the scaling stage. In other words, the probability $g$ is close to 1 and prohibitively hard to substantially reduce. We therefore shouldn't invest resources into futile responses, but instead use them to strengthen both prevention and resilience.

What if each defence layer has a decent chance of stopping a risk? We'll then be best off by allocating a non-zero chunk of funding to all three of them – a strategy of defence in depth, our third conclusion. The reason just is the familiar phenomenon of diminishing marginal returns of resources. It may initially be best to strengthen a particular layer – but once we've taken the low-hanging fruit there, investing in another layer (or in reducing another risk) will become equally cost-effective. Of course, our budget might be exhausted earlier. Defending in depth therefore tends to

be optimal if and only if we can spend relatively much in total.

We close by discussing some limitations of our analysis. First, we remain silent on the optimal allocation of resources *between* different risks (rather than between different layers for a fixed risk or basket of risks); indeed, as we'll argue in section 6, comprehensively answering the question of how to optimally allocate resources intended for extinction risk reduction requires us to look beyond even the full set of extinction risks. We do hope that our work could prove foundational for further research that investigates both the allocation between risks and between defence layers simultaneously. Indeed, it would be straightforward to consider several risks $p_i = c_i \cdot g_i \cdot e_i$, $i = 1, \ldots, n$; assuming specific functional forms for how the probabilities $c_i$, $g_i$, and $e_i$ change in response to invested resources could then yield valuable insights.

Second, we have not considered interactions between different defence layers or different risks (Graham et al., 1995; Baum, 2019; Baum and Barrett, 2017; Martin and Pindyck, 2015). These can present both as tradeoffs or synergies. For example, traffic restrictions in response to a pandemic might slow down research on a treatment that would render the disease non-fatal, thus harming the resilience layer; on the other hand, they may inadvertently help with preventing malicious risk or being resilient against agency risk.

#### Policy implications for resource allocation within risk management

- The most important extinction risks to act on are those that have a non-negligible chance of breaking through all three defence layers – risks where we have a realistic chance of failing to prevent, a realistic chance of failing to successfully respond to, *and* a realistic chance of failing to be resilient against.
- Due to diminishing marginal returns, when budgets are high enough it will often be best to maintain a portfolio of significant investment into each of prevention, response, and resilience.

## Underlying risk factors: risks to the defence layers

In sections 2–4 we have considered ways of classifying threats that may cause human extinction and the pathways through which they may do so. Our classification was based on the three defence layers of prevention, response, and resilience.

Giving centre stage to the defence layers provides the following useful lens for extinction risk management. If our main goal is to reduce the likelihood of extinction, we can equivalently express this by saying that we should aim to strengthen the defence layers. Indeed, extinction can only become less likely if at least one particular extinction risk is made less likely; in turn this requires that it has a smaller chance of making it past at least one of the defence layers.

This is significant because there is a spectrum of ways to improve our defences depending on how narrowly our measures are tailored to specific risks. At one extreme, we can increase our capacity to prevent, respond to, or be resilient against one risk; for example, we can research methods to deflect asteroids. In between are measures to defend against a particular class of risk, as we've highlighted in our policy recommendations. At the other extreme is the reduction of *underlying risk factors* that weaken our capacity to defend against many classes of risks.

Risk factors need not be associated with any potential proximate cause of extinction. For example, consider regional wars; even when they don't escalate to a global catastrophe, they could hinder global cooperation and thus impede many defences.

Global catastrophes constitute one important type of risk factor. We already discussed the possibility of them making earth uninhabitable or removing a capability that would be crucial for long-term survival. But even if they do neither of these, they can severely damage our defence layers. In particular, getting hit by a global catastrophe followed in short succession by another might be enough to cause extinction when neither alone would have done so. There are significant historic examples of such *compound risks* below the extinction level. For instance, the deadliest accident in aviation history occurred when two planes collided on an airport runway; this was only possible because a previous terrorist attack on another airport had caused congestion due to rerouted planes, which disabled the prevention measure of using separate routes for taxiing and takeoff (Weick, 1990). When considering catastrophes we should therefore pay particular attention to negative impacts they may have on the defence layers.

Our capacity to defend also depends on various structural properties that can change in gradual ways even in the absence of particularly conspicuous events. For example, the resilience layer may be weakened by continuous increases in specialisation and global interdependence. This can be compared with the model of synchronous failure suggested by Homer-Dixon et al. (2015). They describe how the slow accumulation of multiple simultaneous stresses makes a system vulnerable to a cascading failure.

It is beyond the scope of this article to attempt a complete survey of risk factors; we merely emphasise that they should be considered. We do hope that our classifications in sections 2–4 may be helpful in identifying risk factors. For example, thinking about preventing conflict and common risks may point us to global governance, while having identified vector and agency risks may highlight the importance of interdependence (even though, upon further scrutiny, these risk factors turn out to be relevant for many other classes of risk as well).

We conclude that the allocation of resources between layers defending against specific risks, which we investigated in section 2, is not necessarily the most central task of extinction risk management. It is an open and important question whether reducing specific risks, clusters of risks, or underlying risk factors is most effective on the margin.

### Policy implications from underlying risk drivers

- Research on smaller-scale risks should pay particular attention to how they might damage the three defence layers against extinction risks. Risk management should aim to mitigate such damage.
- Conversely, the study of extinction risks cannot be limited to individual triggers such as asteroids or specific technologies. It would be desirable to better understand which underlying risk factors contribute to extinction risk by weakening our defences. For example, in what ways does global interdependence make extinction from a global catastrophe more likely, and are there interventions to mitigate this effect?

## Conclusions

The study and management of extinction risks are challenging for several reasons. Cognitive biases make it hard to appreciate the scale and probability of human extinction (Wiener, 2016; Yudkowsky, 2008). Most potential people affected are in future generations, whose interests aren't well represented in our political systems. Hazards can arise and scale in many different ways, requiring a variety of disciplines and stakeholders to understand and stop them. And since there is no precedent for human extinction, we struggle with a lack of data.

Faced with such difficult terrain, we have considered the problem from a reasonably high level of abstraction; we hope thereby to focus attention on the most crucial aspects. If this work is useful, it will be as a foundation for future work or decisions. In some cases our classification might provoke thoughts that are helpful directly for decision-makers that engage with specific risks. However, we anticipate that our work will be most useful in informing the design of systems for analysing and prioritising between several extinction risks, or in informing the direction of future research.

### Data availability statement

Data sharing is not applicable to this article as no new data were created or analysed.

## Notes

[1]We are particularly indebted to Toby Ord for several very helpful comments and conversations. We also thank Scott Janzwood, Sebastian Farquhar, Martina Kunz, Huw Price, Seán Ó hÉigeartaigh, Shahar Avin, the audience at a seminar at Cambridge's Centre for the Study of Existential Risk (CSER), and two anonymous reviewers for helpful comments on earlier drafts of this article. We're also grateful to Eva-Maria Nag for comments on our policy suggestions. The contributions of Owen Cotton-Barratt and Anders Sandberg to this article are part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 669751).

1. In the terminology of the United Nations Office for Disaster Risk Reduction (UNDRR, 2016), response denotes the provision of emergency services and public assistance during and immediately after a disaster. In our usage, we include any steps which may prevent a catastrophe scaling to a global catastrophe. This could include work traditionally referred to as mitigation.
2. The concept of resilience, originally coined in ecology (Holling, 1973), today is widely used in the analysis of risks of many types (e.g. Folke et al., 2010). In UNDRR (2016) terminology, resilience refers to '[t]he ability of a system, community or society exposed to hazards to resist, absorb, accommodate, adapt to, transform and recover from the effects of a hazard in a timely and efficient manner, including through the preservation and restoration of its essential basic structures and functions through risk management.' In this article, we usually use resilience to specifically denote the ability of humanity as a whole to recover from a global catastrophe in a way that enables its long-term survival. This ability may in turn depend on the resilience of many smaller natural, technical, and socio-ecological systems.
3. Strictly knowledge and intentionality are two separate dimensions; however it is essentially impossible to intend the harm without being aware of the possibility, so we treat it as a spectrum with ignorance at one end, intent at the other end, and knowledge without intent in the middle. Again, there is some blur between these: there are degrees of awareness about a risk, and an intention of harm may be more or less central to an action.
4. There are degrees of lack of foresight of the risk. Cases where the people performing the activity are substantially unaware of the risks have many of the relevant features of this category, even if they have suspicions about the risks, or other people are aware of the risks.
5. They may not intend for that damage to cause human extinction – for the purposes of acting on this classification it's more useful to know whether they were trying to cause harm.
6. We thank an anonymous reviewer for suggesting the policy responses of avoiding dangerous technologies and mandating insurance.
7. Global coordination more broadly may however be a double edged tool, since increased interdependency if not well managed can also increase the chance of systemic risks (Goldin & Mariathasan, 2014).
8. We thank an anonymous reviewer for suggesting both the third general recommendation and the example.
9. What about a risk that directly kills, say, 99.9999% of people? Technically this poses only an indirect risk, since to cause extinction it needs to remove the capability of the survivors to recover. However, if the proportion threatened is high enough then we can reason that it must also have a way of reaching essentially everyone, so the analysis of direct risks will also be relevant.
10. Some scholars have argued that humanity expanding into space would increase other risks; see for example an interview (Deudney, n.d.) and an upcoming book (Deudney, forthcoming) by political scientist Daniel Deudney and Torres (2018a). Assessing the overall desirability of space colonisation is beyond the scope of this article.

## References

Armstrong, S. and Sandberg, A. (2013) 'Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox', *Acta Astronautica*, 89, pp. 1–13.

Auplat, C. A. (2012) 'The Challenges of Nanotechnology Policy Making PART 1. Discussing Mandatory Frameworks', *Global Policy*, 3 (4), pp. 492–500.

Auplat, C. A. (2013) 'The Challenges of Nanotechnology Policy Making PART 2. Discussing Voluntary Frameworks and Options', *Global Policy*, 4 (1), pp. 101–107.

Avin, S., Wintle, B. C., Weitzdörfer, J., Ó hÉigeartaigh, S. S., Sutherland, W. J. and Rees, M. J. (2018) 'Classifying Global Catastrophic Risks', *Futures*, 102, pp. 20–26.

Barrett, S. (2016) 'Collective Action to Avoid Catastrophe: When Countries Succeed, When They Fail, and Why', *Global Policy*, 7 (S1), pp. 45–55.

Baum, S. D. (2015) 'Risk and Resilience for Unknown, Unquantifiable, Systemic, and Unlikely/catastrophic Threats', *Environment Systems and Decisions*, 35 (2), pp. 229–236.

Baum, S. D. (2019) 'Risk-risk Tradeoff Analysis of Nuclear Explosives for Asteroid Deflection', *Risk analysis*, 39 (11), pp. 2427–2442

Baum, S. and Barrett, A. (2017) 'Towards an Integrated Assessment of Global Catastrophic Risk'. in B. J. Garrick (ed.), *Catastrophic and Existential Risk: Proceedings of the First Colloquium*. Los Angeles, CA: Garrick Institute for the Risk Sciences, University of California, pp. 41–62..

Baum, S. D. and Barrett, A. M. (2018) 'Global Catastrophes: The Most Extreme Risks', in V. Bier (ed.), *Risk in Extreme Environments: Preparing, Avoiding, Mitigating, and Managing*. New York, NY: Routledge, pp. 174–184.

Baum, S. D. and Handoh, I. C. (2014) 'Integrating the Planetary Boundaries and Global Catastrophic Risk Paradigms', *Ecological Economics*, 107, pp. 13–21.

Beckstead, N. (2015a) 'How Much Could Refuges Help us Recover from a Global Catastrophe?', *Futures*, 72, 36–44.

Beckstead, N. (2015b). 'The Long-term Significance of Reducing Global Catastrophic risks', *The GiveWell Blog*, 2015–08-13 [online]. Available from: https://blog.givewell.org/2015/08/13/the-long-term-significance-of-reducing-global-catastrophic-risks/ [Accessed 3 August 2018.].

Bostrom, N. (2013) 'Existential Risk Prevention as Global Priority', *Global Policy*, 4 (1), pp. 15–31.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bostrom, N. and Ćirković, M. M. (eds.) (2008) *Global Catastrophic Risks*. Oxford: Oxford University Press.

Braun, T., Fiegen, J. A., Wagner, D. C., Krause, S. M. and Guhr, T. (2018) 'Impact and Recovery Process of Mini Flash Crashes: An Empirical Study', *PLoS ONE*, 13 (5), e0196920.

Buchholz, W. and Schymura, M. (2012) 'Expected Utility Theory and the Tyranny of Catastrophic Risks', *Ecological Economics*, 77, pp. 234–239.

Centeno, M. A., Nag, M., Patterson, T. S., Shaver, A. and Windawi, A. J. (2015) 'The Emergence of Global Systemic Risk', *Annual Review of Sociology*, 41 (1), pp. 65–85.

Chernov, D. and Sornette, D. (2015) *Man-made Catastrophes and Risk Information Concealment: Case Studies of Major Disasters and Human Fallibility*. Cham, Heidelberg, New York, Dordrecht, London: Springer.

Clarke, R. A. and Eddy, R. P. (2017) *Warnings: Finding Cassandras to Stop Catastrophes*. New York: Harper Collins.

Dawkins, R. (1976) *The Selfish Gene*. Oxford: Oxford University Press.

Deudney, D. (n. d.) '*An Interview With Daniel Deudney' [online]*. Available from: http://wgresearch.org/an-interview-with-daniel-h-deudney/ [Accessed 08 August 2018].

Deudney, D. (forthcoming) *Dark Skies: Space Expansionism, Planetary Geopolitics, and the Ends of Humanity*. Oxford: Oxford University Press.

Folke, C., Carpenter, S. R., Walker, B., Scheffer, M., Chapin, T. and Rockström, J. (2010) 'Resilience Thinking: Integrating Resilience, Adaptability and Transformability', *Ecology and Society [online]*, 15 (4), art. 20.

Foster, K. R., Vecchia, P. and Repacholi, M. H. (2000) 'Science and the Precautionary Principle', *Science*, 288 (5468), pp. 979–981.

Goldin, I. and Mariathasan, M. (2014) *The Butterfly Defect: How Globalization Creates Systemic Risks, and What to Do About It*. Princeton, NJ: Princeton University Press.

GPP (Global Priorities Project) (2015) '*Policy Brief: Unprecedented Technological Risks' [online]*. Available from: https://www.fhi.ox.ac.uk/wp-content/uploads/Unprecedented-Technological-Risks.pdf [Accessed 08 August 2018].

Graham, J. D., Wiener, J. B. and Sunstein, C. R. (eds.) (1995) *Risk vs. Risk*. Cambridge, MA: Harvard University Press.

Holling, C. S. (1973) 'Resilience and Stability of Ecological Systems', *Annual Review of Ecology and Systematics*, 4 (1), pp. 1–23.

Homer-Dixon, T., Walker, B., Biggs, R., Crépin, A. S., Folke, C., Lambin, E. F. et al. (2015) 'Synchronous Failure: The Emerging Causal Architecture of Global Crisis', *Ecology and Society [online]*, 20 (3), art. 6.

Jebari, K. (2015) 'Existential Risks: Exploring a Robust Risk Reduction Strategy', *Science and Engineering Ethics*, 21 (3), pp. 541–554.

Johnson, N., Zhao, G., Hunsader, E., Meng, J., Ravindar, A., Carran, S. and Tivnan, B. (2012) 'Financial Black Swans Driven by Ultrafast Machine Ecology', *arXiv preprint arXiv:1202.1448*.

Jones, H., O'Brien, M. and Ryan, T. (2018) 'Representation of Future Generations in United Kingdom Policy-making', *Futures*, 102, pp. 153–163.

Lewis, G. (2018) '*Horsepox Synthesis: A Case of the Unilateralist's Curse?' [online]*. Available from: https://thebulletin.org/2018/02/horsepox-synthesis-a-case-of-the-unilateralists-curse/ [Accessed 08 August 2018].

Liu, H., Lauta, K. C. and Maas, M. M. (2018) 'Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research', *Futures*, 102, 6–19.

Maher, T. M. and Baum, S. D. (2013) 'Adaptation to and Recovery from Global Catastrophe', *Sustainability*, 5 (4), pp. 1461–1479.

Martin, I. W. and Pindyck, R. S. (2015) 'Averting Catastrophes: The Strange Economics of Scylla and Charybdis', *American Economic Review*, 105 (10), pp. 2947–85.

Matheny, J. G. (2007) 'Reducing the Risk of Human Extinction', *Risk Analysis*, 27 (5), pp. 1335–1344.

Millett, P. and Snyder-Beattie, A. (2017) 'Existential Risk and Cost-Effective Biosecurity', *Health Security*, 15 (4), pp. 373–383.

Nordhaus, W. D. (2011) 'The Economics of Tail Events with an Application to Climate Change', *Review of Environmental Economics and Policy*, 5 (2), pp. 240–257.

Omohundro, S. M. (2008) 'The Basic AI Drives.' In P. Wang, B. Goertzel and S. Franklin (eds) *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS, pp. 483–492.

Ord, T., Hillerbrand, R. and Sandberg, A. (2010) 'Probing the Improbable: Methodological Challenges for Risks with Low Probabilities and High Stakes', *Journal of Risk Research*, 13 (2), pp. 191–205.

O'Riordan, T. and Cameron, J. (eds) (1994) *Interpreting the Precautionary Principle*. London: Earthscan.

Pamlin, D. and Armstrong, S. (2015) *Global challenges: 12 Risks That Threaten Human Civilization*. Stockholm: Global Challenges Foundation.

Parfit, D. (1984) *Reasons and Persons*. Oxford: Oxford University Press.

Posner, R. A. (2004) *Catastrophe: Risk and Response*. Oxford: Oxford University Press.

Rees, M. J. (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 (AZ).

Rees, M. (2018) *On the Future: Prospects for Humanity*. Princeton, NJ: Princeton University Press.

Roberts, K. H. and Bea, R. (2001) 'Must Accidents Happen? Lessons from High-reliability Organizations', *Academy of Management Perspectives*, 15 (3), pp. 70–78.

Rowe, T. and Beard, S. (2018) *Probabilities, methodologies and the evidence base in existential risk assessments*. Working paper, Centre for the Study of Existential Risk, Cambridge, UK. Available from: http://eprints.lse.ac.uk/89506/ [Accessed 08 August 2018].

Sandin, P. (1999) 'Dimensions of the Precautionary Principle', *Human and Ecological Risk Assessment: An International Journal*, 5 (5), pp. 889–907.

Sandler, T. (2016) 'Strategic Aspects of Difficult Global Challenges', *Global Policy*, 7 (S1), pp. 33–44.

Sunstein, C. R. (2005) *Laws of Fear: Beyond the Precautionary Principle*, vol 6. Cambridge: Cambridge University Press.
Sunstein, C. R. (2007) 'The Catastrophic Harm Precautionary Principle', *Issues in Legal Scholarship [online]*, 6 (3). Available from: https://www.degruyter.com/view/j/ils.2007.6.issue-3/ils.2007.6.3.1091/ils.2007.6.3.1091.xml [Accessed 08 August 2018]
Sunstein, C. R. (2009) *Worst-case Scenarios*. Cambridge, MA: Harvard University Press.
Tonn, B. E. (1991) 'The Court of Generations: A Proposed Amendment to the US Constitution', *Futures*, 23 (5), pp. 482–498.
Tonn, B. E. (2009) 'Obligations to Future Generations and Acceptable Risks of Human Extinction', *Futures*, 41 (7), pp. 427–435.
Tonn, B. E. (2018) 'Philosophical, Institutional, and Decision Making Frameworks for Meeting Obligations to Future Generations', *Futures*, 95, pp. 44–57.
Tonn, B. and Stiefel, D. (2013) 'Evaluating Methods for Estimating Existential Risks', *Risk Analysis*, 33 (10), pp. 1772–1787.
Tonn, B. and Stiefel, D. (2014) 'Human Extinction Risk and Uncertainty: Assessing Conditions for Action', *Futures*, 63, pp. 134–144.
Torres, P. (2016) 'Agential Risks: A Comprehensive Introduction', *Journal of Evolution and Technology*, 26 (2), pp. 31–47.
Torres, P. (2018a) 'Space Colonization and Suffering Risks: Reassessing the "Maxipok Rule"', *Futures*, 100, pp. 74–85.
Torres, P. (2018b) 'Agential Risks and Information Hazards: An Unavoidable But Dangerous Topic?', *Futures*, 95, pp. 86–97.
Trask, A. (2017) '*Safe Crime Prediction: Homomorphic Encryption and Deep Learning for More Effective, Less Intrusive Digital Surveillance' [online]*. Available from: https://iamtrask.github.io/2017/06/05/homomorphic-surveillance/ [Accessed 8 Auguest 2018].
Turco, R. P., Toon, O. B., Ackerman, T. P., Pollack, J. B. and Sagan, C. (1983) 'Nuclear Winter: Global Consequences of Multiple Nuclear Explosions', *Science*, 222 (4630), pp. 1283–1292.
Umbrello, S. and Baum, S. D. (2018) 'Evaluating Future Nanotechnology: The Net Societal Impacts of Atomically Precise Manufacturing', *Futures*, 100, pp. 63–73.
UNDRR (United Nations Office for Disaster Risk Reduction) (2016) '*Report of the open-ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction*'. Document symbol A/71/644 [online]. Available from: http://undocs.org/A/71/644 [Accessed 08 August 2018].
Wagner, G. and Weitzman, M. L. (2015) *Climate Shock: The Economic Consequences of a Hotter Planet*. Princeton, NJ: Princeton University Press.
Weick, K. E. (1990) 'The Vulnerable System: An Analysis of the Tenerife Air Disaster', *Journal of Management*, 16 (3), pp. 571–593.
Weitzman, M. L. (2009) 'On Modeling and Interpreting the Economics of Catastrophic Climate Change', *The Review of Economics and Statistics*, 91 (1), pp. 1–19.
Wiener, J. B. (2011) 'The Rhetoric of Precaution', in J. B. Wiener, M. D. Rogers, J. K. Hammitt and P. H. Sand (eds) *The Reality of Precaution: Comparing Risk Regulation in the United States and Europe*. Abingdon: Earthscan, pp. 3–35.
Wiener, J. B. (2016) 'The Tragedy of the Uncommons: On the Politics of Apocalypse', *Global Policy*, 7 (S1), pp. 67–80.
Yudkowsky, E. (2008) 'Cognitive Biases Potentially Affecting Judgment of Global Risks', in N. Bostrom and M. M. Ćirković (eds) *Global Catastrophic Risks*. New York: Oxford University Press, pp. 91–119.

## Author Information

**Owen Cotton-Barratt** is a Mathematician at the Future of Humanity Institute, University of Oxford. His research concerns high-stakes decision-making in cases of deep uncertainty, including normative uncertainty, future technological developments, unprecedented accidents, and untested social responses.

**Max Daniel** is a Senior Research Scholar at the Future of Humanity Institute, University of Oxford. His research interests include existential risks, the governance of risks from transformative artificial intelligence, and foundational questions regarding our obligations and abilities to help future generations.

**Anders Sandberg** is a Senior Research Fellow at the Future of Humanity Institute, University of Oxford. His research deals with the management of low-probability high-impact risks, societal and ethical issues surrounding human enhancement, estimating the capabilities of future technologies, and very long-range futures.