Reconfiguring Resilience for Existential Risk

Submission of Evidence to the Cabinet Office on the new UK National Resilience Strategy

Prepared by: Matthijs M. Maas, Diane Cooke, Tom Hobson, Lalitha Sundaram, Haydn Belfield, Lara Mani, Jess Whittlestone, Seán Ó hÉigeartaigh,[1] on behalf of The Centre for the Study of Existential Risk (CSER).[2]

Submitted September 27th, 2021

This document presents expert responses to the Cabinet Office's Call for Evidence on the UK National Resilience Strategy, as submitted by CSER researchers to HMG on September 27th, 2021 (response ID: ANON-7FMB-F6JK-W).

Background:

In March 2021, the UK Government published *Global Britain in a Competitive Age: The Integrated Review of Security, Defence, Development and Foreign Policy*, providing a look at the challenges and opportunities the UK faces and will face over the next decade. This Integrated Review commits the Government to develop a new National Resilience Strategy, which is to outline a vision for UK resilience, and establish core objectives for achieving these.

Accordingly, in July 2021, the UK Government announced the '*UK National Resilience Strategy Call for Evidence*', seeking "public engagement to inform the development of a new Strategy that will outline an ambitious new vision for UK National Resilience and set objectives for achieving it." In response, an interdisciplinary team of experts at the Centre for the Study of Existential Risk (CSER) have worked to prepare a concrete response to this call. Through this document, we aim to share the contents of our submission for public deliberation.

---

[1] MMM (mmm71@cam.ac.uk) led the submission, all other authors contributed equally (order randomized).
[2] The Centre for the Study of Existential Risk (https://www.cser.ac.uk/) 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 civilisational collapse. 16 Mill Lane, Cambridge, CB2 1SB.

Approach of CSER responses:

As stated in the Call for Evidence's executive summary, the vision of the UK National Resilience Strategy is that by 2030:

> *[...] we will have a strengthened ability to assess and understand the risks we face. Our suite of systems, infrastructure and capabilities (including international systems) for managing those risks should become more proactive, adaptable and responsive; and there should be fewer regional inequalities in our resilience. As a result, our local communities, businesses, and the UK as a whole, will be more cohesive, resistant to shocks and stresses, and ultimately more adaptable to future threats and challenges.*

In responding to this vision, CSER experts in key risk domains have aimed to provide concrete input on the Strategy, both from a general perspective (in terms of risk-general insights and interventions for improving resilience), as well as in more specific recommendations for improving national resilience in key risk domains such as in biorisk, climate risk, or risks around emerging AI technologies or critical defence systems.

High-level takeaways:

We laud the UK Government's initiative to develop a new National Resilience Strategy, we argue that more work can be done to clarify its approach to resilience, and particularly its approach to global catastrophic and existential risks. As such, while we value the Government's recognition that catastrophic, complex, and existential risk are a separate category of risks which require distinct strategic responses, we argue amongst others that more should be done to categorize and identify global catastrophic and existential risks. We also emphasize the importance of taking a long-term perspective on mitigating and responding to the challenges these risks pose. We argue that such long-term approaches are key as effective resilience requires addressing a wide range of potential risk vectors in parallel to better deal with new and evolving challenges. Finally, we encourage the development of a more comprehensive strategy, as these risks are all intertwined in an interconnected and complex environment.

Structure:

The questions in this Call for Evidence focus on six broad thematic areas: *Risk and Resilience, Responsibilities and Accountability, Partnerships, Community, Investment,* and *Resilience in an Interconnected World.* We have organized our response to these questions below in the same thematic format. We have not answered every question posed, instead choosing to focus on those we are able to best address with our existing research and expertise. Please note that the views gathered and expressed here reflect those of the authors as experts, while drawing on CSER's resources. They may not reflect the views of all those working at CSER and should not be taken as such. We encourage further societal debate on not just this strategy, but the larger questions of UK national resilience into the long-term.

Table of Contents

Vision and Principles

Questions on the UK's proposed National Resilience Strategy.

17-18. To what extent do you agree with the proposed vision of the Resilience Strategy? Please explain your view:

The Centre for the Study of Existential Risk (CSER) congratulates the Government (hereafter 'HMG') for committing to a new National Resilience Strategy (hereafter 'the Strategy'). This is a positive and important step. CSER hopes that through its expertise it can help turn the Strategy into action, while refining its basis and purview.

As such, in the below responses, CSER experts in key risk domains aim to provide concrete input on the Strategy, both from a general perspective (in terms of risk-general insights and interventions for improving resilience), as well as in more specific recommendations for improving national resilience in key risk domains such as in biorisk, climate risk, or risks around emerging AI technologies or critical defence systems. The views gathered and expressed here reflect those of the authors as experts, while drawing on CSER's resources.

With regards to the overall proposed vision of the Resilience Strategy, we agree with a lot of the themes and ideas encapsulated by the overarching goal. In particular, we value and applaud:

(1). The Strategy's vision of making the UK the most resilient nation, one that is "better able to adapt to uncertainty, to proactively address risks, and to withstand adversity."

(2). The Strategy's opening 'case for reform', and its frank reflection and recognition that the COVID-19 pandemic highlighted elements of the resilience approach that need to be strengthened within the coming years.

(3). The Strategy's broad thematic scope, in terms of the range of risks to which HMG pays attention, which includes a wide range of potential risk vectors--including key emerging risks (such as rapid technological developments in AI technology, antimicrobial resistance, and biodiversity loss) which sometimes receive less attention than we believe is warranted.

(4). The Strategy's recognition (especially in Section Theme I) of catastrophic, complex, and existential risks as a critical category which requires its own set of bespoke planning and response measures.

(5). In a broader context, we commend HMG for the degree to which this National Resilience Strategy serves as part of a broader constellation of recent landmark UK initiatives, strategies, and statements, which together reflect HMG's growing awareness of the critical importance of taking a responsible, resilient, and long-term oriented approach to mitigating extreme risks, improving national and global resilience, and securing our common future into the long-term. For

example, it was heartening to hear the Prime Minister quote from our Oxford colleague Toby Ord in his speech to the United Nations (22 September 2021), and emphasize that follow-on commitment and swift action will now be crucial.

(6). We particularly value the way in which these different Strategies can strengthen one another in dealing with new and evolving threats to resilience, especially in the field of potential evolving risks from emerging technologies such as AI. As an example of this, we commend HMG's commitment, in the recently published National AI Strategy (22 September 2021), that "government takes the long term risk of non-aligned Artificial General Intelligence, and the unforeseeable changes that it would mean for the UK and the world, seriously", and that accordingly "The Office for AI will coordinate cross-government processes to accurately assess long term AI safety and risks". We value the recognition of these important goals (Ó HÉigeartaigh and Ord 2021), as well as the emphasis that the AI strategy puts, in this context, on supporting governmental research & monitoring infrastructures for AI, which have been a key theme of CSER's work over the past years. We see this as a key step, one that can not only secure technological innovation, but can also contribute to the goal of national resilience. As such, as HMG recognizes, it will be important to implement and coordinate these approaches and programs across government. It is excellent to see the National AI Strategy explicitly set out the importance of working with the National Resilience Strategy at the strategic level; we would be happy to provide input on integrating these strategies, and on aligning cross-governmental work on these issues more broadly.

---

19. Is there anything that you would add, amend or remove?

While we think the strategy's vision takes important steps in the right direction, there are a number of changes we recommend:

(1). We recommend that the Strategy's vision articulates that extreme, complex, and catastrophic risks are considered priorities, and are a key and explicit part of our understanding of a resilient UK. To build true resilience, global risks, including existential risks, must be a key part of the UK's Strategy for the coming decades. While such risks are currently briefly discussed in the Strategy (in the first thematic section on 'Risk and Resilience'), we encourage an explicit mention of these as part of the Strategy's opening Vision.

(2). We recommend that the vision clarifies the Strategy's long-term orientation and relevance. To its credit, while the Strategy's vision currently focuses on certain goals to achieve by 2030, it notes that 'the endeavour of improving national resilience will stretch far beyond this timeframe.' It would be important to clarify what this means, and what role the Strategy could play in either taking into account longer timeframes, or ensuring it is more adaptive. It would also be valuable for the Strategy to clarify how a continued 'long-term focus' relates not only to considering long-term risk trends, or long-term planning and investment, but also to questions such as the long-term interests of future generations.

(3). We recommend the Strategy, either in this vision or elsewhere, takes more space to explicitly articulate its relation (conceptually, organizationally, and concretely) to the constellation of other interconnected initiatives that are currently (being) articulated by HMG, in order to ensure this Strategy does not stand alone, but purposefully supports and amplifies the other, interconnected work of government on themes around ensuring the UK's long-term resilience and world-leading approach to emerging technologies. For instance, such links are explicitly articulated (with regards to the National Resilience Strategy) in the recently released National AI Strategy.

(4). We recommend the Strategy's vision clarifies the dependencies and timeframes for the core goals it sets out. This vision currently sets out two related goals--(a) to make the UK the world's most resilient nation (par 23); and (b) to achieve a set of intermediate goals to attain by 2030, which are to ensure a 'strengthened ability to assess and understand the risks we face' (par 26). These are valuable goals, but it would be useful for HMG to clarify how these are related or how these feed into one another: is the aim that, through achieving the Strategy's intermediate goals by 2030, that this makes the UK the world's most resilient nation on this same timeframe?

20-21. To what extent do you agree with the principles laid out for the strategy? Please explain your view. Please explain your view

We find the enumerated list of core principles to offer a valuable departure point for the strategy. In particular:

(1). We value the emphasis on a holistic understanding of the risk landscape, through a lens that highlights not just potential hazards, but also pre-existing vulnerabilities, the intersection of risks, and potential 'geographic and socioeconomic variations' in the impacts and consequences. This resonates with much of the approach and research taken in our own work (Avin et al. 2018; 2021; Liu, Lauta, and Maas 2018).

(2). We value the emphasis on investing in prevention, mitigation, and recovery to risks across the entire risk lifecycle; as well as the emphasis on "developing generic capabilities which can be used in many different scenarios" in order to ensure greater efficiency and adaptability.

(3). We value the recognition of the importance of transparency around risks, including through strategic communication mechanisms to all stakeholders.

---

22. Is there anything you would add, amend, or remove?

There are a number of additional changes and additions that we would recommend to the Strategy's principles:

(1). We recommend that the Strategy clarifies its underlying understanding of resilience, and explicitly how this ties into both its principles and its overarching vision. For instance, the 'vision' (see Q18-Q19) currently proposes a United Kingdom that is "better able to adapt to uncertainty, to proactively address risks, and to withstand adversity." This reflects a concept of resilience that focuses on the ability to operate under pervasive uncertainty, and to master contingencies through broad capabilities (which requires institutional capacity, foresight, and investment in basic capabilities). In other parts of the strategy, some policies that are put forward however that seem to emphasize identifying specific, discrete risks and raising emergency response should they occur. It would be valuable for the Strategy (in both principles and overarching vision) to set out a closer link to HMG's understanding of resilience. We believe that a holistic and proactive approach to fostering resilience capacity is likely to be most effective.

(2). We recommend that the Strategy's principles clarify HMG's understanding of how UK national resilience relates to and is predicated upon global resilience. In particular, we emphasize the importance of highlighting the key insights of the Thematic Section on 'Resilience in an Interconnected World' to a top-level principle. What this means is that it is key for the Strategy's Principles, and HMG's resilience policy, to recognize the fact that, where it concerns global risks, true resilience is impossible in isolation. It would be valuable to highlight the importance of proactive foreign policy efforts (including capacity building, aid, and soft power leadership (Hobson and Edwards 2021)), which will be a principal component to realizing the Strategy's vision of a resilient UK by 2030.

(3). As discussed previously (see Q19), the Strategy's principles also should embed the importance of approaching resilience as a long-term endeavour, and clarify the sort of institutional provisions that should ensure that the Strategy can adapt beyond its initial 2030 timeframe.

(4). As discussed previously (see Q19), the Strategy's first principle ('We should understand the risks we face, including the impacts they could have, and our exposure to them') should expand its list of 'interconnected factors' that must be surveyed. Specifically, it could add in an additional factor: 'the specific challenges around understanding-, evaluating, and preparing to mitigate global catastrophic- and existential risks'. The nature of global catastrophic and existential risks (complex and at times even unprecedented) make them difficult to assess and address, in comparison to more regularly occurring events such as floods, earthquakes or terrorist attacks. For that reason, we recommend HMG pay special attention to ensuring that any government risk assessment process is also able to keep global catastrophic and existential risks in scope (Avin et al. 2021).

(5). We recommend that HMG's National Resilience Strategy commits to embedding justice and fairness throughout its proposals. Point 26 of the Call for Evidence highlights HMG's vision of addressing regional inequalities in resilience. CSER's work has highlighted that addressing global catastrophic and existential risks will often require us to consider moral issues of global justice. Global and regional injustice may be a driver of such extreme risk: it is both an exacerbating

factor for specific hazards like climate change and global conflict, and a systemic factor driving societal vulnerability and hampering efforts to address risks. We therefore further recommend that HMG should take this opportunity to lead on this both at home and internationally.

Risk and Resilience

Questions on strengthening the UK Government's ability to manage an evolving risk landscape, by improving its capabilities to both predict and adapt to identified and unexpected challenges.

---

23. Is there more that the Government can do to assess risk at the national and local levels? If so, what?

Yes. We believe that the Strategy's approach to resilience is a promising and detailed one. However, we recommend a number of further approaches to the Strategy, both in its general approach, and with regards to specific risk domains (AI, bio, defence), which we will discuss below.

OVERALL RECOMMENDATIONS

On the sub-theme 'Risk Assessment':

(1): We recommend the Strategy articulates a systemic approach to extreme risks, that takes into consideration both the interaction between hazards, exposures and vulnerabilities (Avin et al. 2021; Liu, Lauta, and Maas 2018). We believe that by taking a systemic approach to identifying and mitigating extreme risk, and by considering the interaction between extreme risks and global justice, the government will be able to more comprehensively be able to assess risks at the national level. Instead of focusing on individual hazards which could precipitate a catastrophe, taking a systemic approach can both (a) help us to identify a wider range of both emergent and structural risks, as well as their drivers, and (b) enable us to find more effective mitigation strategies for reducing risk. A systemic approach involves three lenses, each recognizing how:

- Technological risks should be assessed in their social, political and environmental contexts.
- Extreme risks tend to be complex, with a significant potential for indirect harm, which should be assessed. This includes consideration of how individual hazards or vectors, even if each individually fails to rise to the level of a catastrophe, can nonetheless interact with one another (Beard et al. 2021)--such as in areas like (global) food insecurity, international conflict, or future geoengineering technologies (A. Tzachor 2020)--in ways that can increase aggregate risk to an extreme level, and threaten both national and international resilience.
- Mitigation strategies work better when they address society's structural vulnerability to catastrophes (Liu, Lauta, and Maas 2018).

(2). We recommend the Strategy highlights the relation between global risks and global justice:

Our research has found that global risk and global justice are closely related, and that tackling global risk requires tackling many questions of global justice. For example:

- Global injustice can often serve as an underlying driver of global risks, or barrier to their effective and coordinated mitigation
- Addressing global risks requires us to consider moral issues of global justice
- Global justice raises important questions about (risk) distribution

We have proposed various concrete policies for addressing global justice concerns around global risk, including:

- The formation of an All-party Parliamentary Group on Future Generations, which now exists, (Jones 2017; Jones, O'Brien, and Ryan 2018) and a Future Generations Commissioner, as proposed in Lord John Bird's Future Generations Bill.[3]
- Inclusion of obligations to consider the long-term risks or impacts of governmental policies.
- Promoting dialogues on global risks across diverse groups who may represent or emphasize different conceptions of justice or ethics.

On the sub-theme 'Risk Appetite', we recommend further consideration for the methods for assessing and shaping risk appetite:

(3). Prioritisation of risk by publics is not straightforward with complexities associated with how publics value risk; not all communities are affected by risks in the same way and society cannot be considered as homogenous. The Strategy should acknowledge that some risks that may be considered as a high priority for mitigation by some publics may not align with those prioritised by the government. Participatory methodologies, such as a Citizens' Assembly, should be adopted to inform governmental decision making for mitigation of risk, ensuring that the views and perceptions of publics are incorporated. Where a misalignment in risk priority exists between government and publics, it is essential that this be communicated with transparency, in order to not undermine public sentiment and harm trust in those responsible for the decision-making.

On the sub-theme 'Handling catastrophic and complex risks', we recommend a number of changes to the Strategy. In particular, we recommend HMG clarifies its approach to global

[3] Post-submission addition: see (Lord Bird 2021).

catastrophic and existential risks, ensures these are in-scope, and articulates clearer institutional responsibilities.

(4). It is key for the Strategy to clarify terminology around concepts such as 'global catastrophic risk' or 'existential risk' (see also the discussion of these terms in Q39).

- '*Global catastrophic risks*' (GCRs) are those risks which could lead to significant loss of life or value across the globe, impacting all of humanity. While a clear delineation of the category has yet to emerge in the academic field, key works refer to disasters that inflict a loss of 10% or more of the human population, or (on lower thresholds) to more than 10 million deaths (Rhodes et al. 2016; Global Challenges Foundation 2017). While these are extreme scenarios that have not been experienced in living memory, they are certainly not historically unprecedented. Moreover, several scientifically-plausible scenarios have been identified which could lead to such losses today or in the future, including the use of nuclear or biological weapons in warfare, catastrophic climate change, and pandemics (Rhodes et al. 2016; Bostrom 2013; Ord 2020).

- '*Existential risks*' are those which could lead to 'the premature extinction of earth-originating intelligent life, or the permanent and drastic destruction of its potential for desirable future development' (Bostrom 2013; 2002). Unlike global catastrophic risks, existential risk scenarios do not allow for meaningful recovery, and are therefore, by definition, unprecedented in human history. Existential risk studies seeks to understand and mitigate events and processes that threaten the survival or welfare of large parts of the world population (Avin et al. 2021) and/or the destruction of humanity's long-term potential (Ord 2020). The term refers to any risk that could lead to human extinction or civilisational collapse - such as climate change, nuclear war, some pandemics, unaligned artificial general intelligence, and some natural risks (such as asteroid impacts or supervolcanoes).

(5). It is key for the strategy to clarify and reconfigure its approach to existential and catastrophic risks, to recognize that not all such risks are in fact very statistically unlikely; and to accordingly consider the mitigation of such risks as in principle being in-scope for HMG and for national resilience policy. Critically, the Strategy currently foresees very little planning role for HMG in responding to what it refers to as 'statistically unlikely' existential risks, on the argument that it might 'not be practicable for Government to plan for them' (Par 39). We do not believe this stance to be tenable, and instead hold that a Strategy that fails to reckon with existential and global catastrophic risks will not suffice to ensure a truly resilient UK into the long-term. This is for various reasons:

- A risk appearing statistically unlikely does not prima facie imply that government action is clearly unwarranted.[4] For instance, while a large asteroid strike is unlikely over the next century (perhaps a 1 in a million chance (Ord 2020)), NASA and the ESA nevertheless plan for this eventuality, and have 'planetary defence' programs - with which the UK Space Agency might productively collaborate.

- A range of other existential and global catastrophic risks are unfortunately far from 'statistically unlikely'. Rather, their likelihoods (while uncertain and low year-on-year) may be cumulatively large, enough so as to warrant government action. For instance, models of nuclear war risk indicate that the annualized likelihood of nuclear war may be as high as 1.17%--which would suggest that for a child born today, the compounding chance of her living through a nuclear war during their lifetime would be nearly 60% (S. Baum, de Neufville, and Barrett 2018; Rodriguez 2019). Several global catastrophic or existential risks may therefore have a likelihood that is low year-on-year, but over longer time periods become likely, or at least sufficiently probable that the potential stakes mean that these risks still require planning from HMG.

- Other growing threats present a risk profile that combines high probabilities of pervasively harmful impacts (especially under continued business-as-usual), with low-likelihood but extreme-impact 'tail risk' scenarios that could result in true global catastrophic risks. For instance, the accumulating hazards posed by climate change, resource exhaustion and environmental degradation are set to have large impacts on the UK and the world under most mainline scenarios. The expected impacts in terms of climate change, resource shortages, and biodiversity loss are, on present trajectories, likely to inflict significant global harms. Even if the resulting harms still fall short of the high threshold for a 'global catastrophic risk' (as we discuss in (4), above, and in Q39), they will certainly be 'catastrophic' in the sense of the Strategy's own definition (Annex B), Such impacts, by themselves, would therefore warrant increased efforts to prevent and mitigate these trends. Moreover, these trends also create small but certainly significant chances of inflicting globally catastrophic impacts (Beard et al. 2021). For instance, while understandings of earth's climate sensitivity are still continuously evolving (Sherwood et al. 2020), some climate research has estimated a 10% chance of exceeding a temperature rise of 6 °C by 2100, which would be catastrophic (given GHG concentrations of 700 ppm). Yet in public dialogue, many higher end warming scenarios (3 °C and above) remain severely neglected (Jehn et al. 2021). Here, again, HMG can and should consider work that could help mitigate the more likely impacts, which will plausibly help reduce the likelihood also of these worst-case outcomes that could plausibly lead to global

[4] Post-submission addition: a similar case has been recently made by (Wiblin and Harris n.d.), suggesting that since the risk of many existential risks remains alarmingly high, yet can be reduced at a reasonable cost, investments to reduce them pass mundane cost-benefit analyses--which has been one historical rationale for asteroid defense programs.

catastrophic consequences. With such risks, it is paramount that HMG's approach to national resilience avoids 'betting on the best case' only.

In sum, while there may be a few types of existential risks (such as supernovae) which really are so extremely rare that it may not be practicable for the Government to plan for them, this is unfortunately not the norm: there may be many existential risks that are (unfortunately) likely enough for the Government to plan for them. Many of these risks can be mitigated against and and prepared for with suitable policy and governance. It is critical for HMG to pursue mitigation strategies in coordination with various stakeholders at home, with other states and international institutions.

In line with other recommendations in this evidence submission, we advise the HMG should take a proactive approach to addressing the drivers and mitigating the effects of existential risks.

(6). More generally, it is important for the strategy to clarify and reconfigure its approach to existential and global catastrophic risks. Rather than only consider the (ex ante) statistical likelihood of existential and global catastrophic risks (as discussed in the Strategy, Par. 39), it is important to expand the approach to exploring pre-existing vulnerabilities and exposures (Liu, Lauta, and Maas 2018), infrastructural 'pinch points' (Mani, Tzachor, and Cole 2021), and how these feed into the UK's societal capacity to cope with extreme risk. This would simply align the Strategy's approach to global catastrophic risks with the more nuanced approach it is already taking to other types of risks.

Instead, it is both possible and important to classify global catastrophic risks not just in terms of its source 'hazard' (and whether that appears 'likely' or not), but also by the (1) critical systems which are seeing their safety boundaries exceeded; (2) spread mechanisms, and (3) prevention and mitigation failures. This can provide an analytical tool for studying both systemic risks as well as global catastrophic risks, without reducing their inherent complexity and helps identify underlying drivers (Beard and Torres 2020), policy levers, and other opportunities for mitigating them (Avin et al. 2021).

(7). The Strategy should recognize the international institutions, instruments, or governance structures that are relevant to monitoring, managing, and responding to a range of existential and global catastrophic risks (Kemp and Rhodes 2020) (see also Q61). These provide an emerging (if still fragmented) global governance regime which could gain from sustained engagement from HMG, and which would in turn contribute to UK national resilience.

(8). The Strategy should clarify roles and responsibilities for organizations with regards to the response to handling global catastrophic or existential risks. Currently, the Strategy suggests organizations such as CSER and the Future of Humanity Institute, are expected to play an 'important role' in 'monitoring these [existential] risks and indicating any changes in their likelihood". It would be helpful for HMG to clarify what kind of monitoring and reporting it would

hope for organizations such as ours to provide, and how they would expect to react to such organizations indicating (sudden) changes in the estimated likelihood of these risks.

(9). In line with the above, we stress that the work of institutions such as CSER and FHI need not be reserved to monitoring likelihoods of existential risks, but can also extend to various other roles, such as analysing, conducting and proposing: (a) resilient strategies under uncertain likelihoods or pervasively uncertain epistemic environments; (b) broadly beneficial policies that could help prepare for these problems while also addressing other societal challenges (S. D. Baum 2015); (c) the ways in which hazards could interact with exposures and vulnerabilities (the systemic approach sketched above); (d) the role that foresighting methodologies can play in studying rare or unprecedented (but potentially extreme) catastrophes (Rios Rojas et al. 2021) (e) horizon-scanning exercises (Kemp et al. 2020), and 'problem-finding' explorative research (Liu and Maas 2021), in order to improve the basis for more 'creative' scientific approaches which are fit for the particular challenges around studying extremely rare or unprecedented, but extremely high-stake catastrophic risks (Currie 2019). We are ready and willing to extend and deepen our work with HMG on resilience.

RISK DOMAIN-SPECIFIC RECOMMENDATIONS

Beyond these general changes to the Strategy's overall approach to assessing risks, we also recommend HMG pursues a number of domain-specific policies to improve resilience in particular domains.

In the domain of biological risks, we recommend:

(10). HMG should adopt the Recommendation, from the 'Future Proof' Report, to task one body with ensuring preparedness for the full range of biological threats the UK faces. It is important to use a wide definition of biosecurity here - as is done in the UK's Biosecurity Strategy (including threats from invasive species, laboratory accidents and deliberate harm) - to ensure that nothing is missed. Housing this work within a single body could help break the 'panic and neglect' cycle and ensure there is clear responsibility for long-term biological security in HMG (Ord, Mercer, and Dannreuther 2021). This is in line with oral evidence we (and others) have presented in the House of Lords (Avin et al. 2021; Sutherland et al. 2021).

(11). HMG should update the Biosecurity Strategy: in light of the pandemic, obviously, but also because several of the governmental bodies referenced therein have changed or shifted roles and new bodies are being planned. As a result, greater clarity as to roles and responsibilities is needed, with a much greater emphasis on implementation. In contrast to other countries' biosecurity strategies (as we note in Q60 below) relatively little attention is paid in the Strategy to implementation, but this is how we can ensure monitoring and accountability. Part of this implementation should, we recommend, be the adoption of evidence-based research agendas

into key areas of biosecurity, as we have suggested in our paper, "80 Questions for UK Biological Security" (Kemp et al. 2021).

(12). We further recommend that HMG takes a proactive approach towards capacity building for technology assessment and oversight within civil society, academia and practice communities. Robust and sustainable capacity to assess the consequences and hazards of technology development and scientific research must form a central component of any efforts to minimise risks from emerging technology, misuse or misapplication. Fostering these capabilities also presents an opportunity to embed democratic science principles into the national resilience strategy, and takes seriously the call for involvement from all members of society.

In the domain of risks from AI, while we are greatly encouraged by the steps made by HMG in other initiatives, such as most notably the recent National AI Strategy, to accurately assess long-term AI safety and risks (Ó HÉigeartaigh and Ord 2021) (see also Q18(6)), there is still more the government can do to assess risks at the national level, especially under the aegis of the National Resilience Strategy itself:

(13). We recommend HMG invests in policies that improve various stakeholders' ability to make verifiable claims about the (safety and robustness) properties of their AI systems, especially those deployed in Critical National Infrastructures (CNI). This can draw on the proposals and hardware, software, and institutional provisions surveyed in the report 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims' (Brundage et al. 2020). These policies include: (a) institutional mechanisms such as third-party auditing, red team exercises, bias and safety bounties, and the sharing of AI incidents; (b) software mechanisms, such as audit trails, interpretability solutions, and privacy-preserving machine learning solutions; and (c) hardware mechanisms, such as secure hardware, high-precision compute measurement systems, and compute subsidies for academia.

(14). We recommend HMG invests in policies to improve foresight and progress in tracking of AI research, as recommended in the Future Proof report (Ord, Mercer, and Dannreuther 2021). As part of this, we especially recommend investment in monitoring infrastructures to aggregate and track progress in, and impacts of, AI technology, as discussed in (Whittlestone and Clark 2021). Better information about the underlying aspects of AI technology and diffusion is essential in ensuring the government is not surprised by technological progress and have time and knowledge to prepare the tools to intervene before harm is realized. This includes major accidents, societal risks, malicious attacks (Brundage et al. 2018), and other types of harmful or structural forms of sociotechnical impact that can actively or passively threaten national or global resilience (Maas 2021). However, many governments today do not yet systematically utilize metrics and measures to govern AI in a systematic manner. The processes governments currently use to get information about AI, such as by convening experts, are insufficient due to their lack of speed and informal, piece-meal nature.

A push on the national level to build measurement and monitoring infrastructure would allow the government to better understand AI technology and its impacts, while also helping to create tools to intervene earlier. We therefore propose governments invest in initiatives to measure and monitor various aspects of AI research, deployment, and impacts. This would speed up governments' ability to regulate this technology, while also creating tools to intervene earlier and in ways with a lighter touch than regulation. Such measurements and monitoring would include assessing:

The capabilities and impacts of deployed systems:

- Continuously analyzing deployed systems for potential harms, as well as developing better ways to measure the impacts of deployed systems where such measures do not already exist.
- Developing better ways to measure the societal impacts of deployed systems.

The development and deployment of new AI capabilities:

- Tracking activity, attention, and progress in AI research by using bibliometric analysis, benchmarks and open-source data.
- Assessing the technical maturity of AI capabilities relevant to specific domains of policy interest.
- Developing better ways to assess progress

In the domain of risks around defence technologies, we recommend:

(15). HMG improve defence procurement systems around any military use of AI technologies (Belfield, Jayanti, and Avin 2020); such as by: (a). Improving systemic risk assessments in defence procurement; (b). Ensuring clear lines of responsibility; (c). Consider how shifts in international standards for autonomous systems will affect UK standards and practices, and build flexible procurement standards; (d). Updating the MoD's current definition of 'lethal autonomous weapons systems' to be in line with that of its allies (Belfield, Jayanti, and Avin 2020; Ord, Mercer, and Dannreuther 2021). The current definition remains idiosyncratic and sets so high a bar to cross (nearly equivalent to human-level intelligence) as to be almost meaningless in informing debates about actual, real-world or prospective applications of AI systems to military operations. This risks holding the UK back from providing global leadership, as well as creating uncertainty for the UK's defence procurement decisions, defence industry, and exports.

---

24. Is there more that the Government can do to communicate about risk and risk appetite with organisations and individuals? If so, what?

Yes. To improve communication about risk, we recommend that HMG undertake a number of steps:

(1). In order to better understand how to communicate about risk and risk appetite, the Government should work towards establishing how publics and organisations think about risk. Risk perceptions work for extreme risk is due to take place in CSER over the coming months and can contribute to this understanding. This work will seek to establish how publics prioritise extreme risk, how best we can communicate risk to publics and organisations, and what the expectations are for the mitigation and prevention of extreme, global catastrophic, and existential risks. However, further work should be conducted to explore public perceptions of risk, such as through participatory methods such as Citizens' Assemblies, providing a platform for two-way inclusive dialogues.

(2). HMG should take into consideration that there is no one method-fits-all solution for effective risk communication, and accordingly adopt well-established and evidence-based methodologies, such as scenario-based exercises, wargaming, role-playing and narrative-based tools. Different methods and tools will be more effective with different audiences and applied to different messages. Development of a robust communication strategy for risk communication and risk appetite, along with message testing can ensure a more rigorous approach.

(3). A key theme for communication is to raise awareness of- and engagement around the close connection between extreme risks and (global) justice (Avin et al. 2021). This highlights the importance of addressing injustices as key measure for global and national resilience, not just in order to prevent and mitigate risks of disasters, but also by improving resilience. This can be done by engaging with a diverse group of individuals and organizations who may represent or emphasize different conceptions of justice or ethics. In particular it would be worth communicating concerns about:

- How rising inequality and power differentials might undercut programs and efforts to mitigate extreme risks;
- How national or global policies to mitigate extreme risks can be developed in a way that is legitimate and able to elicit meaningful and authentic support from across society and diverse stakeholders.

(4). For example, we recommend HMG establish more participatory methods to establish and explore risk appetite with different stakeholders around the impacts- and risks of emerging technologies (Cremer and Whittlestone 2021). These can take the form of group exercises and interactive games (Avin, Gruetzemacher, and Fox 2020).

(5). Furthermore, we suggest that HMG also inform and instill a culture of systemic risk awareness amongst 'universal owners', the class of institutional investors that by their nature cannot stock-pick their way out of a crisis, thus aligning significant financial interest and resources with broad risk management priorities (Quigley 2020).

---

26. How does your organisation assess risks around unlikely or extreme events, when there is limited or no data?

As discussed in previous reports (Avin et al. 2021), identification and assessment of global catastrophic or existential risks is a core activity at CSER; we have gained a better understanding of the challenges involved in foresight for these risks, and have developed a range of methods to overcome these.

(1). For critical systems that are essential for survival yet subject to constantly-evolving transformations and threats (such as critical ecosystems), we adopt a horizon scanning method based on the Investigate, Discuss, Estimate, Aggregate (IDEA) protocol.[5]

(2). To help direct research activities towards the most pressing topics, we use modified expert elicitation to identify specific questions that are of sufficient breadth and importance to set field-wide research agendas, as for biosecurity in the UK (Kemp et al. 2020).

(3). For exploration of near-term developments in technological domains, such as biotechnology or misuse of artificial intelligence, we use regular expert elicitation exercises which emphasise a diversity of experts, and incorporate a "red team" approach to increase the range and creativity of scenarios considered. Concretely, we therefore recommend that HMG follow the recommendation from the 'Future Proof' report (Ord, Mercer, and Dannreuther 2021), to normalise red-teaming in Government, including by creating a dedicated red team to conduct frequent scenario exercises.

(4). To assist in the exploration of longer-term technological developments we combine theoretical analysis and survey work to identify key themes and milestones that can structure future foresight exercises.

(5). To keep track of the fast-expanding and intrinsically interdisciplinary literature on global catastrophic and existential risks, we have developed a scientific literature crawling system which combines crowdsourcing and machine learning elements, to identify and curate potentially relevant scientific work as soon as it is published (Shackelford et al. 2019).

[5] Post-submission addition: see (Hanea et al. 2017).

Responsibilities and Accountability

Questions on building resilience to have a clear understanding of when, where and how to apply tools, processes and relationships effectively.

---

28. Do you think that the current division of resilience responsibilities between Central Government, the Devolved Administrations, local government and local responders is correct?

Yes. In terms of the division of resilience responsibilities between different levels of government, we would particularly emphasize that the responsibility for monitoring AI to reduce risk and increase resilience should be predominantly owned by the national level of government, to ensure centralized and uniform measurement and monitoring. Moreover, owning this at a national level enables a more unified response when leveraging the data collected from monitoring to address potential challenges (Whittlestone and Clark 2021).

---

31. The primary legislative basis for emergency management is the Civil Contingencies Act 2004 (CCA). Specific questions on the CCA are covered in Annex A. The UK's resilience also depends on legislation covering specific risk areas including, for example, the Terrorism Act 2000 and the Climate Change Act 2008, amongst others. What do you consider the advantages and disadvantages of the current legislative basis for resilience?

We urge caution with regards to any revisions of the CCA that would provide HMG with the ability to raise more emergency powers in the absence of further provisions for democratic, parliamentary, or public rights to oversight or recourse.

A theme we have highlighted throughout our submissions to this Call for Evidence is that HMG's National Resilience Strategy should focus on developing resilience and preparedness through: foresight, building capacity and providing resources within institutions and communities; funding research to better understand the drivers of extreme risks; and international leadership to address them and foster resilience globally. The constitutional democratic basis of existing emergency management legislation is an important safeguard for the resilience of the UKs governmental institutions and its populace.

Partnerships

Questions on how other parts of society play an essential role in building our collective resilience.

---

32. Do you think that the resilience of CNI can be further improved? If so, how?

Yes. As discussed in CSER's previous policy submissions (Avin et al. 2021), it is possible to further mitigate risk and enhance resilience across society, including in CNI systems, by developing policies that address the prevention and mitigation of risk systemically.

This can be done through better identifying and understanding the risks posed, and developing tools that address these risks. We recommend that the UK should develop policies both to address specific risks and to reduce systemic vulnerabilities, which can be driven by environmental, technological, or social issues. CSER has developed number of detailed policy recommendations that address a range of these specific risks that would be relevant to the CNI (see below).

Furthermore, as many extreme risks are global in nature, national risk mitigation efforts should also include pursuing international agreements and action, which the UK is in a strong position to do (see also response to Qs59-63). Finally, the resilience of CNI can be improved by understanding our awareness of which of these systems (both national and international) converge at 'pinch points' of heightened vulnerability (Mani, Tzachor, and Cole 2021). Incorporating these can contribute significantly to overall resilience of CNI.

In the specific domain of risks related to the integration of AI in CNI, we emphasize two policies:

(1). Care should be taken to take on board the lessons from historical experience with cascading 'normal accidents' (Perrow 1984). Such failure modes are likely to occur not just in existing CNIs, but also (or especially) in future applications of AI systems in various domains (Maas 2018). As such, care should be taken to ensure that automated fail-safes do not inadvertently contribute to human over-trust, automation bias, and the exacerbation of failure modes.

(2). Where it comes to the intersection of CNI with AI, the deployment of, for instance, reinforcement learning-based AI systems to CNI systems should receive careful scrutiny (Whittlestone, Arulkumaran, and Crosby 2021). Simultaneously, HMG can use the information generated by measurement and monitoring exercises to exert greater influence over improving the resilience of the CNI (Whittlestone and Clark 2021). This includes HMG playing a greater role in discussions about what traits or features in AI should be measured, how they should be measured, and what should be prioritized and when. Similarly, gathering information about the state of deployed AI systems, their capabilities, and where they're being deployed, can give governments a greater ability to identify areas where it may wish to support further deployments, or areas where it may want to take a more active regulatory role.

---

33. Do you think the introduction of appropriate statutory resilience standards would improve the security and resilience of CNI operators? Why? How would such standards define the necessary levels of service provision? Are there any risks associated with implementing such standards?

Yes. We suggest that:

(1). Overall CNI resilience can be improved if the government drives to develop a number of regulatory standards (Avin et al. 2021), including through:

- Integration of risk assessment into the earliest stages of developing and procuring novel technologies, especially for safety-critical or defence-related systems.
- Ensuring throughout-lifetime accountability for high-technology systems, particularly those used in security contexts.
- Investing in systematic and regular auditing of CNI systems and operators.

(2). In the domain of AI, we expect that statutory resilience standards could play a big role in improving the security and resilience of CNI operators. Much of the CNI is owned by the private sector, where currently AI technological development is progressing unmonitored. The unregulated nature of this market increases the potential for harm or misuse. By being the driving force behind national monitoring and assessment efforts, the government would be able to ensure operators are held to an explicit set of standards in how they are required to develop and incorporate AI technology to ensure minimum levels of safety. Furthermore, further regulatory involvement would allow for the identification of operators that provide better AI technological products and subsequently would be preferred operators over those who are unable to conform to these standards.

(3). In the domain of CNI systems integrated in defence roles, we recommend HMG follow the recent recommendations, in the 'Future Proof' report (Ord, Mercer, and Dannreuther 2021), proposing that:

- The UK Government refrains from incorporating AI systems into NC3 (nuclear command, control, communications) systems; and leads on establishing this norm internationally, and in emphasizing to other states the particular risks to strategic stability and mutual resilience that could emerge from such arrangements (see also (Avin and Amadae 2019)).
- HMG sets up throughout-lifetime stress-testing of computer and AI system security.
- HMG establishes a new Defence Software Safety Authority as a sub-agency of the Defence Safety Authority, to protect UK defence systems from emerging threats.

---

34. What do you think is the most effective way to test and assure the resilience of CNI?

We recommend a combination of institutional, analytical, and policy instruments to test and encourage the resilience of the CNI against extreme risks. These include tools for foresight, intervention, implementation, and governance, and are discussed in more detail in various CSER reports (Avin et al. 2021; Rios Rojas et al. 2021). It is necessary to implement these tools together in parallel to be as effective as possible.

In the specific case of ensuring the resilience of CNIs involving AI systems, we recommend policies that include: tests for risks from automation bias; test to ensure robustness against adversarial inputs or hacking; special care around integration of reinforcement learning-based AI systems in critical infrastructures (Whittlestone, Arulkumaran, and Crosby 2021); the measurement and monitoring of AI technological development, driven by HMG, in order to:

- Test deployed systems to see if they conform to regulation
- Incentivize positive applications of AI via measuring and ranking deployed systems
- Engage in more rigorous and coordinated approaches to impact assessment and assurance.

---

35. To what extent do you think regulators should play a role in testing the resilience of CNI systems and operators? [Multiple choice]

A substantial role.

---

36. During an emergency, what do you think should be the role of the operators of CNI in ensuring continued provision of essential services (e.g. water, electricity, public transport)?

Continuity of CNI services is incredibly significant - particularly where suspension or removal of access to these services would lead to degradation of public health or localised deprivation from essential resources - such as water or electricity. At the same time, there may arise situations where (brief) interruptions in system service provision may need to be accepted rather than held out against, in order to intercept or arrest accident cascades, and prevent in-progress disasters from getting (even) worse. In order to navigate this tension, we recommend that HMG's National Resilience Strategy works towards:

(1). Developing extensive redundancy and spare capacity within CNI networks, so that service interruptions are mitigated against as far as possible,

(2). Acknowledging that resilience may at times require prioritisation of long-term or sustainable continuity of service, and therefore building response mechanisms that can effectively manage any necessary short term or localised service interruptions. Modular interruptions such as these could serve effectively as "fire breaks" that insulate systems. In any event, HMG should prioritise preparation for effective public communication, building public trust in relevant institutions and building capacity to provide the necessary alternative services and to restore CNI rapidly.

---

37. How can the Government support CNI owners or operators during an emergency?

We expect regulators would be useful in playing a significant role in testing the resilience of CNI systems and operators. As already established stakeholders engaging in the assessment of CNI systems and operators, regulators could work with government and government-partnered organizations to ensure necessary data collection for effective measurements and monitoring. In addition, they are well placed to audit CNI systems and operations, as well as enforce minimum standards built based off these monitoring and measurements. Finally, they would have the authority to take the necessary steps to address potential issues found via these audits.

Moreover, regulation in areas such as CNI are necessary in order to enable further resilience against risks. System-wide regulations led by the government can build an ecosystem that is able to hold developers of emerging technologies accountable, thus creating an environment where user trust can be placed in trustworthy actors.

Finally, HMG can support CNI owners during an emergency, by identifying not just the sources of potential dangers, but also their potential dissemination (spread) mechanisms and links (Avin et al. 2018; Cotton-Barratt, Daniel, and Sandberg 2020).

---

38. What role, if any, does your business or sector play in national resilience?

CSER's mission is to study and mitigate global catastrophic and existential risks. The emphasis here is on understanding our risk environment. As noted by the Strategy, such understanding is a key initial step when developing resilience. There are a number of concrete assets which an institution such as CSER can offer to support national resilience: (a). rigorous analysis of sources of risk; (b). support in establishing monitoring infrastructure; (c). expert advice on specific decisions; (d). shaping informed national conversation.

In the domain of resilience to risks from AI technology, currently, the academic sector plays a large role, if not the largest, in the ongoing monitoring of AI technological development. This has enabled risks or potential issues resulting from the use of AI technology to be identified earlier and more effectively. Examples include racial biases found in major facial recognition systems. (i.e. Gender Shades project (Buolamwini and Gebru 2018)). However, increasing

compute resource costs and requirements for leading AI projects are making it harder for academics to adequately verify or check a number of private AI projects, which could be a site of greater government support for providing computing resources to such academic actors (Brundage et al. 2020). Finally, academic actors are also well-placed to convene multi-stakeholder groups that include not just scientists and policymakers but also citizens, in order to carry out 'Participatory Technology Assessments' (Cremer and Whittlestone 2021).

---

39. What are the risks that your business or organisation is most concerned about?

As also touched on above (in Q23), our work is structured by a taxonomy of extreme risks. It should be noted here that in the National Resilience Strategy, the current definition of 'Catastrophic Risk' in Annex B. [Glossary] is given as:

> "Those risks with the potential to cause extreme, widespread and/or prolonged impacts, including significant loss of life, and/or severe damage to the UK's economy, security, infrastructure systems, services and/or the environment. Risks of this scale would require coordination and support from the central Government. Examples include: the widespread dispersal of a biological agent, severe flooding, or the detonation of an improvised nuclear device."

(1). While a valuable starting point for a definition of 'catastrophic risk', we recommend HMG expands this Glossary--and its approach in the overall Strategy--to order to also take account of additional categories of risks. Specifically, Annex B. could include additional definitions for 'global catastrophic risks' and 'existential risks', following the definitions developed in work by CSER (Avin et al. 2021) (see the definitions provided in Q23(4)).

(2). In addition to these core categories of global catastrophic and existential risks, CSER and related institutions such as the Centre for Long-Term Resilience more broadly use 'extreme risks' to refer to both categories of risks (Ord, Mercer, and Dannreuther 2021). What matters practically is that the nature of global catastrophic risks and existential risks (complex and unprecedented) makes them difficult to assess and address, in comparison to more regularly occurring events such as floods, earthquakes or terrorist attacks. We argue that, because of this, it is especially warranted for HMG and the National Resilience Strategy to pay significant attention to these kinds of risks. This is not just because of their potential extreme stakes, but also because a range of cognitive biases mean that we are prone to underestimating the likelihood and/or impacts of such risks (Yudkowsky 2011; Wiener 2016; Liu, Lauta, and Maas 2020); moreover, even risk that are discussed, such as extreme global warming scenarios, still receive structurally less attention than is warranted given their probabilities (Jehn et al. 2021).

(3). Beyond these risks, CSER's work is also concerned over a range of contributory and indirect risk factors, which contribute to our exposure and vulnerability to risks (Avin et al. 2018; Liu, Lauta, and Maas 2018). These include amongst others threats to our societal 'epistemic security' (Seger

et al. 2020) and our ability to coordinate internationally to mitigate extreme risks; as well as intermediate risk scenarios in certain domains, should policy not come to grips with these challenges. For instance, in the domain of AI, we predict that if the government does not build a cohesive monitoring infrastructure to track AI developments, we will see some version of the following over the coming years (Whittlestone and Clark 2021):

- Private sector interests will exploit the lack of measurement and monitoring infrastructure to deploy AI
- AI technology will inflict negative externalities, and HMG will lack the tools available to address these.
- Information asymmetries between the government and the private sector will widen, causing deployments to occur that negatively surprise policymakers, which will lead to hurried, imprecise, and uninformed lawmaking.
- Other interests will step in to fill the evolving information gap; most likely, the private sector will fund entities to create measurement and monitoring schemes which align with narrow commercial interests rather than broad, civic interests.

---

43. What can the Government do to make collaboration between academic and research organisations more effective?

In general, we recommend HMG could undertake new initiatives to involve academics in the emerging extreme and existential risks community in resilience assessments, given that many of these researchers are highly motivated to support however they can. This could take the form of:

- Advisory positions or secondment of researchers into government, in order to advise on how to do risk monitoring
- Setting up pilot systems for early warning systems and monitoring infrastructures
- Red teaming policy, scenario and tabletop exercises.

For collaboration on improving resilience to risks from AI technology, we have two suggestions:

(1). We support the recent recommendation of the 'Future Proof' report, for HMG to bring more technical AI expertise into Government through a scheme equivalent to TechCongress (at an estimated annual cost of £1.5 million) (Ord, Mercer, and Dannreuther 2021).

(2). For AI monitoring developments, while it may be useful to subcontract out some aspects of measurement and monitoring to third parties in the private sector or in academia (especially where deeper technical expertise may be needed), we nonetheless recommend that governments need to have a large degree of ownership over and visibility into this work in order

for it to strengthen policymaking (Whittlestone and Clark 2021). In particular, governments should set the objectives for projects and ensure that core infrastructure (e.g., aggregated datasets, search tools, indexes) remains within government (while ensuring it can also be accessible to third parties where needed). Measurement and monitoring infrastructure can both help target research funding in the most effective areas, and provide policymakers with more effective tools to incentivize research. If governments can robustly measure the things they care about, they can more easily create incentives for research and industry to build systems that perform better on these measures, through funding or competitions.

Examples of useful collaborations in this space could include:

- Partnering with research institutions to prioritize areas of specific interest
- Incentivizing research by hosting competitions
- Funding projects to improve assessment methods in commercially important areas (e.g. certain types of computer vision, to accelerate progress and commercial application in these areas.)

---

44. Are there areas where the role of research in building national resilience can be expanded?

In terms of research relevant to enhancing overall resilience, HMG could improve implementation of risk management. In particular, domain-specific research areas could include a series of themes identified in the recent 'Future Proof' report (Ord, Mercer, and Dannreuther 2021). These include (a) investment in AI safety R&D; (b) investment in applied biosecurity R&D; (c) further investment in improving long-term forecasting and planning.

In the domain of resilience to emerging risks from AI technologies, productive areas of research could include:

- Identifying new areas of measurement and monitoring in AI technology. Much of the current auditing and monitoring efforts have arisen from ongoing research in the academic sector, such as with facial recognition bias. Therefore, investing in research into new methods of monitoring would be invaluable. By investing in measurement and monitoring, policymakers will be better equipped to identify areas where more research can support a policy need -- and that research, in turn, is likely to generate more useful information for policymakers. For example, a better understanding of the metrics currently used in research to assess the fairness of AI systems would enable policymakers to identify specific types of fairness that aren't being evaluated, and push for more work to reduce these gaps.

- The creation of a pool of machine learning-relevant computation resources to provide free of charge for socially beneficial application and AI safety, security, and alignment research (£35 million annually) (Ord, Mercer, and Dannreuther 2021; see also Brundage et al. 2020).

In the domain of resilience to biological risks, productive areas of research could include: new horizon scans (Kemp et al. 2020; Sutherland et al. 2021), studies of risk governance lessons from COVID-19 which are not just narrowly tailored to responding to future coronaviruses, but to a broader range of uncertain but potentially high-stake risks (Liu, Lauta, and Maas 2020).

More generally, we highlight the importance of HMG in investing in support of academic research, which could enable various academic organizations to:

- Improve the processes for identifying and understanding extreme, global catastrophic or existential risks;

- Help build an ecosystem that is better able to hold developers of emerging technologies accountable, thus creating an environment where user trust can be placed in trustworthy actors;

- Help explore ways in which emerging technologies such as artificial intelligence can support crisis response, while also investing in mechanisms to speed up or pre-prime ethical review processes (‘doing ethics with urgency’) for the rapid rollout of such tools during crises (A. Tzachor et al. 2020).

Community and Local Resilience

Questions on a whole-of-society approach to strengthening the UK's resilience, emphasizing a revived effort to inform and empower all parts of society who can make a contribution.

---

45. Do you agree that everyone has a part to play in improving the UK's resilience? If not, why not?

Yes. We emphasize extensively, however, that the National Resilience Strategy should take a nuanced approach to the expectations it sets for all parties: everyone has a part to play, but not everyone is in the same position, or with access to the same resources, to play every part, or play their part fully. This should be attended to.

(1). For instance, people living in poverty and deprivation are likely to be at greater risk to life and livelihood from various extreme risks; however, the double-bind nature of the poverty trap means they may be less likely to engage with governmental organisations, due to low expectations of improvement to their situation. Breaking the barrier of low expectation is essential for ensuring all parts of society are in an empowered position where they can engage for the longer term, and truly play their part in building national resilience.

(2). The call for evidence for the National Resilience Strategy places great emphasis on the role of local communities, businesses and individuals in developing and maintaining UK resilience. Building capacities for resilience will require HMG to provide local communities, care providers and other core elements of LRFs with the resources they require to fulfil this mandate. Moreover, it is important that "resilience" not be seen simply as "responding". Local communities may well (if sufficiently resourced) be the best place to implement response and recovery strategies but they must also be empowered to take an active role in preparation and mitigation.

(3). Taking into account the need to take a holistic view of resilience, we recommend that HMG conducts further research as to how local resilience fostered in a fair and equitable fashion. While embedding resilience at a local level is likely to be of great value in the event of crisis, we also note that a focus on localised abilities to respond to crisis situations should not come at the expense of efforts to prevent or mitigate against extreme (particularly global catastrophic or existential) risks at the national or international level. Resilience is not only about emergency response, and many of the drivers of extreme risks that will affect the UK in the coming decades -- from climate change to resource exhaustion, and from new arms races to emerging disease -- will require global cooperation. The UK should seek to lead on these efforts.

---

53. Have recent emergencies (e.g. COVID-19 pandemic, flooding, terrorist attacks) made you think differently about risks or changed the way you prepare for emergencies?

Yes. The emergencies and experiences over the past years have in the first instance reinforced our institutional priority to examining rare but potentially catastrophic global risks. Catastrophic events like Covid have highlighted the need to discuss and prepare for extreme risks, and continuing CSER's research into the areas of global catastrophic and existential risk. We believe this has also shown how mitigation strategies work better when they address society's structural vulnerability to catastrophes. Direct or stopgap solutions to mitigate specific hazards or sources of risks may address part of the threat, but may do little to reduce overall risk (or increase resilience in a meaningful way) if underlying cross-domain vulnerabilities are not addressed (Liu, Lauta, and Maas 2018).

For instance, one such cross-cutting vulnerability, which the pandemic has brought to greater attention for CSER, is the challenge of reaching informed collective decisions during times of acute crisis. This can be difficult in any time, but becomes a growing challenge in our ability to respond to global catastrophic risks more generally, given that many democratic societies have in recent years seen their 'epistemic security' eroded by political polarization, disinformation, and emerging (media) technologies (Seger et al. 2020).

The Strategy currently places great emphasis on the role of local communities, businesses and individuals in developing and maintaining UK resilience. Building capacities for resilience will require HMG to provide local communities, care providers and other core elements of LRFs with the resources they require to fulfil this mandate. Building on CSERs work on systemic risk, and the need to take an holistic view of resilience, we recommend therefore that HMG conducts further research as to how local resilience fostered in a fair and equitable fashion.

Investment

Questions on the challenges of where to place investment in the risk cycle is one that affects the public and private sectors alike.

---

55. How does your organisation invest in your approach to the risks outlined in this document? Is your investment focussed on particular stages of the risk lifecycle (for example, on prevention)?

CSER is focused on the study and mitigation of global catastrophic and existential risks. While this means that a lot of our organization's work focuses on the prevention or mitigation of such risks ever arising or manifesting in the first place, we are also beginning to undertake work on measures that contribute to resilience by identifying existing 'pinch points' (Mani, Tzachor, and Cole 2021) in our infrastructures, by reducing vulnerabilities and exposures (Liu, Lauta, and Maas 2018), by improving the ability to intercept risk cascades (Cotton-Barratt, Daniel, and Sandberg 2020), and by improving the capacity to respond and adapt.

---

58. Are there examples of where investment (whether by the Government, by businesses or by individuals) has driven improvements in resilience?

In many risk domains, the UK has led on crucial international agreements and policies on risk mitigation, particularly in the field of arms control, through the Biological Weapons Convention, and also in relation to emerging technologies and nuclear weapons. These are great examples of the UK having driven the building of infrastructure and governance to enable further strengthening of resilience against potentially devastating risks. HMG can draw lessons from such past efforts, in shaping investment in a new focal point for a UK global technology assessment strategy. We have recently proposed a role for a coordinating institution (Hobson and Edwards 2021), to play a pivotal role in linking up capacities domestically in the area of innovation strategy and governance, with the UK's foreign policy agenda. This body could (1) systematically track developments of relevance to UK foreign policy, and (2) support the development and evolution of a more explicit and consolidated policy on the issue of global technology assessment.

Resilience in an Interconnected World

Questions on UK resilience in regards to the wider global context.

---

59. Where do you see the UK's resilience strengths?

As a global leader, the UK is well placed to lead efforts in resilience in insulating states against extreme risk. Since most or even all extreme risks will have substantial impacts globally, it is crucial for the UK to drive the building of resilient infrastructures internationally if the UK itself wishes to be more resilient domestically.

Indeed, it is key for HMG to consider how the success of the vision laid out in this National Resilience Strategy, will both depend on the approaches and efforts taken by other countries. The UK is well placed to take a proactive role in leading multilateral efforts to develop regional and global approaches to fostering resilience.There is an opportunity for the UK to establish a global reputation in setting international norms and proactive policy addressing various global catastrophic or existential risks, such as climate change, adequate preparation for future pandemics, and AI.

For example, the UK is one of the global leaders in AI technology, making it well placed to collect and measure the development of AI technology. Monitoring infrastructure would allow for a comparative analysis of the strength of countries' AI ecosystems, which would be useful in improving the UK's own resilience as well as advising other countries on how to improve their own resilience (Whittlestone and Clark 2021).

---

60. Are there any approaches taken by other countries to resilience that you think the UK could learn from?

There is a lot happening worldwide that the UK could productively take note of. For instance:

In the domain of biosecurity, the US's Federal Bureau of Investigation has recognised the need to keep up with not just technical developments in the biological sciences but also the growing and diversifying world of actors in this space (including, for example, those engaged in DIY-biology). This is being achieved through a programme of internal expertise-building and community-outreach and engagement (Evans et al. 2020).

Moreover, the US Biodefense Strategy is one that we could usefully use as a comparator. As noted in our response to Q23, implementation of a biological security strategy is key: the US's National Biodefense Strategy devotes fully one third of its text to this. We at CSER are in the process of planning a workshop (based on our previous work on Biosecurity Governance (CSER 2019)), to compare US and UK perspectives on biosecurity.

Another (international) effort in biosecurity is one from the International Genetically Engineered Machines Competition (a yearly initiative with 5,000 young synthetic biologists), which instills in its participants the principles of biosecurity and 'Human Practices' throughout. This form of governance, it is hoped, percolates through the research community (iGEM itself has produced an alumni community of approximately 40,000 synthetic biologists). It also enables a core team of biosecurity experts to keep track of new developments in the field (Millett et al. 2021).

---

61. Which of the UK's international relationships and programmes do you think are most important to the UK's resilience?

In general, as articulated in previous work, we recommend HMG formulate clearer guiding principles on the UK's foreign policy in monitoring developments in science and technology (Hobson and Edwards 2021).

Beyond this, recent CSER work has done much to map the existing global 'cartography' of international initiatives, regimes, and governance instruments that are pertinent to mitigating many global catastrophic or existential risks (Kemp and Rhodes 2020). This global initiative atlas found that there are clusters of dedicated regulation and action, including in nuclear warfare, climate change and pandemics, biological and chemical warfare. Despite these concentrations of governance their effectiveness is often questionable. For others risk vectors, such as catastrophic uses of AI, asteroid impacts, solar geoengineering, unknown risks, super-volcanic eruptions, inequality and many areas of ecological collapse, the global legal landscape remains littered more with gaps than effective policy.

On this basis, we therefore suggest the following steps to help advance the state of global GCR governance and fill the gaps:

(1). Work to identify instruments and policies that can address multiple risks and drivers in tandem.

(2). Closer research into the relationship between drivers and hazards to create a deeper understanding of our collective 'civilizational boundaries'. This should include an understanding of tipping points and zones of uncertainty within each governance problem area;

(3). Exploration of the potential for 'tail risk treaties': agreements that swiftly ramp-up action in the face of early warning signals of catastrophic change (particularly for environmental GCRs);

(4). Closer examination on the coordination and conflict between different GCR governance areas. If there are areas where acting on one GCR could detrimentally impact another than a UN-system wide coordination body could be a useful resource.

(5). Further work on building the foresight and coordination capacities of the UN for GCRs.

(6). Undertake all of the above informed by an awareness that addressing extreme risks will require HMG to address issues of global justice and equality.

Specifically, we therefore recommend the UK could play a forward looking role in exploring how its National Resilience Strategy can be embedded into a broader multi-level global governance architecture for global catastrophic risks (Avin et al. 2021). To do this, it will be important for the UK to better map the existing global governance architecture for different global catastrophic risk areas, understanding these regimes' maturity, overlaps, and gaps, in order to identify opportunities to patch, support or strengthen this architecture. A starting point for this could be found in early UK support for the new 'Common Agenda' that has been recently set forth by UN Secretary-General Guterres (United Nations 2021).

Finally, in the specific risk domain for AI, there is currently a window of opportunity for the UK to help shape a global institutional governance landscape for AI that is in some flux (Cihon, Maas, and Kemp 2020a; 2020b). We recommend the UK highlight and pursue rapid action in coalitions such as the Global Partnership for AI (GPAI), the G20 (Jelinek, Wallach, and Kerimi 2020), the range of multilateral UN initiatives (Garcia 2020), and the AI Partnership for defence (Trabucco 2020), in order to help set shared norms and expectations, and draw clear red lines around destabilizing or systemically risky uses of AI technologies (such as for instance in defence roles adjacent to nuclear command and control NCI (Ord, Mercer, and Dannreuther 2021).

---

62. What international risks have the greatest impact on UK resilience?

As is the nature of global catastrophic risks, these will likely pose substantial or even impossible challenges for UK resilience to overcome in isolation. This is especially true in the case of existential risks, which by their definition do not allow for meaningful recovery. That is why we emphasize the importance of investing in the prevention and mitigation of said risks.

---

63. How can the UK encourage international partners to build resilience to global risks?

Global risk mitigation requires us to reassess both national and international governance structures. As a global leader, the UK is well situated to both encourage partners and allies to take up best practices, as well as lead in the development of shared international governance architectures and coordination systems to build not just disseminated, but joint resilience to global risks. We recommend (Avin et al. 2021):

(1). HMG should examine how the National Resilience Strategy can be embedded into a broader multi-level global governance architecture for global risks. By mapping the existing global governance architecture for global risk areas, this can help the government better understand these regimes' maturity, overlaps, and gaps, and identify opportunities to patch, support or strengthen this architecture.

(2). HMG should examine how UK action can play a key role during the current window of opportunity, after the COVID-19 pandemic and on the cusp of rapid technological changes, to set down the appropriate norms and collaboration frameworks amongst global and national stakeholders. For instance, HMG can update the Ministry of Defence's definition of "lethal autonomous weapons systems" to be more in line with best international practice (Ord, Mercer, and Dannreuther 2021; Belfield, Jayanti, and Avin 2020) (see also Q23); HMG can also take steps to foster cross-cultural cooperation on issues such as around AI governance (ÓhÉigeartaigh et al. 2020), or to promote productive debates amongst different stakeholders in epistemic community, in ways that ensure productive policies around both existing and future risk policy issues (Stix and Maas 2021).

(3). HMG could set up a new Government Office of Risk Management, headed by a Chief Risk Officer (CRO) with specialist risk management expertise (Ord, Mercer, and Dannreuther 2021), in order to help bring the UK into line with current best practice from industry and elsewhere, while in turn enabling it to become a world leader in addressing global risks domestically, then sharing best practice internationally.

(4). HMG can invest in developing leading sociotechnical solutions for detecting risks and improving resilience, which can be disseminated amongst partners. For instance, in the domain of resilience to AI societal impacts, HMG can ensure it becomes a global leader in developing and disseminating infrastructure to systematically measure and monitor the capabilities and impacts of AI systems (Whittlestone and Clark 2021), enabling the UK to set the standard and best practices for monitoring internationally. This would have multiple benefits to resilience both domestically and internationally.

- Domestically, it would create an international infrastructure that the UK could leverage to better assess its own resilience as well as identify where it itself is a technological leader or where areas within AI would benefit from further measurement or investment.

- Internationally, it would encourage resilience efforts in AI technology in other states by providing an existing blueprint to work from. This would be especially useful for states who do not have the existing resources to engage in this kind of resilience efforts themselves.

(5). One key overarching question that will need input from both the UK and its allies, is the challenge of ensuring sufficient flexibility in institutions (both domestic and international), to ensure governance and resilience responses can evolve as the character (or our understanding) of global risks changes (Maas 2019). In particular, the UK can play a role in organizing the emerging global international 'regime architecture' around new technological risks. In doing so, it should take stock of various trade-offs (such as political power; inclusiveness; adaptiveness; brittleness) in considering the merits of centralizing governance in single institutions. Taking such action is especially urgent in areas such as in the governance of AI technology, where the global

governance architecture is currently in a window of opportunity to set norms and organize cooperation frameworks (Cihon, Maas, and Kemp 2020a; 2020b).

References and further readings

Avin, Shahar, and S. M. Amadae. 2019. "Autonomy and Machine Learning at the Interface of Nuclear Weapons, Computers and People." In *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, edited by V. Boulanin. Stockholm International Peace Research Institute. https://doi.org/10.17863/CAM.44758.

Avin, Shahar, Ross Gruetzemacher, and James Fox. 2020. "Exploring AI Futures Through Role Play." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8–14. New York NY USA: ACM. https://doi.org/10.1145/3375627.3375817.

Avin, Shahar, Lalitha Sundaram, Jessica Whittlestone, Matthijs Maas, and Thomas Hobson. 2021. "Submission of Evidence to The House of Lords Select Committee on Risk Assessment and Risk Planning." Report. https://doi.org/10.17863/CAM.64180.

Avin, Shahar, Bonnie C. Wintle, Julius Weitzdörfer, Seán S. Ó hÉigeartaigh, William J. Sutherland, and Martin J. Rees. 2018. "Classifying Global Catastrophic Risks." *Futures*, Futures of research in catastrophic and existential risk, 102 (September): 20–26. https://doi.org/10.1016/j.futures.2018.02.001.

Baum, Seth D. 2015. "The Far Future Argument for Confronting Catastrophic Threats to Humanity: Practical Significance and Alternatives." *Futures* 72: 86–96.

Baum, Seth, Robert de Neufville, and Anthony Barrett. 2018. "A Model for the Probability of Nuclear War." *Global Catastrophic Risk Institute Working Paper*, no. 18–1. https://doi.org/10.2139/ssrn.3137081.

Beard, S. J., Lauren Holt, Asaf Tzachor, Luke Kemp, Shahar Avin, Phil Torres, and Haydn Belfield. 2021. "Assessing Climate Change's Contribution to Global Catastrophic Risk." *Futures* 127 (March): 102673. https://doi.org/10.1016/j.futures.2020.102673.

Beard, S.J., and Phil Torres. 2020. "Identifying and Assessing the Drivers of Global Catastrophic Risk." Centre for the Study of Existential Risk. https://www.cser.ac.uk/resources/identifying-assessing-drivers/.

Belfield, Haydn, Amritha Jayanti, and Shahar Avin. 2020. "Written Evidence to the UK Parliament Defence Committee's Inquiry on Defence Industrial Policy: Procurement and Prosperity." Cambridge, UK: Centre for the Study of Existential Risk. https://committees.parliament.uk/writtenevidence/4785/default/.

Bostrom, Nick. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9 (1). https://nickbostrom.com/existential/risks.html.

———. 2013. "Existential Risk Prevention as a Global Priority." *Global Policy* 4 (1): 15–31.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. 2018. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." http://arxiv.org/abs/1802.07228.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. 2020. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." *ArXiv:2004.07213 [Cs]*, April. http://arxiv.org/abs/2004.07213.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of Machine Learning Research*, 81:1–15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 2020a. "Should Artificial Intelligence Governance

Be Centralised?: Design Lessons from History." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 228–34. New York NY USA: ACM. https://doi.org/10.1145/3375627.3375857.

———. 2020b. "Fragmentation and the Future: Investigating Architectures for International AI Governance." *Global Policy* 11 (5): 545–56. https://doi.org/10.1111/1758-5899.12890.

Cotton-Barratt, Owen, Max Daniel, and Anders Sandberg. 2020. "Defence in Depth Against Human Extinction: Prevention, Response, Resilience, and Why They All Matter." *Global Policy* 11 (3): 271–82. https://doi.org/10.1111/1758-5899.12786.

Cremer, Carla Zoe, and Jess Whittlestone. 2021. "Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI." *International Journal of Interactive Multimedia and Artificial Intelligence* 6 (5): 100–109.

CSER. 2019. "Novel Practices of Biosecurity Governance (Event)." 2019. https://www.cser.ac.uk/events/novel-practices-biosecurity-governance/.

Currie, Adrian. 2019. "Existential Risk, Creativity & Well-Adapted Science." In *Studies in the History & Philosophy of Science.*, 76:39–48. https://www.sciencedirect.com/science/article/abs/pii/S0039368117303278?via%3Dihub.

Evans, Sam Weiss, Jacob Beal, Kavita Berger, Diederik A. Bleijs, Alessia Cagnetti, Francesca Ceroni, Gerald L. Epstein, et al. 2020. "Embrace Experimentation in Biosecurity Governance." *Science* 368 (6487): 138–40. https://doi.org/10.1126/science.aba2932.

Garcia, Eugenio V. 2020. "Multilateralism and Artificial Intelligence: What Role for the United Nations?" In *The Global Politics of Artificial Intelligence*, edited by Maurizio Tinnirello, 18. Boca Raton: CRC Press. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3779866.

Global Challenges Foundation. 2017. "Global Catastrophic Risks 2017." Global Challenges Foundation. https://www.cser.ac.uk/resources/global-catastrophic-risks-2017/.

Hanea, A. M., M. F. McBride, M. A. Burgman, B. C. Wintle, F. Fidler, L. Flander, C. R. Twardy, B. Manning, and S. Mascaro. 2017. "I Nvestigate D Iscuss E Stimate A Ggregate for Structured Expert Judgement." *International Journal of Forecasting* 33 (1): 267–79. https://doi.org/10.1016/j.ijforecast.2016.02.008.

Hobson, Tom, and Brett Edwards. 2021. "Submission of Evidence to the Foreign Affairs Committee Inquiry on Tech and the Future of UK Foreign Policy." Report. https://doi.org/10.17863/CAM.73736.

Jehn, Florian Ulrich, Marie Schneider, Jason Ruochen Wang, Luke Kemp, and Lutz Breuer. 2021. "Betting on the Best Case: Higher End Warming Is Underrepresented in Research." *Environmental Research Letters*. https://doi.org/10.1088/1748-9326/ac13ef.

Jelinek, Thorsten, Wendell Wallach, and Danil Kerimi. 2020. "Policy Brief: The Creation of a G20 Coordinating Committee for the Governance of Artificial Intelligence." *AI and Ethics*, October. https://doi.org/10.1007/s43681-020-00019-y.

Jones, Natalie. 2017. "Representing Future Generations: Why Politics Needs to Look beyond the Short Term." *In the Long Run* (blog). October 31, 2017. http://www.inthelongrun.org/index.php/articles/article/representing-future-generations-why-politics-needs-to-look-beyond-the-short.

Jones, Natalie, Mark O'Brien, and Thomas Ryan. 2018. "Representation of Future Generations in United Kingdom Policy-Making." *Futures*, Futures of research in catastrophic and existential risk, 102 (September): 153–63. https://doi.org/10.1016/j.futures.2018.01.007.

Kemp, Luke, Laura Adam, Christian R. Boehm, Rainer Breitling, Rocco Casagrande, Malcolm Dando, Appolinaire Djikeng, et al. 2020. "Bioengineering Horizon Scan 2020," May. https://doi.org/10.17863/CAM.52994.

Kemp, Luke, David C. Aldridge, Olaf Booy, Hilary Bower, Des Browne, Mark Burgmann, Austin

Burt, et al. 2021. “80 Questions for UK Biological Security.” *PLOS ONE* 16 (1): e0241190. https://doi.org/10.1371/journal.pone.0241190.

Kemp, Luke, and Catherine Rhodes. 2020. “The Cartography of Global Catastrophic Governance.” Global Challenges Foundation. https://globalchallenges.org/the-cartography-of-global-catastrophic-governance/.

Liu, Hin-Yan, Kristian Cedervall Lauta, and Matthijs Michiel Maas. 2018. “Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research.” *Futures*, Futures of research in catastrophic and existential risk, 102 (September): 6–19. https://doi.org/10.1016/j.futures.2018.04.009.

Liu, Hin-Yan, Kristian Lauta, and Matthijs Maas. 2020. “Apocalypse Now?: Initial Lessons from the Covid-19 Pandemic for the Governance of Existential and Global Catastrophic Risks.” *Journal of International Humanitarian Legal Studies* 1 (aop): 1–16. https://doi.org/10.1163/18781527-01102004.

Liu, Hin-Yan, and Matthijs M. Maas. 2021. “‘Solving for X?’ Towards a Problem-Finding Framework to Ground Long-Term Governance Strategies for Artificial Intelligence.” *Futures* 126 (February): 22. https://doi.org/10.1016/j.futures.2020.102672.

Lord Bird. 2021. *Wellbeing of Future Generations Bill [HL]*. https://bills.parliament.uk/bills/2531.

Maas, Matthijs M. 2018. “Regulating for ‘Normal AI Accidents’: Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment.” In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 223–28. AIES ’18. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3278721.3278766.

———. 2019. “Innovation-Proof Governance for Military AI? How I Learned to Stop Worrying and Love the Bot.” *Journal of International Humanitarian Legal Studies* 10 (1): 129–57. https://doi.org/10.1163/18781527-01001006.

———. 2021. “Aligning AI Regulation to Sociotechnical Change.” In *The Oxford Handbook on AI Governance*, edited by Justin Bullock, Valerie Hudson, Baobao Zhang, Yu-Che Chen, Johannes Himmelreich, Matthew Young, and Anton Korinek. https://papers.ssrn.com/abstract=3871635.

Mani, Lara, Asaf Tzachor, and Paul Cole. 2021. “Global Catastrophic Risk from Lower Magnitude Volcanic Eruptions.” *Nature Communications* 12 (1): 4756. https://doi.org/10.1038/s41467-021-25021-8.

Millett, Piers, Christopher R. Isaac, Irina Rais, and Paul Rutten. 2021. “The Synthetic-Biology Challenges for Biosecurity: Examples from IGEM.” *The Nonproliferation Review* 0 (0): 1–16. https://doi.org/10.1080/10736700.2020.1866884.

Ó HÉigeartaigh, Seán, and Toby Ord. 2021. “Comment on the UK Government’s National AI Strategy.” Centre for the Study of Existential Risk. September 23, 2021. https://www.cser.ac.uk/news/uk-government-set-out-its-national-ai-strategy/.

ÓhÉigeartaigh, Seán S., Jess Whittlestone, Yang Liu, Yi Zeng, and Zhe Liu. 2020. “Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance.” *Philosophy & Technology*, May. https://doi.org/10.1007/s13347-020-00402-x.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. Illustrated Edition. New York: Hachette Books.

Ord, Toby, Angus Mercer, and Sophie Dannreuther. 2021. “Future Proof: The Opportunity to Transform the UK’s Resilience to Extreme Risks.” The Centre for Long-Term Resilience. https://www.longtermresilience.org/futureproof.

Perrow, Charles. 1984. “Normal Accidents: Living with High Risk Technologies.” 1984. http://press.princeton.edu/titles/6596.html.

Quigley, Ellen. 2020. “Universal Ownership in Practice: A Practical Positive Investment

Framework for Asset Owners." SSRN Scholarly Paper ID 3638217. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3638217.

Rhodes, Catherine, Seth Baum, Ariel Conn, Sebastian Farquhar, Malini Mehra, Magali Reghezza, Ama Van Dantzig, Alexandra Wandel, Robert Wiblin, and Kevin Wong. 2016. "Resetting The Frame: Global Challenges Quarterly Risk Report - August 2016." Global Challenges Foundation. https://www.cser.ac.uk/resources/resetting-frame/.

Rios Rojas, Clarissa, Catherine Rhodes, Shahar Avin, Luke Kemp, and Simon Beard. 2021. "Foresight for Unknown, Long-Term and Emerging Risks: Approaches and Recommendations." Report. Evidence for the Risk Assessment and Risk Planning Committee at the House of Lords. https://doi.org/10.17863/CAM.64582.

Rodriguez, Luisa. 2019. "How Likely Is a Nuclear Exchange between the US and Russia?" Rethink Priorities. https://rethinkpriorities.org/publications/how-likely-is-a-nuclear-exchange-between-the-us-and-russia.

Seger, Elizabeth, Shahar Avin, Gavin Pearson, Mark Briers, Seán Ó hÉigeartaigh, and Helena Bacon. 2020. "Tackling Threats to Informed Decisionmaking in Democratic Societies: Promoting Epistemic Security in a Technologically-Advanced World." The Alan Turing Institute. https://www.turing.ac.uk/research/publications/tackling-threats-informed-decision-making-democratic-societies.

Shackelford, Gorm E., Luke Kemp, Catherine Rhodes, Lalitha Sundaram, Seán S. Ó hÉigeartaigh, Simon Beard, Haydn Belfield, et al. 2019. "Accumulating Evidence Using Crowdsourcing and Machine Learning: A Living Bibliography about Existential Risk and Global Catastrophic Risk." *Futures*, December, 102508. https://doi.org/10.1016/j.futures.2019.102508.

Sherwood, S. C., M. J. Webb, J. D. Annan, K. C. Armour, P. M. Forster, J. C. Hargreaves, G. Hegerl, et al. 2020. "An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence." *Reviews of Geophysics* 58 (4): e2019RG000678. https://doi.org/10.1029/2019RG000678.

Stix, Charlotte, and Matthijs M. Maas. 2021. "Bridging the Gap: The Case for an 'Incompletely Theorized Agreement' on AI Policy." *AI and Ethics* 1 (3): 261–71. https://doi.org/10.1007/s43681-020-00037-w.

Sutherland, William, David Aldridge, Tom Hobson, Luke Kemp, Philip Martin, Clarissa Rios Rojas, and Lalitha Sundaram. 2021. "House of Lords Committee on Risk Assessment and Risk Planning: BioRISC Submission on Improving Risk Assessment and Planning." Report. https://doi.org/10.17863/CAM.69982.

Trabucco, Lena. 2020. "AI Partnership for Defense Is a Step in the Right Direction – But Will Face Challenges." *Opinio Juris* (blog). October 5, 2020. http://opiniojuris.org/2020/10/05/ai-partnership-for-defense-is-a-step-in-the-right-direction-but-will-face-challenges/.

Tzachor, Asaf. 2020. "Famine Dynamics: The Self-Undermining Structures of the Global Food System." *GLOBAL RELATIONS FORUM YOUNG ACADEMICS PROGRAM ANALYSIS PAPER SERIES*, 44.

Tzachor, Asaf, Jess Whittlestone, Lalitha Sundaram, and Seán Ó hÉigeartaigh. 2020. "Artificial Intelligence in a Crisis Needs Ethics with Urgency." *Nature Machine Intelligence* 2 (7): 365–66. https://doi.org/10.1038/s42256-020-0195-0.

United Nations. 2021. "Our Common Agenda: Report of the Secretary-General." United Nations. https://www.un.org/en/content/common-agenda-report/.

Whittlestone, Jess, Kai Arulkumaran, and Matthew Crosby. 2021. "The Societal Implications of Deep Reinforcement Learning." *Journal of Artificial Intelligence Research* 70 (March): 1003-1030-1003–30. https://doi.org/10.1613/jair.1.12360.

Whittlestone, Jess, and Jack Clark. 2021. "Why and How Governments Should Monitor AI Development." *ArXiv:2108.12427 [Cs]*, August. http://arxiv.org/abs/2108.12427.

Wiblin, Robert, and Keiran Harris. n.d. "Carl Shulman on the Common-Sense Case for Existential Risk Work and Its Practical Implications." 80,000 Hours Podcast. Accessed October 11, 2021. https://80000hours.org/podcast/episodes/carl-shulman-common-sense-case-existential-risks/.

Wiener, Jonathan B. 2016. "The Tragedy of the Uncommons: On the Politics of Apocalypse." *Global Policy* 7 (S1): 67–80. https://doi.org/10.1111/1758-5899.12319.

Yudkowsky, Eliezer. 2011. "Cognitive Biases Potentially Affecting Judgment of Global Risks." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan Cirkovic. New York: Oxford University Press. https://intelligence.org/files/CognitiveBiases.pdf.