THE CENTRE FOR
LONG-TERM RESILIENCE

# FUTURE PROOF

## THE OPPORTUNITY TO TRANSFORM THE UK'S RESILIENCE TO EXTREME RISKS

June 2021

THE CENTRE FOR
LONG-TERM RESILIENCE

# Lead Authors

**Toby Ord**
Toby is an expert in extreme risks, Senior Research Fellow at Oxford University's Future of Humanity Institute, and author of *The Precipice*. He focuses on the big-picture questions facing humanity: what are the most important issues of our time, and how can we best address them? Toby has advised the United Nations, the World Health Organisation, the World Bank, the World Economic Forum, the UK Prime Minister's Office, Cabinet Office, and Government Office for Science.

**Angus Mercer**
Angus is Chief Executive of the Centre for Long-Term Resilience and a Research Affiliate at the Centre for the Study of Existential Risk at University of Cambridge. He is a former Policy Adviser in the Secretary of State's Office at the UK Department for International Development.

**Sophie Dannreuther**
Sophie is Director of the Centre for Long-Term Resilience and a Research Affiliate at the Centre for the Study of Existential Risk at University of Cambridge. She is a former Strategy Consultant in the Cabinet Office.

Cover photo credit — NASA
Please cite this report as 'Future Proof, The Centre for Long-Term Resilience (2021)'

“This report offers significant recommendations about how the UK can enhance its resilience to extreme risks — one of the most important challenges of our time. It should be required reading for policymakers at all levels.”

***Sir Oliver Letwin, former Chancellor of the Duchy of Lancaster and author of ‘Apocalypse How?’***

“Upgrading our risk management system to include extreme risks is both crucial to our national resilience and a financial no-brainer. The cost of implementing many of these proposed changes would be a rounding error in any Spending Review, but highly beneficial to the UK’s national security. As Covid-19 has shown, it is a false economy not to spend millions now to save billions later.”

***Baroness Pauline Neville-Jones, Member of the Joint Committee on National Security Strategy***

“It is heartening — and timely — to see experts set out a research agenda for improved UK resilience to the extreme threats we face, and offer their support to the Government. This is an offer that policymakers should seize with open arms.”

***Lord Martin Rees, Astronomer Royal and co-founder of the Centre for the Study of Existential Risk at the University of Cambridge***

“Extreme risks present a significant security threat to both current and future generations. I hope that the risk management community in government makes use of the expertise that this report, and its contributors, have to offer.”

***Lord Des Browne, Visiting Researcher at the Centre for the Study of Existential Risk at the University of Cambridge, and Former Secretary of State for Defence***

“Adopting these recommendations would be like taking out an insurance policy against some of the biggest threats we face. It is an idea whose time has come.”

***Wera Hobhouse, MP***

“This report’s key observation — that our national resilience is all-too-vulnerable to a range of extreme risks — is spot on. It is encouraging to see such focus on this critically important issue.”

***Lord Toby Harris, Chair of the National Preparedness Commission***

“In its response to Covid-19, the UK has a once-in-a-generation opportunity to become a world leader in its resilience to biosecurity and other extreme risks. This report provides an excellent roadmap for doing so.”

***Andrew Weber, Senior Fellow, Council on Strategic Risks and Former U.S. Assistant Secretary for Nuclear, Chemical and Biological Defence Programs***

“The prevention of the supreme catastrophe ought to be the paramount object of all endeavour.”

— **Winston Churchill, 1924**

# CONTENTS

# EXECUTIVE SUMMARY

Out of the wreckage of the Second World War, the UK transformed itself. It rebuilt its shattered economy. It founded the NHS. It created national insurance. And it helped establish international institutions like the United Nations so that the world would never endure a tragedy on this scale again.

Human progress doesn't come in straight lines. Instead, there are rare moments where transformative change is possible — where decades' worth of progress can be achieved in a matter of months.

**Such an opportunity for transformative change may now be upon us**. As the UK begins to emerge from Covid-19, which has cost tens of thousands of lives and over £300 billion in 2020 alone,[1] we have a similar opportunity to that which existed in 1945.

While the scale of national tragedy is alive in our minds, the Government must seize this opportunity, and ensure we are much better prepared for the next extreme risk event that will devastate lives and economies on a global scale. **The UK must become a global leader in ensuring long-term resilience to extreme risks**, and keep pace with the significant steps the United States is taking in this area.

We do not know which extreme risk event will come next — it might be another pandemic, or it might be something completely different. But we do know what many of the most extreme risks are, and how best to prepare for them. This report offers a roadmap for how to do just that — **it provides an insurance policy for Britain against the biggest threats we face**.

In this report, leading experts set out the **key extreme risks we face**, analyse the **UK's current level of focus** on each of them, and provide **recommended actions for the Government** to take over the next 12 months. Estimated costs of implementing these recommendations in the upcoming Comprehensive Spending Review are included in this report where possible, and further information is available on request. A great deal of progress can be made with Government investment in the tens of millions — an insignificant sum in the context of any Spending Review.

Contact us to meet the experts, discuss their recommendations, or schedule a policy workshop.

[1] The Office for Budgetary Responsibility's forecast for the 2020/21 UK deficit changed dramatically in the wake of Covid-19, increasing from £55 billion in March 2020 to £394 billion in November 2020. The Institute for Government describes this £339 billion difference as a way of calculating the UK's "cost of coronavirus so far." https://www.instituteforgovernment.org.uk/explainers/cost-coronavirus

# A ROADMAP FOR LONG-TERM RESILIENCE TO EXTREME RISKS

| | Policy area | Current Government Focus[2] | Recommended actions and estimated costs |
|---|---|---|---|
| BIOSECURITY | We are highly vulnerable to biological threats from bioweapons and accidental laboratory leaks, which risk even worse consequences than naturally occurring pandemics like Covid-19.<br>Rapid developments are being made in synthetic biology and biotechnology. These bring great benefits, but also offer harrowing prospects of misuse. | Amber/ Red | 1. Task one body with ensuring preparedness for the full range of biological threats the UK faces (£1 million annually).<br>2. Launch a prize to incentivise development of clinical metagenomics (£3 million one-off cost).<br>3. Establish a Biosecurity Leadership Council and appoint a liaison officer to improve coordination between the biosciences and security communities (£1 million annually).<br>4. Ensure that all DNA synthesis is screened for dangerous pathogens, and regulate DNA synthesis machines. |
| ARTIFICIAL INTELLIGENCE | The capabilities of artificial intelligence (AI) systems have increased significantly in recent years. Though there is still widespread debate and uncertainty, some AI experts have estimated that there is a significant chance that AI ultimately achieves human-level intelligence in the coming decades.<br>A human-level artificial intelligence that is not aligned with the objectives and values of humans poses extreme risk, as does widespread deployment of today's existing capabilities. | Amber | 1. Improve foresight and progress tracking in AI research (£600k annually).<br>2. Bring more technical AI expertise into Government through a scheme equivalent to TechCongress (£1.5 million annually).<br>3. Ensure that the UK Government does not incorporate AI systems into NC3 (nuclear command, control, communications), and lead on establishing this norm internationally.<br>4. Set up throughout-lifetime stress-testing of computer and AI system security (£200k annually).<br>5. Update the Ministry of Defence's definition of "lethal autonomous weapons systems". |
| RISK MANAGEMENT | The UK does reasonably well at risk identification compared to other countries. Nevertheless, there remain a number of technical flaws in the National Security Risk Assessment and the National Risk Register which must be addressed, including difficulties with capturing extreme risks.<br>There also needs to be greater cross-government accountability to ensure that these risks are addressed, that adequate plans are drawn up and that the latest science and research leads to changes in risk policy. | Amber | 1. Improve extreme risk assessment and ownership across government by updating the NSRA, applying 'three lines of defence' model to risk management and installing a Chief Risk Officer (£8.3 million annually).<br>2. Lead the way to ensure global resilience to all extreme risks, not just pandemics, post-Covid-19.<br>3. Normalise red-teaming in Government, including creating a dedicated red team to conduct frequent scenario exercises (£800k annually).<br>4. Revise the Green Book's discount rate and ensure the Treasury adopts key recommendations on intergenerational fairness.<br>5. Establish a new Defence Software Safety Authority as a sub-agency of the Defence Safety Authority, to protect UK defence systems from emerging threats (£5 million annually).<br>6. Fund a comprehensive evaluation of the actions required to increase the resilience of the electrical grid. |
| R&D | The UK has world-leading strengths in academic research in areas relevant to long-term and extreme risk such as technology regulation, AI, and biotechnology.<br>Nevertheless, research on extreme risks remains significantly underfunded given its importance. Total funding on AI safety, for example, is significantly smaller than the total funding going into private investment to accelerate its capabilities. | Green/ Amber | 1. Create 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).<br>2. Invest further in AI safety R&D.<br>3. Invest further in applied biosecurity R&D.<br>4. Invest further in improving long-term forecasting and planning. |

[2] We rate risks as green when, comparatively, they are already an extensive focus of Government policy, amber when there is some (but limited) policy focus, and red when they are almost entirely neglected.

# INTRODUCTION

# Extreme risks — the defining challenge of our time

At several points in humanity's long history, there have been great transitions in human affairs that accelerated our progress and shaped everything that would follow.

Ten thousand years ago, we had the Agricultural Revolution. Farming could support 100 times as many people on the same piece of land, making much wider cooperation possible. We developed writing, mathematics, engineering, and law. We established civilization.

Four hundred years ago, we had the Scientific Revolution. The scientific method replaced deference to perceived authorities, with careful observation of the natural world, seeking simple and testable explanations for what we saw.

Two hundred years ago, we had the Industrial Revolution. This was made possible by the discovery of immense reserves of energy in the form of fossil fuels. Productivity and prosperity accelerated, giving rise to the modern era of sustained growth.

But there has recently been another transition more important than any that has come before. **With the detonation of the first atomic bomb, a new age of humanity began**. We finally reached the threshold where we might be able to destroy ourselves — the first point when the threat to humanity from within exceeded the threats from the natural world.

These threats to humanity — which we in this report refer to as 'extreme risks' — define our time.

**We are currently living with an unsustainably high level of extreme risk.**

With the continued acceleration of technology, and without serious efforts to boost our resilience to these risks, there is strong reason to believe the risks will only continue to grow.

### What are extreme risks? [1 2 3]

**Extreme risks** are high-impact threats which have a global reach, and include both global catastrophic risks and existential risks. The nature of extreme risks makes them difficult to assess and address, compared to more regularly occurring events, such as floods, earthquakes, or terrorist attacks.

**Global catastrophic risks** are those which could lead to significant loss of life or value across the world. For a rough sense of scale, many research papers refer to risks of disasters that result in a loss of 10% or more of the human population.

**Existential risks** are those which could lead to the premature extinction of humanity or the permanent and drastic destruction of its potential. Unlike global catastrophic risks, existential risk scenarios do not allow for meaningful recovery and are, by definition, unprecedented in human history. In his recent book, The Precipice, Toby Ord, one of this report's authors, estimates the likelihood of the world experiencing an existential catastrophe over the next one hundred years at one in six.[4]

# The post-Covid-19 opportunity

Covid-19 has given us a sense of the devastating impact that extreme risks would have on our health and economy. In any given year, the likelihood of an extreme risk materialising is relatively small, but the odds that we — or our children and grandchildren — will face one of them are uncomfortably high.

We do not know which extreme risk event will come next — it might be another pandemic, or it might be something completely different. But we do know what many of the most extreme risks are, and how best to prepare for them.

**The cost of better preparation for extreme risks pales in comparison to the cost of Covid-19 so far, which has been estimated at over £300 billion in 2020 alone.[5] Government spending on extreme risk resilience is the best kind of investment — one**

[1] We have footnoted all papers and reports which we cite and, for the convenience of online readers, also included hyperlinks in the body of the report for key resources.

[2] https://www.repository.cam.ac.uk/handle/1810/317070;jsessionid=FC51BC52C0FFF10D3C39534F57AB2DF1

[3] https://www.cser.ac.uk/resources/global-catastrophic-risks-2017/

[4] https://www.bloomsbury.com/uk/the-precipice-9781526600219/

[5] The Office for Budgetary Responsibility's forecast for the 2020/21 UK deficit changed dramatically in the wake of Covid-19, increasing from £55 billion in March 2020 to £394 billion in November 2020. The Institute for Government describes this £339 billion difference as a way of calculating the UK's "cost of coronavirus so far." https://www.instituteforgovernment.org.uk/explainers/cost-coronavirus

**which will save not only countless lives, but hundreds of billions of pounds.**[6]

To do justice to the seriousness of these risks and the unsustainably high level of risk we currently face, the Government must go beyond simply 'fighting the last war' and focusing solely on better preparedness for naturally occurring pandemics like Covid-19. **It needs to transform the UK's resilience to extreme risks across the board.**

The UK is already an academic leader in the field of extreme risks, with world-class organisations such as Oxford University's Future of Humanity Institute[7] and Oxford Martin School[8], Cambridge University's Centre for the Study of Existential Risk[9] and Centre for Risk Studies[10], and Warwick University's Anthropogenic Global Catastrophic Risks project.[11] **The UK Government now has the perfect opportunity to match academic excellence in extreme risks with policy leadership.**

Just as Covid-19 triggers an immune response in each individual, protecting them from reinfection, so the pandemic has triggered a social immune response across the UK, where there is public will to prevent the next extreme risk. But like the individual immune response, this social immune response will fade over time. Before it does, we need to seize this opportunity to put in place lasting protections to safeguard the country from extreme risks — both at a risk-specific level and at a systemic level.

Encouragingly, the new Integrated Review highlights the need for "low-probability, catastrophic-impact threats"[12] to be at the heart of the UK's efforts to build national resilience. It also commits the UK to developing a "comprehensive national resilience strategy in 2021 to prevent, prepare for, respond to and recover from risks."[13]

**The next step is to set out exactly how the Government can embed extreme risks into its resilience planning — and into its upcoming AI strategy and biosecurity review.** This report provides a roadmap for how to do so.

We welcome engagement with the Government on any aspect of this report, via info@longtermresilience.org.

[6] https://www.ft.com/content/8f7ac2da-3902-11e9-b72b-2c7f526ca5d0. Also see: Calculating Catastrophe, Gordon Woo, 2011

[7] https://www.fhi.ox.ac.uk/

[8] https://www.oxfordmartin.ox.ac.uk/

[9] https://www.cser.ac.uk/

[10] https://www.jbs.cam.ac.uk/faculty-research/centres/risk/

[11] https://warwick.ac.uk/fac/soc/pais/research/researchcentres/ierg/research/agcr

[12] Integrated Review, pp 88-89: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/969402/The_Integrated_Review_of_Security__Defence__Development_and_Foreign_Policy.pdf

[13] Integrated Review, pp 88-89

# OVERVIEW

Firstly, we provide an **overview of two of the most extreme risks** the UK faces in the twenty-first century — those relating to **biosecurity** and **artificial intelligence**. We assess to what extent they are a focus of current Government policy. We also provide recommended actions for the Government to take in the next twelve months to help safeguard the UK against these extreme risks.

These recommendations are by no means comprehensive — rather, they focus on issue areas that are relatively neglected in the existing policy conversation and the key actions to take within those issue areas. While **climate change** and **nuclear security** are perhaps the best-known extreme risks, this report focuses on biosecurity and AI on the basis that the extreme risks they pose currently receive much less attention in policy circles.

We then turn our attention to **two cross-cutting policy recommendations** to improve resilience to extreme risks at a systemic level.

The first cross-cutting recommendation is to **improve the UK Government's process for managing extreme risks**. The second is to **increase funding for research into extreme risks** and cement the UK's leadership in this area. This would ensure that we better understand the nature of the threats we face and how best to deal with them.

We conclude with a short section setting out **further information on extreme risks**.

Estimated costs of implementing these recommendations in the upcoming Comprehensive Spending Review are included in this report where possible, and further information is available on request.

# THE EXTREME RISKS

## OVERVIEW AND RECOMMENDATIONS

# Issue-specific policy recommendations

# BIOSECURITY

| CURRENT GOVERNMENT FOCUS | AMBER / RED |
|---|---|

## Section Contributors

Dr Cassidy Nelson
Acting Co-Lead, Biosecurity Research Group, Future of Humanity Institute, University of Oxford and Board Member of the Centre for Long-Term Resilience

Dr Gregory Lewis
Acting Co-Lead, Biosecurity Research Group, Future of Humanity Institute, University of Oxford

Dr Piers Millett
Senior Research Fellow, Future of Humanity Institute, University of Oxford; consultant to the World Health Organisation

## Summary of the risk

**The tragic events of the Covid-19 pandemic have highlighted the need for the UK to transform its level of preparedness against biological threats.**

But in our response, we cannot simply ‘fight the last war’ and focus solely on preparedness for future naturally occurring pandemics. We know from national security risk assessments and the UK Biological Security Strategy that we remain vulnerable to accidental and deliberate biological threats which risk even worse consequences than Covid-19.[1] Even more concerning are the very rapid developments that are being made in synthetic biology and biotechnology, which offer harrowing prospects of misuse.

Fortunately, there are concrete steps that can be taken to ensure that the UK leads the world in efforts to mitigate these risks and prepare for all forms of pandemics.

## Current level of Government focus

Responding to Covid-19 has rightly been the Government's core priority since early 2020. Alongside its day-to-day response to the pandemic, there have been several changes to the machinery of government to enhance pandemic preparedness going forward — including the creation of the UK Health Security Agency (formerly the National Institute of Health Protection).

There are encouraging signs that the long-term protection of the UK's biological security is now higher on the political agenda than it was prior to Covid-19. The Government has committed to holding a public inquiry into Covid-19, which is due to start next year. Further, the Joint Committee on National Security Strategy's December 2020 report has called on the Government to "plan for unexpected futures", and recommended the creation of a "dedicated national centre for biosecurity".[2] The Government's response, published in March 2021, did not accept this recommendation specifically, but did recognise the "need for a resilient and enduring approach to biological security".[3] It also announced that urgent work is ongoing to identify where responsibility for biosecurity should sit long term. The Government has also announced a forthcoming review of the UK's Biological Security Strategy, though no specific date has been set for this.

In the Integrated Review, we learned that the cross-government approach to biosecurity is being "reviewed and reinforced", and we anticipate further changes to the machinery of government. The Government committed to review its national stockpile of clinical countermeasures and consumables, such as personal protective equipment, expanded testing capability and laboratory equipment. It also set out the UK's ambitions for biosecurity, including international leadership; for instance, by increasing funding for the World Health Organisation and reducing priority vaccine development and deployment time to 100 days.[4]

These are important steps to take. **Yet much more needs to be done now to protect**

[1] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/730213/2018_UK_Biological_Security_Strategy.pdf

[2] https://publications.parliament.uk/pa/jt5801/jtselect/jtnatsec/611/61110.htm#_idTextAnchor076

[3] https://committees.parliament.uk/publications/4870/documents/49008/default/

[4] Integrated Review, p93-94

**the UK from a broader range of biosecurity threats. We currently run the risk of simply 'fighting the last war' and preparing for the previous pandemic instead of what might come next.** The UK has made this mistake in the past: pandemic preparedness planning prior to Covid-19 presumed influenza, and was wrong-footed when confronted by a coronavirus. The UK's pandemic strategy did not include any plans for a lockdown, despite this having become one of the dominant strategies for responding to Covid-19.

The threat of human-caused pandemics — by either accident or attack — is growing in step with the rapid march of biotechnological progress. This threat is even more of a concern than naturally-arising pandemics over this century, and therefore warrants similar attention from the Government. Though the Integrated Review indicates that an improved approach to pandemic preparedness is in train, it is both striking and disappointing that these human-caused disasters, such as accidental laboratory outbreaks and deliberate bioweapons, are barely mentioned in the review, while much more emphasis is placed on nuclear weapons.

The UK has a poor track record of laboratory accidents. The 2007 foot and mouth disease outbreak was caused when it escaped from what was supposed to be the most secure level of lab in the whole world. Likewise the final victims of smallpox were in the UK, when it escaped from an insecure lab. Building a better safety culture in biotechnology is essential, and requires knowledge of the rate and causes of laboratory accidents. Yet these are currently poorly understood.

The UK rivals the United States in terms of its bioscience capability, but we currently do not make the most of the expertise we have. Unlike in the United States, there are (to our knowledge) no permanent biosecurity experts on the UK National Security Council. The UK also has fewer senior Government officials with a technical background in biosecurity compared with the United States, and less of an established health security community.

## Key biosecurity policy recommendations

### 1. Task one body with ensuring preparedness for the full range of biological threats the UK faces (estimated cost: £1 million annually)

**Why this matters:** Biological security is critical to the UK's national security. Covid-19 has shown that urgent reforms are needed, but these reforms must go beyond preparation for naturally occurring pandemics and avoid the well-known pattern of 'panic and neglect' in global health security. One designated body should have accountability to lead these efforts.

We do not know when the next major pandemic will hit the UK — it could be decades away. This is plenty of time to sink into complacency or be distracted by other short-term crises.

Ideally, **a new National Centre for Biosecurity would be tasked with prevention of, and preparedness for, future large-scale and high-priority biological threats faced by the UK, regardless of their origin**. It would provide strategic direction over policy and technical solutions, along with national-level coordination and integration of expertise from a wide range of disciplines to prevent and increase preparedness for biological threats. It would complement the proposed new research agency, ARIA, by fulfilling a think-tank-like function that delivers insights on new areas of opportunity and promising technical solutions.[5] It would also usefully draw on the latest foresight work in the areas of biosecurity and bioengineering.

In short, such a centre's mission would be to ensure the biological security of the UK.

To achieve this, it would focus on the four areas of highest priority:

1. **Preventing and countering the threat of biological weapons from both state and non-state actors**, treating them as a comparable security challenge to nuclear weapons;
2. **Developing effective defences to biological threats**, helping bring horizon technologies (especially pathogen-blind diagnostics) to technical readiness;
3. **Promoting responsible biotechnology development** across the world; and
4. **Developing talent and collaboration** across the UK biosecurity community, cementing the UK as a world leader in safe and responsible science and innovation.

For further information, see Oxford's Future of Humanity Institute's working paper on the proposed Centre.

It may be that the recently announced UK Health Security Agency takes on many or all of these priority areas, which would be another viable solution. What matters most is that UK biosecurity focuses not just on immediate threats, but also on prevention of and preparedness for the full range of biological threats we face.

[5] https://policyexchange.org.uk/wp-content/uploads/Visions-of-Arpa.pdf

## 2. Launch a prize to incentivise the development of clinical metagenomics (estimated one-off cost: £3 million)

**Why this matters:** Clinical metagenomics has the potential to identify new, unexpected pathogens in the first few infected patients, rather than months later. This would be game-changing, as it would allow for a much earlier, targeted response to an outbreak of an unknown virus.

Metagenomic sequencing takes a sample from a patient, sequences the DNA of all organisms in it, and automatically compares these to a known database of pathogens, finding the closest matches. With coming technologies, this will likely be affordable to the point where doctors can routinely send in a sample from any case in the UK that they cannot diagnose with standard techniques.

It would cost approximately £100 per sample for the metagenomic sequencing to be performed and to identify the closest matches. This capability could be integrated into the existing public health laboratory network, or take place in a central laboratory. This would be extremely helpful for both regular diagnoses and for novel pathogens. In the case of Covid-19, such technology would have immediately shown that the closest match was SARS, but that it was sufficiently different to be a novel SARS-like pathogen. Put simply: if metagenomic sequencing had been widely available at the start of 2020, the trajectory of Covid-19 may well have been very different.

**A prize of around £3 million could be awarded to the first group that can develop an interface that takes raw metagenomics data and turns out potentially clinically relevant results.**

Prize challenges are a wonderful innovation. They increase the number of minds tackling a particular problem without having to predict which team or approach is most likely to succeed. They are also an efficient means of identifying talented individuals and teams, who can be seconded for future programs. They tend to be about 10 times more cost-effective than traditional research projects, and to prompt a higher degree of spending on research by the contestants. This is both because competitors tend to overestimate the probability of winning, and because they tend to place a significant value on the reputational reward for winning or being shortlisted.

## 3. Establish a Biosecurity Leadership Council and appoint a liaison officer to improve coordination between the biosciences and security communities (estimated cost: £1 million annually)

**Why this matters:** Biotechnology is often 'dual use', meaning that advances can be used for harm as well as good. For example, an individual could build live viruses 'from scratch' for legitimate research, but also to conduct bioterrorism. This makes biotechnology a highly challenging area to navigate, and one which requires a far greater degree of coordination than that which currently exists.

**This proposed new Biosecurity Leadership Council's role would be to develop biosecurity policy through collaboration** between the Government, academia, business, and other relevant stakeholders. It would provide an official channel of coordination to ensure that there is dialogue between these stakeholders. The UK Synthetic Biology Leadership Council is a possible model.[6]

The Council would help ensure adequate resourcing, both in terms of funding and expertise, and a liaison officer would improve coordination between the biosciences and security communities. This officer would provide advice and build relationships across Government, law enforcement, intelligence agencies, academic researchers, and private sector researchers. An equivalent position already exists in the United States.

## 4. Ensure that all DNA synthesis is screened for dangerous pathogens and regulate DNA synthesis machines

**Why this matters:** Malicious biological threats warrant equal concern to natural pandemics, but they are receiving considerably less policy attention following the Covid-19 outbreak. As the Secure DNA Project notes: "A world in which many thousands of people have access to powerful and potentially dangerous biotechnologies is unlikely to flourish... Historical pandemics killed tens of millions of people, and engineered agents could be even more destructive."[7]

Unless active controls are present, gene synthesis machines provide a way for individuals to get their hands on dangerous or novel pathogens. The export of desktop versions of these machines already require a license on biosecurity grounds. **Gene**

[6] https://www.gov.uk/government/groups/synthetic-biology-leadership-council

[7] https://www.securedna.org/main-en

**synthesis companies should therefore be required to adhere to biosecurity guidelines for screening DNA orders for dangerous pathogens**, such as those released by the Secure DNA Project.

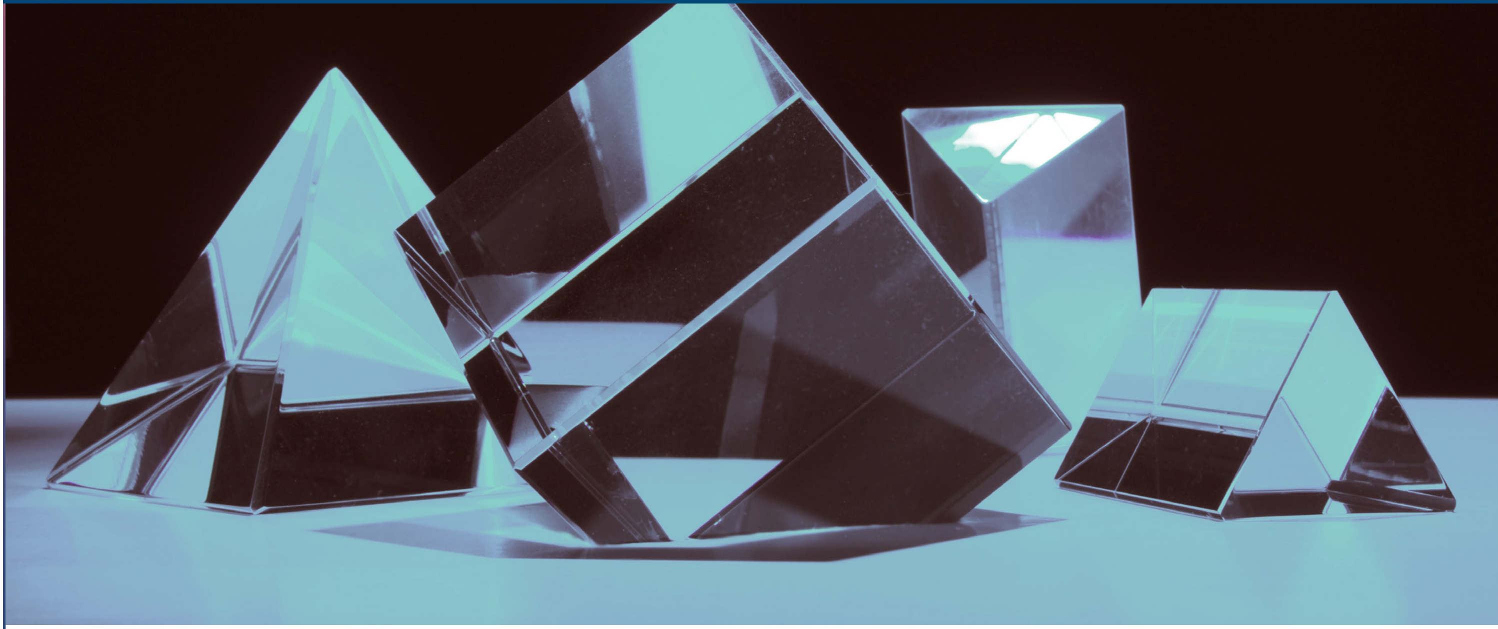
These guidelines go beyond the most commonly used International Gene Synthesis Consortium protocol to reflect rapid advancements in the field and current technological capabilities.[8] Imported DNA orders should adhere to the same biosecurity screening guidelines, and the UK should be a leader in the international community on further improving these initiatives and make screening more universal and robust.
If full coverage cannot be achieved through self-regulation by gene synthesis companies, the UK should take a leading role in pushing for domestic and international regulation in this area.

The UK is already at the forefront of 'biofoundries' which enable the rapid design, construction, and testing of genetically reprogrammed organisms for biotechnology applications and research.[9] The founding members of the Global Biofoundries Alliance are leading UK synthetic biologists. This is therefore a natural leadership area for the UK to adopt.

[8] https://genesynthesisconsortium.org/wp-content/uploads/IGSCHarmonizedProtocol11-21-17.pdf

[9] https://biofoundries.org/about-the-gba

# ARTIFICIAL INTELLIGENCE



| | |
|---|---|
| CURRENT GOVERNMENT FOCUS | AMBER |

## Section Contributors

Dr Jess Whittlestone

Senior Research Fellow, AI Ethics and Policy, Centre for the Study of Existential Risk, University of Cambridge

Dr Jade Leung

Governance and Policy Advisor, OpenAI and Board Member of the Centre for Long-Term Resilience

Markus Anderljung

Project Manager, Operation & Policy Engagement, Centre for the Governance of AI, Future of Humanity Institute, University of Oxford

## Summary of the risk

As anyone who watched DeepMind's AlphaGo machine defeat the Go world champion in 2016 will know, the capabilities of AI systems have increased significantly in recent years.

AI progress is rapid, unprecedented, and transformative. But what is perhaps most striking is that we do not even need to look to what AI might do in the future to see the extreme risks it poses. **The widespread deployment of even our current AI capabilities could lead or contribute to extreme risks**. These risks can be subdivided into three categories:

- **Misuse risks** result from using AI in an unethical manner. For example, realistic but synthetic images or videos generated using AI can be used for activities like political disruption. And since AI capabilities are so broadly applicable, they may often have unforeseen harmful uses.
- **Accident risks** involve harms arising from AI systems behaving in unintended ways, such as a self-driving car collision caused by a car failing to understand its environment. The more AI is integrated into safety-critical systems such as vehicles and energy systems, the higher the stakes are for accident risks.[10]
- **Structural risks** will likely appear in the longer term. They involve the increasing use of AI to change political, social and economic structures and incentives. Widespread use of AI systems could exacerbate existing inequalities by locking in patterns of historical discrimination, provoking rapid and wide-scale unemployment, or dramatically concentrating power in the hands of a few companies and states.[11]

Strikingly, when hundreds of scientists were surveyed about when they believed AI would reach general human-level intelligence, the median response was around 35 years from now.[12] And, of course, there is no reason to think that AI would stop at human levels of intelligence. Such technology will likely be highly beneficial to humanity in countless ways, but a human-level AI that is not aligned with human objectives and values also constitutes an extreme risk.

Policymakers need to act now to ensure that AI is developed, used and governed responsibly, as an asset rather than a threat to human potential.

## Current level of Government focus

The UK Government has so far demonstrated laudable proactivity in the field of AI policy and governance, establishing the Office for Artificial Intelligence, the Centre for Data Ethics and Innovation (CDEI), the AI Council, NHSX, the Regulatory Horizons Council, and becoming a member of the Global Partnership on AI.

The Integrated Review, though surprisingly light on AI, does hint at significant forthcoming activity: the creation of a new AI strategy and a "centre to accelerate adoption of this technology across the full spectrum of our capabilities and activities."[13]

[10] https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure

[11] https://ainowinstitute.org/discriminatingsystems.pdf

[12] See Diagram 2 of https://arxiv.org/pdf/1705.08807.pdf

[13] Integrated Review: pp 73 - 74

This planned activity makes the prioritisation of AI ethics vitally important. It is therefore heartening to see GCHQ's recent report confirm that it has commissioned work from the Alan Turing Institute to study the implications of AI for ethics.[14] The Government's published guidance on AI ethics and safety is another encouraging development, as is the National Data Strategy's emphasis on using data in an ethical and responsible way.[15]

Initiatives from regulators such as the AI Auditing Framework and Project ExplAIn are also noteworthy examples of AI safety being taken seriously, at least with regards to accuracy, fairness and transparency.[16] The AI Council's AI roadmap puts a significant amount of emphasis on the importance of good governance and regulation to build public trust in AI, noting that "making the UK the best place in the world to research and use AI requires it to be world-leading in the provision of responsive regulation and governance". Finally, the Turing AI Fellowship programme is worth highlighting, as it enables further research into AI safety and robustness.[17]

These initiatives represent an encouraging start. **However, the UK's efforts on AI safety remain incomplete in some areas and embryonic in others**. While risks of bias, transparency and accountability are frequently highlighted by policymakers, currently very few resources are invested to foresee, understand and mitigate the full spectrum of risks that AI safety researchers are concerned about.

## Key AI policy recommendations

### 1. Improve foresight and progress tracking in AI research (estimated cost: £600k annually)

**Why this matters:** AI capabilities and their potential applications in society are growing fast. The UK risks falling behind and taking an overly reactive approach if it does not develop its capability to monitor AI progress.

**The UK Government should establish its own capacity to anticipate and monitor AI progress and its implications for society.**

[14] https://www.gchq.gov.uk/files/GCHQAIPaper.pdf

[15] https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety
https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy#data-3-5

[16] https://ico.org.uk/about-the-ico/news-and-events/ai-auditing-framework/#!; https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf

[17] https://www.turing.ac.uk/people/researchers/ai-fellows

Such monitoring will help inform future AI policy and regulation to help manage the societal implications of AI. It will also mitigate the risks of increasingly widely deployed AI applications in critical areas. The work would complement and work closely alongside initiatives like the Centre for Data Ethics and Innovation's AI barometer, the OECD AI Observatory[18] and Stanford's AI Index initiative.[19]

We recommend funding:

- **Greater capacity inside Government for establishing metrics and mechanisms to assess progress in AI, its applications and impacts on society.** This could be achieved either by a new body (which could be housed within CDEI or the Alan Turing Institute), or by substantially increasing the scope and funding for existing initiatives, such as CDEI's existing AI-monitoring capacity.
- **Research projects** in AI foresight and progress tracking, to be awarded through the mechanisms set out above.

## 2. Bring more technical AI expertise into Government through a scheme equivalent to TechCongress (estimated cost: £1.5 million annually)

**Why this matters:** As AI systems become more capable, their impacts will grow and become more cross-cutting, increasing the need for technical expertise across the UK Government. Such expertise is currently sorely lacking.

The Government is steadily bolstering the supply of tech talent within Government through initiatives such as the Number 10 Innovation Fellowship, the Data Science Graduation Programme and the Data Science Campus Faculty. However, there is more it could do to plug the current gap in technical AI expertise, including:

- **Setting up a TechCongress-equivalent scheme** (potentially as part of the cross-government Open Innovation Team) aimed at enabling the UK Government to recruit and gain access to AI expertise in fields like AI governance and ethics.[20] The scheme should place experts in Parliament and also embed them within the Civil Service.
- **Creating specific technical roles in key departments**, including the Ministry of Defence, the Information Commissioner's Office, the Home Office and the Department for Business, Energy, and Industrial Strategy. These roles would be

[18] https://oecd.ai/

[19] https://hai.stanford.edu/research/ai-index-2019

[20] https://www.techcongress.io/

targeted at experts in AI, machine learning and cyber security, and would focus on assuring the safety and security of AI systems that are deployed in specific sectors, particularly those that serve critical functions to society (such as critical infrastructure, law enforcement, finance, and defence).

- **Setting up a fund** that agencies can apply for to cover the salaries of additional technical experts where necessary.
- **Providing funding for existing civil servants to develop training and expertise in AI and machine learning**. The Treasury currently provides scholarships for civil servants to study economics; an equivalent scheme should be devised for AI.

## 3. Ensure that the UK Government does not incorporate AI systems into NC3 (nuclear command, control, communications), and that the UK leads on establishing this norm internationally

**Why this matters:** As evidenced by the sobering history of nuclear near misses, introducing AI systems (or automation) into NC3 increases the risk of an accidental launch, without proportional benefits.

We recommend that an appropriate body or individual at the **Ministry of Defence investigates the process that the UK would need to undertake to make a credible commitment that it will not incorporate AI systems into NC3**, and then makes this commitment.

We also recommend avoiding (and publicly committing to avoid) cyber operations — including intelligence-gathering operations — that target the NC3 of Nuclear Proliferation Treaty signatories.[21] A recent report by the US National Security Commission on Artificial Intelligence made a similar recommendation.[22]

We further recommend that the UK advocates for this policy norm internationally; for example, by establishing a multilateral agreement to this effect.

[21] https://www.un.org/disarmament/wmd/nuclear/npt/

[22] p98, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf: “Clearly and publicly affirm existing U.S. policy that only human beings can authorize employment of nuclear weapons, and seek similar commitments from Russia and China.”

## 4. Set up throughout-lifetime stress-testing of computer and AI system safety and security (estimated cost: £200k annually)

**Why this matters:** For the UK's national security, it is important to stress-test computer systems thoroughly to test their resilience and identify flaws. To do this well, there must be an incentive structure to point out problems rather than underplay them.

The Government announced in November 2020 that it would create a new AI agency and re-orient its defence capability towards emerging threats.[23] It is vitally important that any new AI-related bodies prioritise the safe development of AI, and in particular that they set up throughout-lifetime stress-testing of computer and AI system safety and security.

**Stress-testing allows systems to be assessed for flaws and vulnerabilities before they can be exploited by adversaries, or before accidents involving new systems occur.** This is particularly important given the Government's decision to incorporate AI into its defence capabilities — for example through the Royal Air Force's AI and drone technology.

We recommend that computer and AI systems be stress-tested during development, testing, training, early deployment, at regular intervals and before retirement of relevant systems. The Government should also have dedicated 'white hats' stress test their systems by attempting to compromise their software and hardware vulnerabilities, through social engineering and by designing adversarial environments. We also recommend making adversarial testing and red teaming part of military exercises.

## 5. Update the Ministry of Defence's definition of 'lethal autonomous weapons systems'

**Why this matters:** The UK should take a leading role in the ethics and implications of lethal autonomous weapons and promote international dialogue. The Ministry of Defence's definition of 'lethal autonomous weapons systems' differs from that of other governments, making international dialogue more difficult.

[23] This was announced as part of the UK's November 2020 announcement that it would spend £16.5 billion on defence spending: https://www.bbc.co.uk/news/uk-54988870

Within the wide set of defence systems that integrate increasingly capable AI and machine learning, particular attention is rightly paid to lethal autonomous weapons systems. These systems raise important questions of ethics and international humanitarian law, and are the focus of arms-control negotiations at the United Nations.

The Ministry of Defence's current definition of 'lethal autonomous weapons systems' is quite different from that used by many other nations. It defines an "autonomous" system as "capable of understanding higher-level intent and direction", "capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control" and "able to take appropriate action to bring about a desired state."[24]

**This definition is a very high bar to reach — almost human-level intelligence — and is so high as to be almost meaningless.** No system currently under research or development would be capable of meeting this definition. The UK should adopt a definition similar to that used by other governments and international organisations in order to improve its ability to consider and protect against foreseeable risks associated with these systems, and to act as a global leader in setting international standards for this emerging technology.

[24] p13, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/673940/doctrine_uk_uas_jdp_0_30_2.pdf

# Cross-cutting policy recommendations

# IMPROVING THE UK GOVERNMENT'S RISK MANAGEMENT PROCESSES

| | |
|---|---|
| CURRENT GOVERNMENT FOCUS | AMBER |

## Section Contributor

Sam Hilton
Research Affiliate, Centre for the Study of Existential Risk, University of Cambridge

## Current level of Government focus

The UK does reasonably well at risk identification compared to other countries. The National Security Risk Assessment (NSRA) and the National Risk Register (NRR) provide detailed analysis of the risks that the UK faces, and there is good use of horizon-scanning and foresight capacity.

Nevertheless, there remain a number of technical shortcomings with the NSRA which need to be addressed in the review of its methodology which is currently taking place:

- **The NSRA does not sufficiently capture high uncertainty risks like extreme risks**. Future risks and low-probability risks tend to be excluded from its assessment.

- **The way the NSRA delineates risks and highlights uncertainty is potentially confusing for decision makers** — in particular, its unclear definition of "reasonable worst-case scenario".
- **The NSRA should include a robust vulnerability assessment**, asking in particular how effective and well-documented existing mitigations and crisis management capabilities are.

**The shortcomings of the current approach to assessment of extreme risks have been made clear by Covid-19**. The most recent National Risk Register failed to adequately estimate the scale of the crisis, estimating that emerging non-influenza infectious diseases could lead "to up to 100 fatalities". This was clearly a very large underestimation of the scale of the current crisis, as well as being out of line with academic evidence and other risk reports. It led to Government plans which focused too heavily on influenza rather than other diseases.

Beyond the NSRA, there appears to be **limited cross-government accountability to ensure that risks are mitigated, that adequate plans are drawn up and that the latest science and research leads to changes in risk policy**. For instance, it is notable that the UK's pandemic influenza strategy did not make any plans for a lockdown, despite this being one of the dominant strategies for responding to Covid-19.[25]

**That being said, the Government has indicated an encouraging will to learn lessons for UK risk management in the aftermath of Covid-19**. A new Cabinet Office National Situation Centre aims to improve 'situational awareness for crises and national security issues' by collating data and insights. And the new Integrated Review leaves readers in no doubt that building resilience to risk is now a top national priority. Ensuring "a more robust position on security and resilience" is defined as one of the four main components of Global Britain.

The Integrated Review contained a range of promising announcements:

- A **new approach to preparedness and response to risks**, which fully recognises that natural hazards and other extreme risks can cause as much disruption to the UK's core interests as security threats.
- An **ongoing external review of the NSRA** and its underlying methodology;
- A **comprehensive national resilience strategy**, with a commitment to include threats and hazards and all kinds of risk including the possibility of "low-probability, catastrophic-impact events".
- A new **Performance and Planning Framework** and an **Evaluation Taskforce** (to drive change across Government and assess impact as per the Integrated Review's recommendations).

[25] Samuel Hilton and Caroline Baylon's recent paper on risk management provides further details on the state of UK risk management, the lessons we can learn from Covid-19, and ensuring the UK is prepared for the next disaster. https://www.cser.ac.uk/resources/risk-management-uk/

We welcome this progress, and hope that the Government will continue to engage academic and risk management experts as they develop this new risk management infrastructure.

We also hope that the Integrated Review's approach of setting out a ten-year strategy based on broad public consultation becomes a more regular part of the policymaking process. There should be a legal requirement for each new government to confirm their long-term vision and strategy in this way, including their strategy to build resilience to extreme risks. Governments also need to be held accountable for delivering on these strategies, through mechanisms such as annual reviews, clearly-defined performance indicators, and the oversight of a designated Minister and Select Committee.

## Key risk management policy recommendations

### 1. Improve extreme risk assessment and ownership across Government by updating the NSRA, applying the 'three lines of defence' model to risk management and installing a Chief Risk Officer (estimated cost: £8.3 million annually)

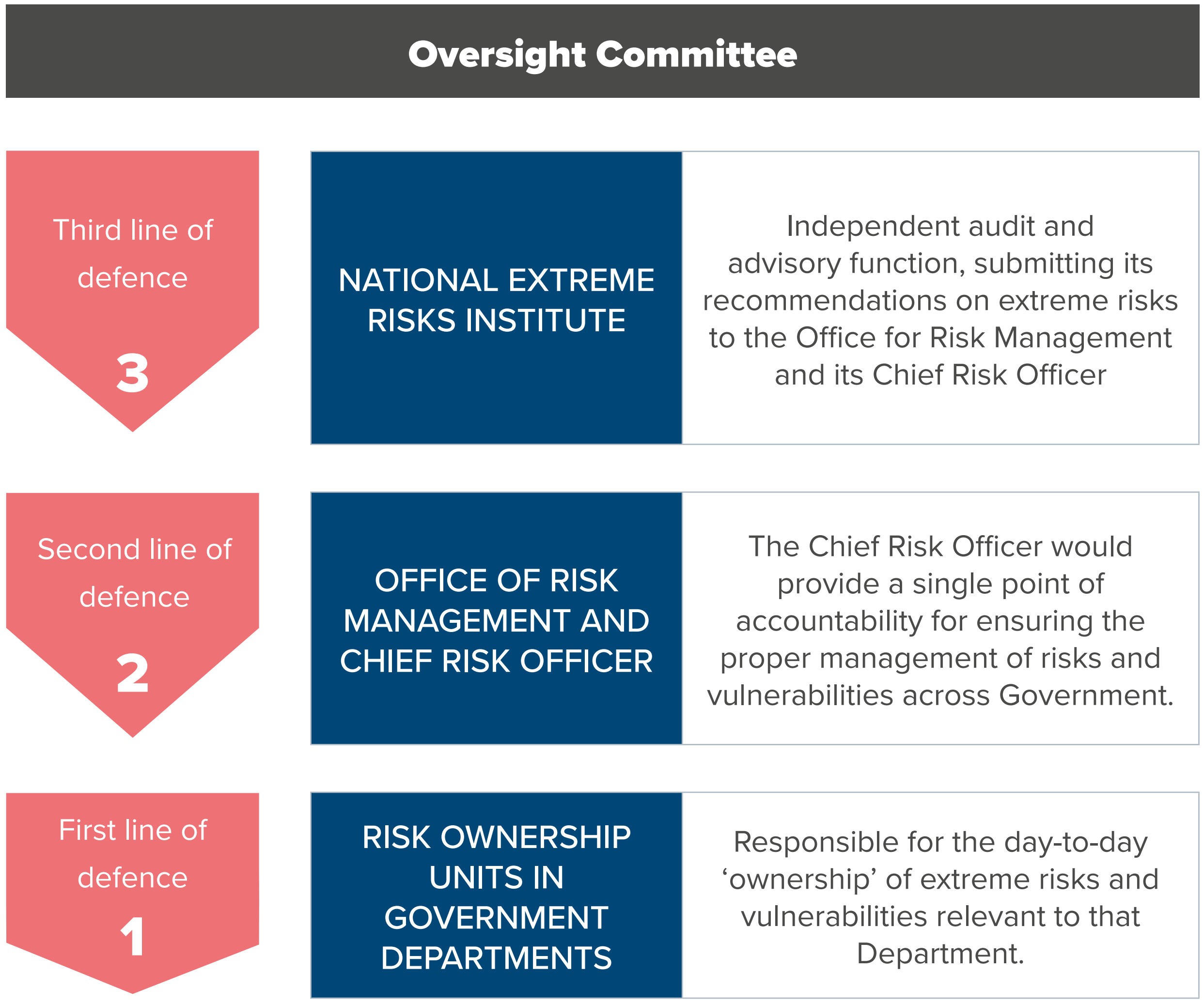
**Why this matters:** As we have seen from Covid-19, there are clear lessons to be learned about how to better assess and manage risks. A clearly defined single point of accountability in Government for risks will help transform the UK's resilience to future extreme risks.

Improving the Government's approach to risk management will need strong leadership from the centre of Government, along with iterative work between policy officials, politicians, and risk experts from a range of sectors.

The NSRA review provides scope to ensure that extreme risks are given appropriate focus within that document, and the Integrated Review sets out a promising long-term vision for national resilience. **The next step is to set out a plan for achieving this vision which has extreme risks firmly embedded within it.**

We suggest a plan below based on the 'three lines of defence' structure, which is standard practice across industry. The plan sets out the ideal implementation model, though this is flexible and alternative options are possible depending on the Government's appetite for large-scale change in this area.

**Proposed model for the Three Lines of Defence approach to risk management**



Oversight Committee

Third line of defence 3

NATIONAL EXTREME RISKS INSTITUTE

Independent audit and advisory function, submitting its recommendations on extreme risks to the Office for Risk Management and its Chief Risk Officer

Second line of defence 2

OFFICE OF RISK MANAGEMENT AND CHIEF RISK OFFICER

The Chief Risk Officer would provide a single point of accountability for ensuring the proper management of risks and vulnerabilities across Government.

First line of defence 1

RISK OWNERSHIP UNITS IN GOVERNMENT DEPARTMENTS

Responsible for the day-to-day 'ownership' of extreme risks and vulnerabilities relevant to that Department.

### a) The first line of defence: Strengthening departments' ability to manage extreme risks

**Why this matters:** Government departments should be responsible for the day-to-day management of extreme risks relevant to their department. This is currently done quite effectively for non-extreme risks, but much less so for extreme risks. Risk Ownership Units covering both extreme and non-extreme risks would ensure that a culture of risk ownership is championed across Government and help policymakers consider extreme risks when making policy decisions.

**We recommend an assessment of which Government departments are best suited to manage individual extreme risks**. Once complete, we suggest building up Risk Ownership Units of between two and six civil servants. These would be housed in relevant departments and constitute the 'first line of defence.'[26] Ministers would continue to be held accountable for the risks designated to their departments.

For example, an Electrical Grid Risk Unit might sit in BEIS. It would focus on boosting the resilience of the UK's electrical grid against extreme terrestrial and solar storms, human-made electromagnetic pulses and malicious digital intrusions.

These Units must be completely embedded in their departments and seen as part of those departments rather than extensions of the second line of defence (see immediately below). The relevant minister should also fully support the Units and understand their purpose.

**b) The second line of defence: Creating a new Government Office of Risk Management, headed by a Chief Risk Officer**

**Why this matters:** The Civil Contingencies Secretariat provides a good risk identification function, but as far as we are aware there is currently limited focus on extreme risks, and no cross-government accountability mechanism to ensure action is taken to check the quality and viability of risk planning and mitigation strategies. Without this, there is a strong chance that the UK will not be well-prepared for future extreme risks.

**A new Government Office of Risk Management, headed by a Chief Risk Officer (CRO) with specialist risk management expertise, would help bring the UK into line with current best practice from industry and elsewhere.**

This Office would ideally be an extension of the current Civil Contingencies Secretariat. However, other arrangements would also work.

Its responsibilities would include:

- **Having overall responsibility for risk management** across Government.
- **Having powers to assign responsibility for risks to ministers** and hold them to account for their risk response strategy.
- **Playing a leadership role in ensuring that risk planning, risk mitigation, and risk preparedness improves across Government.** This would include ensuring that Departmental risk plans are fit for purpose and providing a body of expertise who can support Departments with risk planning.
- **Playing a leadership role** in ensuring that risk management improves globally.
- **Running regular vulnerability assessments**. Calibration of risk severity should be

[26] Our initial assessment suggests eight new Risk Ownership Units: https://www.fhi.ox.ac.uk/wp-content/uploads/three_lines_defence.pdf.

combined with a rating of vulnerability (not just likelihood). The assessment should examine the strength of existing mitigations and crisis management capabilities, how external the threat is, and its velocity should it occur. This vulnerability assessment helps identify further mitigations required and actions to be taken by relevant risk owners.

- **Implementing the recommendations of the proposed new National Extreme Risks Institute** (see the recommendation below).

### c) The third line of defence: Establishing an independent National Extreme Risks Institute.

**Why this matters:** There is currently no government body which focuses exclusively on extreme risks. This would be the first UK public body to be exclusively incentivised to focus on public sector decision-making around dealing with extreme risks, many of which are not clearly under the management of any particular Secretary of State.

**A National Extreme Risks Institute would be tasked with providing independent advice on assessing and red-teaming the Government's approach to identifying and preparing for extreme risks, and making recommendations to the UK Government** for how it can improve its management of these risks.

This Institute would focus on identifying and supporting Government efforts to boost resilience to catastrophic events, as promised in the Integrated Review. It could be created as a What Works Center and part of the broader What Works Network.[27]
The Institute's role would include:

- **Carrying out independent, evidence-based assessments** of extreme risks. This would allow for a greater focus on risks that are low-probability but highly destructive. It would mirror Switzerland's approach[28] of an independent institute offering a separate perspective on risks to that which exists in Government, thereby reducing the chance of groupthink.
- **Carrying out issue-specific risk assessments to audit and 'red-team' Government departments** in areas of particular concern.
- **Submitting recommendations to the Government and to a new Government Office of Risk Management,** which would oversee the identification, assessment, and mitigation of risk.
- **Collating and presenting research** on extreme risks on risk management policy to decision makers.

[27] https://www.gov.uk/guidance/what-works-network

[28] https://www.oecd-ilibrary.org/governance/national-risk-assessments_9789264287532-en

- **Identifying and highlighting extreme risks** that are not clearly under the management of any particular Secretary of State.
- **Issuing a flagship report** alongside each National Security Risk Assessment and Spending Review.

The Institute should have independence from the Government and be accountable to Parliament. It should ideally be funded by way of endowment to protect it from cost-cutting exercises in future Spending Reviews. The Government should ensure that the Institute's expert staff can access relevant confidential information by providing security clearances to staff.

### Why the 'three lines of defence'?

One perceived drawback of a 'three lines of defence' structure is that it risks creating a siloed Government institution. But without a new CRO, there is no single point of responsibility for risk management. This means that it tends to be deprioritised amidst the 'tyranny of the urgent'.

And without the 'three lines' structure set out above, checks and balances are lacking, and risk owners don't get held to account to mitigate or plan for their risks. **The problem we are seeking to solve is less one of coordination, and more one of establishing clear accountability.** A lighter-touch option would be expanding the remit of the National Audit Office to include the 'third line of defence'. This would retain the audit function of the third line but lack the extreme risks expertise that an Institute would provide.

To ensure proper coordination, we would also recommend an **Oversight Committee**. The Committee would bring the three lines of risk infrastructure together, with the CRO reporting to its Chair, as well as to the appropriate departmental head (e.g. the Permanent Secretary). It could be chaired by the Institute's head or by a parliamentarian to provide independent oversight.

For more information, see the Future of Humanity Institute's proposal for a new 'three lines of defence' approach to UK risk management (2021).[29]

[29] https://www.fhi.ox.ac.uk/wp-content/uploads/three_lines_defence.pdf

## 2. Lead the way to ensure global resilience to all extreme risks — not just pandemics — post-Covid-19

**Why this matters:** The UK cannot address transnational challenges alone. But it can use its position as a diplomatic superpower to lead the journey towards global resilience to extreme risks. As countries begin to form their longer-term policy responses to Covid-19, there may never be a better moment to put extreme risks at the top of the international agenda and go beyond simply 'preparing to fight the last war' of pandemic preparedness.

During its G7 presidency and in other international fora, the UK should make the economic case to its counterparts for improving international extreme risk governance and preparedness. We set out a number of international asks below.

### a) Encouraging international long-term spending commitments on extreme risks, such as spending a target percentage of GDP on extreme risk management

This could mirror the NATO agreement to spend 2% of Gross Domestic Product on Defence, or the OECD commitment to spend 0.7% of Gross National Income on International Development. Commitments should be made now to ramp up spending on catastrophic risk prevention once the fiscal situation allows.

This would ensure that the necessary spending to provide global resilience to extreme risks is locked in, regardless of whether political and public attention fades over time.

### b) Pushing for an international agreement on pre-agreed finance and a "Crisis Lookout" function

More than 40 leaders from across the humanitarian, development and private sectors have signed a Statement of Support for the Crisis Lookout campaign. This statement calls on G7 leaders to start a new global approach for predicting crises, preparing for them, and ensuring more people are better protected.

By making pre-arranged finance the primary way to pay for crises by 2030, the G7 could ensure that funding gets to where it is needed faster and with greater impact. A global Crisis Lookout function would also better synthesise, prioritise, quantify and communicate all crisis risks and their potential costs. It would also ensure better global responses to disasters by identifying financial protection gaps at local, regional and global levels. This work would focus initially on a group of the most vulnerable countries to pilot better prediction of (and coordinated protection from) crises.

The UK Government should lead calls for its adoption this year.

### c) Creating and then leading a global extreme risks network

If the UK shows domestic leadership by introducing a Government Chief Risk Officer, it can lead the way internationally, too. Having announced its intention to create a new 'three lines of defence' risk management structure (see the recommendation immediately above this one), the UK could then call on other countries to do the same. It could encourage annual meetings between Chief Risk Officers from around the world, where countries share information and learn from each other's risk assessments. This would include exploring why some countries were better prepared for Covid-19 than others, and ensure that all national risk assessments and foresight programmes draw on global expertise. The Bank for International Settlements, which facilitates Central Bank coordination, is a useful equivalent organisation in the finance sector.

### d) Calling for a Treaty on the Risks to the Future of Humanity

Some serious risks, like climate change or nuclear weapons, are covered by at least some international law, but there is currently no regime of international law in force that is commensurate with the gravity of risks such as global pandemics, or that has the breadth needed to deal with the changing landscape of risks.

A new Treaty on Risks to the Future of Humanity has been recommended by Guglielmo Verdirame QC.[30] He argues that such a Treaty would provide a framework for identifying and addressing such risks, and that international diplomacy and domestic politics must be engaged at the highest level to achieve it. A new Treaty should have a series of UN Security Council resolutions to place this new framework on the strongest legal footing.

The UK could take a global leadership position on this issue by starting to build an alliance towards a treaty with like-minded countries, such as Australia, Japan and New Zealand.

[30] https://unherd.com/2020/04/for-china-a-legal-reckoning-is-coming/

## 3. Normalise red-teaming in Government, including creating a dedicated red team to conduct frequent scenario exercises (estimated cost: £800k annually)

**Why this matters:** It is vital to scan for and discover vulnerabilities to UK infrastructure, thus avoiding and anticipating as many disasters as possible. A team recruited with the skills to do this well can reduce groupthink and question key assumptions. The Government's Integrated Review sets red-teaming as a "reform priority", highlighting the need to "foster a culture that encourages more and different kinds of challenge, further developing capabilities such as red-teaming to mitigate the cognitive biases that affect decision-making".[31]

Not only does the Integrated Review include red-teaming and challenge as a reform priority; it ran its own red-teaming exercise to "challenge and test emerging thinking from the perspective of third parties", which is to be commended.[32]

We welcome red-teaming being used when developing policy, particularly around civil and defence risks. **We recommend maintaining a red team of seasoned experts with the relevant background checks and security clearances, tasked with running scenario exercises and then implementing the recommendations from their findings.**

The red team would focus on scenarios such as:

- A major cyber attack on UK infrastructure.
- The non-availability of one or more major cloud providers in the UK for an extended period of time.
- An accidental or deliberate release of a virus.[33]
- Cut-off from the internet for an extended period of time.

This would help ensure that the most important scenario exercises are conducted frequently and that clear lessons learned are 'owned' by senior policy makers. The results could be reported to a designated body tasked with ensuring that the findings are used, such as to Parliament, a new Government Office of Risk Management or National Extreme Risks Institute.

We would also sound a note of caution on red teaming:

- For certain issues (such as biosecurity), red teaming can result in the identification

[31] Integrated Review: p98.

[32] Integrated Review: p108.

[33] https://www.gov.uk/government/publications/foot-and-mouth-disease-2007-a-review-and-lessons-learned

of new vulnerabilities, the publication of which should be carefully thought through.

- It may also fail to meet its objectives whilst providing a false sense of security around a particular project, especially if red teaming is not undertaken by people with knowledge of the institutions or issues under question.

More generally, the Government should broadly ensure that civil servants are skilled in the art of decision-making in situations of high uncertainty, and can apply a range of appropriate tools beyond red-teaming, including robust decision-making, cluster thinking, exploratory thinking, scenario planning, and adaptive planning. When key decisions are made in Government — such as decisions relating to management of extreme risks — it is essential that a variety of different approaches are used to evaluate those decisions.

## 4. Revise the Green Book's discount rate and ensure the Treasury adopts key recommendations on intergenerational fairness

**Why this matters:** The policy decisions we make today have huge implications for future generations, yet are often overly influenced by short-term pressures. Our children and grandchildren, as well as current generations, deserve to be treated equitably, and we need to consider the long-term consequences of today's policy decisions.

Certain technical changes to Treasury processes would significantly improve incentives for decision makers to act in the interests of the long term, which would in turn improve management of extreme risks. The Institute for Government has noted that within the Treasury "there is too little focus on the long term and on the trends — and foreseeable problems — which may affect these plans."

We therefore recommend that the Treasury revise the Green Book (HM Treasury's guidance on how to appraise policies, programmes and projects) and the discount rate.

Discounting is a way of comparing costs and benefits with different time spans. The Treasury uses a discount rate of 3.5% for future costs and benefits, in part to adjust for 'social time preference' (i.e. the value society attaches to present, as opposed to future, consumption).[34]

**We recommend lowering the Green Book's discount rate to ensure that today's policies do not make future generations disproportionately worse off**. The discount

[34] https://www.gov.uk/government/collections/the-green-book-and-accompanying-guidance-and-documents

rate should decline more quickly in the long run, the ‘pure time preference’ part of the discount rate should be set at 0%, and the Green Book should acknowledge that the current discount rate formula does not work for estimating the costs of significant disasters (for instance, because they could lead to significant economic decline).

The Green Book should also have more detail on how to account for second-order effects (the further consequences of an action, beyond the desired initial consequence).

## 5. Establish a new Defence Software Safety Authority as a sub-agency of the Defence Safety Authority, to protect UK defence systems from emerging threats (estimated cost: £5 million annually)

**Why this matters:** The procurement and development of defence systems that integrate increasingly capable AI, machine learning and autonomy is vital to national security. But as this area grows in importance and complexity, ensuring the good governance, safety, and security of these systems becomes ever more important to avoid vulnerabilities or accidents that could harm service people or citizens or lead to inadvertent escalation.

The Defence Safety Authority has a number of sub-agencies that ensure the safety and good governance of risks such as Land (DLSR), Ordnance and Explosives (DOSR), Medical Services (DMSR), and Nuclear (DNSR).

A new Defence Software Safety Authority would be tasked with regulating the safety of defence systems that integrate increasingly capable AI, machine learning and autonomy. This should involve adopting a new regulation in the form of a Joint Service Publication. This would require a targeted increase in funding for additional hiring, and training to judge the limitations, risks, and overall safety and security of new defence systems.

Key priorities when procuring these systems include improving systemic risk assessment in defence procurement and ensuring clear lines of responsibility so that senior officials are held responsible for errors caused in the defence procurement chain.[35]

[35] https://www.cser.ac.uk/resources/written-evidence-defence-industrial-policy-procurement-and-prosperity/

## 6. Fund a comprehensive evaluation of the actions required to increase the resilience of the electrical grid

**Why this matters:** The electrical grid is at risk from an array of both natural and human-made threats, any one of which could cause widespread disruption and thousands of avoidable deaths. If the grid is damaged or disabled, perishables such as food and medicine will expire, communication networks will collapse, oil and gas distribution will halt, water purification and distribution will cease functioning, and effective governance will likely disappear. In the worst-case scenario, nuclear reactors will also melt down. The grid's ability to withstand the impact of these threats is a major concern for national security and the ability to maintain basic services for the larger population.

There is currently a window of opportunity to make Britain a global leader in electrical grid resilience and preparedness for the next century of risks. The grid is the country's largest machine, and it's also 100 years old. It is currently undergoing the process of being completely changed by renewables. As this transition occurs, **we have the opportunity to upgrade the physical and digital integrity and resilience of the grid, and we can do it in a way that also helps promote clean energy generation and a Green recovery.**

This effort should produce specific policies, procedures and technological solutions, together with implementation timelines and an estimate of required resources.

It should include action plans in the following areas:

- **Improving the UK's ability to identify threats and vulnerabilities**: Produce standards and guidelines for threat identification and emergency response planning and preparation which are accepted and implemented by the energy sector.
- **Increasing the ability to protect against threats and vulnerabilities**: Establish a nationwide network of resiliency test platforms that are long-duration, blackout-survivable microgrids. These should be located in facilities controlled by the Government, in stable areas that are free from flooding, severe weather and other high-impact disasters.
- **Improving recovery capacity and time:** Design ultra-secure, low-power, self-healing wireless networks capable of bypassing compromised network components while maintaining essential connectivity to critical grid assets. This should be designed to preserve fail-safe operations that engage within minutes of a cyber attack.

# INCREASING FUNDING FOR RESEARCH INTO EXTREME RISKS

| CURRENT GOVERNMENT FOCUS | GREEN/ AMBER |
|---|---|

## Section Contributor

Haydn Belfield
Research Associate & Academic Project Manager, Centre for the Study of Existential Risk, University of Cambridge

## Current level of Government focus

The UK has world-leading strengths in academic research in areas relevant to long-term and extreme risk, such as AI and biotechnology.[36] The announcement of the £800 million Advanced Research and Invention Agency (ARIA) and the commitment to increase UK investment in R&D to 2.4% of GDP by 2027 are both highly welcome.

The new Integrated Review sets an encouraging tone, too, with the UK aiming to be a "science and tech superpower by 2030" through our comparative advantage in areas like AI, and announcing a new Office for Talent.[37]

[36] https://www.scimagojr.com/countryrank.php

[37] Integrated Review: p39.

As the Integrated Review recognises, science and technology plays a crucial role in our national resilience and in our ability to address transnational challenges. Unfortunately, **publicly funded research on extreme risks remains significantly underfunded compared to its importance**. Total funding on AI safety, for example, is significantly less than the total funding going into private investment to accelerate its capabilities.[38] This could significantly impair our ability to respond to future crises.

## Key extreme risks research policy recommendations

### 1. Create a pool of machine-learning-relevant computational resources to provide free of charge for socially beneficial applications and AI safety, security, and alignment research (estimated cost: approx. £35 million per year)

**Why this matters:** Access to large amounts of AI computational resources (‘compute’) — for instance, computing clusters of machine learning-optimised computer chips — is critically important for AI safety and to maintain UK scientific and economic leadership.

Many recent machine-learning breakthroughs and expected advances in this area are reliant on large compute budgets that are currently beyond the reach of academia and civil society. This has led to research being skewed towards short-term aims rather than developing socially beneficial applications or AI safety, security and alignment.

**We recommend creating a ‘compute fund’ to provide free or subsidised computation resources to researchers working on socially beneficial AI applications or AI safety, security and alignment**. This idea is currently being investigated in the United States.[39] [40]

This would benefit critical research in the following areas:

[38] https://aiimpacts.org/changes-in-funding-in-the-ai-safety-field/

[39] The National Defence Authorization Act 2021 includes a provision that a National AI Research Resource Task Force should investigate setting up a compute fund or cluster. A summary of the relevant provisions is available here.

[40] The US National Security Commission on Artificial Intelligence made a similar recommendation in its recent report: “To bridge the “compute divide,” the National AI Research Resource would provide verified researchers and students subsidized access to scalable compute resources, co-located with AI-ready government and non-government data sets, educational tools, and user support. It should be created as a public-private partnership, leveraging a federation of cloud platforms.” See p191: https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf

- **Human-centric and beneficial AI applications**: e.g. medical research and diagnostics, energy optimisation and decarbonisation, and AI for the Sustainable Development Goals.
- **AI safety research**: e.g. into areas such as transparency, privacy, accountability, technical robustness, and fairness.
- **Providing open-source alternatives to commercial AI models** (i.e. non-proprietary and freely available).
- **Increasing scrutiny**: enabling the scrutiny, auditing and certification of commercial and government AI systems.
- **Using AI to test AI**: using AI to rigorously identify, test and eliminate potential bugs, hazards, and failures.

## 2. Invest further in AI safety R&D

**Why this matters:** Promoting technical AI safety research is critically important — not only due to the dangers of unsafe systems, but because it will bolster the UK's competitiveness as states and companies seek to acquire safe and beneficial AI systems.

**We strongly recommend the UK increase its funding for technical AI safety research.** This could be done via the Alan Turing Institute, through ARIA, or through the new autonomous systems research hub at Southampton University.

Funding should be made available for four broad areas of research:

- **Alignment**: For very capable AI systems, pursuit of an incorrectly specified goal would not only lead an AI system to do something other than what we intended, but could lead the system to take harmful actions. Can we design training procedures and objectives that will cause AI systems to learn what we want them to do?
- **Robustness**: As AI systems become more influential, reliability failures could be very harmful, especially if failures result in an AI system learning an objective incorrectly. Can we design training procedures and objectives that will cause AI systems to perform as desired on inputs that are outside their training distributions or that are generated adversarially?
- **Interpretability**: If an AI system's internal workings could be inspected and interpreted, we might be able to better understand how its models work and why we should or should not trust the model to perform well. Interpretability could help us to understand how AI systems work and how they may fail, misbehave, or otherwise not meet our expectations.

- **Assurance of deep learning systems**: It is not enough for models to be robust. In particularly high-stakes situations, we also need assurance that this is the case. However, traditional methods for gaining such high assurance typically cannot be applied to deep-learning AI systems. New methods need to be developed.

## 3. Invest further in applied biosecurity R&D

**Why this matters:** With investment in highly promising horizon technologies, the UK's research community could dramatically improve how well and how quickly the UK can respond to new biological threats.

**Currently, interventions to tackle a novel biological threat can either be rapidly deployed (e.g. non-pharmaceutical interventions) or highly effective (e.g. vaccines), but not both.** Innovative technologies can help close this gap, and should be urgently prioritised for development.

As mentioned in the Biosecurity section of this report, one promising area for development is in the field of metagenomic sequencing — a potentially game-changing process which can identify new, unexpected pathogens in the first few patients who become infected. This would allow for a much earlier targeted response to an outbreak of an unknown virus.

This significant investment in biosecurity R&D would shore up Britain's status as a biosecurity world leader, as showcased by the Oxford / AstraZeneca vaccine, and make future breakthroughs more likely. It would also help keep pace with developments in the United States, where a Bipartisan Committee recently proposed an Apollo Program for Biodefense, which includes recommendations for long-term multi-year funding in this area.[41]

**We would also recommend increased caution before undertaking 'gain of function' research.** In the context of biosecurity, this is a type of research that aims to increase the virulence and lethality of pathogens and viruses.

Such research can improve scientific understanding in an important area, as it helps better understand how natural pandemics could evolve to become worse. However there are important risks to factor into any cost-benefit analysis of gain of function research. For instance, the newly acquired information generated by the research could be obtained and misused by hostile actors. Alternatively, newly created pathogens and

[41] https://biodefensecommission.org/reports/the-apollo-program-for-biodefense-winning-the-race-against-biological-threats/

viruses could accidentally escape from the laboratories in which they were created. Dual-use research of concern in other high-risk domains — such as the field of artificial intelligence — should be treated with similar caution.

## 4. Invest further in improving long-term forecasting and planning

**Why this matters:** The relatively new field of forecasting has the potential to substantially improve how well the UK predicts the probabilities of future disasters, thus helping allocate resources towards risks that are the most serious and the most likely. Thus far, forecasting has mainly focused on generating near-term predictions, but there is ample scope for this to change.

**We recommend extensive research into improving forecasting techniques**, for example, through the use of quantified falsifiable predictions and full inference cycle tournaments, as proposed by Professor Philip Tetlock.[42] Cosmic Bazaar, the internal Civil Service forecasting platform, may be able to lead this work.

In terms of current good practice, the Office of Budget Responsibility produces publicly available fiscal and economic forecasts, and reviews annually (in a report to Parliament) how well their forecasts match reality and how they can improve.

The Government should build on work done in the US by the Intelligence Advanced Research Projects Activity's intelligence community prediction market, and from the Center for Security and Emerging Technologies' new policy-forecasting project, Foretell.

The Government should also research long-term planning and risk planning. RAND has conducted significant work on improving the robustness of decision-making and improving adaptive planning for high-uncertainty events (including work with governments on risk preparedness). The UK has adopted these ideas in some domains (for instance, the Thames Estuary TE2100 Plan), but there is scope to build up civil service expertise in other areas.

[42] https://www.sas.upenn.edu/tetlock/

# FURTHER INFORMATION AND ACKNOWLEDGEMENTS

To learn more about the field of extreme risks, see the Centre for the Study of Existential Risk's research work,[43] the Future of Humanity's Institute's research areas,[44] and the Global Challenges Foundation's annual Catastrophic Risk Reports.[45]

For further details of these recommendations and the evidence base which sits behind them, or to arrange meetings with the authors and contributors to discuss these ideas, please email info@longtermresilience.com.

We would like to thank all our section contributors, along with Tildy Stokes, Niel Bowerman, Seb Krier, Jack Clark, James Ginns, Simon Beard, Shahar Avin, Henry Elkus, Sean Roche, Sam Feinburg, Liam Glass, Jake Swett, Caroline Jeanmaire, Laura Pomarius, Bella Soares, Natalie Martin, Alex Hill, Seán Ó hÉigeartaigh, Nick Bostrom, Allan Dafoe, Ula Zarosa, Richard Pyle, John Fogle and Denise Ferreras for their help developing aspects of this report and associated materials.

## Recommended Reading

The Centre for the Study of Existential Risk's Risk management in the UK: What can we learn from COVID-19 and are we prepared for the next disaster? (2020), University of Cambridge. By Sam Hilton and Caroline Baylon.

The Future of Humanity Institute's Proposal for a new 'three lines of defence' approach to UK risk management (2021), University of Oxford. By Toby Ord.

The Future of Humanity Institute's Proposal for a National Institute for Biological Security (2020), University of Oxford. By Cassidy Nelson and Gregory Lewis.

The Centre for the Study of Existential Risk's Submission of Evidence to The House of Lords Select Committee on Risk Assessment and Risk Planning (2021), The University of Cambridge. By Shahar Avin, Lalitha Sundaram, Jess Whittlestone, Matthijs Maas and Tom Hobson.

The Centre for the Study of Existential Risk's Foresight for unknown, long-term and emerging risks, Approaches and Recommendations (2021), University of Cambridge. By Clarissa Rios Rojas, Catherine Rhodes, Shahar Avin, Luke Kemp, Simon Beard.

The US National Security Commission on Artificial Intelligence's Final Report (2021).

[43] https://www.cser.ac.uk/research/

[44] https://www.fhi.ox.ac.uk/research/research-areas/

[45] https://globalchallenges.org/initiatives/analysis-research/reports/

THE CENTRE FOR
LONG-TERM RESILIENCE