EXTREME RISKS WORKING PAPER 2021-1

Proposal for a New 'Three Lines of Defence' Approach to UK Risk Management

Toby Ord
Macrostrategy Research Group
Future of Humanity Institute, University of Oxford

2021

Cite as: Ord, T. (2021): "Proposal for a New 'Three Lines of Defence' Approach to UK Risk Management", Extreme Risks Working Paper 2021-1, Future of Humanity Institute, University of Oxford

Executive Summary

The coronavirus pandemic highlights the devastating impact that extreme risks can have on our health and economy. Extreme risks—high-impact threats that have a potentially global reach—need our urgent attention.

We do not know which extreme risk will come next, but we do know what many of the most extreme risks are, and that our preparation needs to be much better. This paper proposes a new Three Lines of Defence system to ensure that extreme risks are sufficiently captured in UK risk management. It suggests going beyond simply 'fighting the last war' and focusing solely on better pandemic preparedness, instead of transforming the UK's resilience to extreme risks across the board.

There is no better time for the UK to give extreme risks the attention they deserve, but rarely receive. Oxford's *Future of Humanity Institute* hopes that this time-critical opportunity to put in place lasting protections for UK citizens will be taken. We welcome discussion and feedback about this paper.

Three Lines of Defence Structure[1]:

- **Line 1: Eight new Government Risk Ownership Units**. These eight units would be responsible for day-to-day risk management within departments, and embedding the right risk culture, with a particular focus on extreme risks in areas including AI safety, biological security and electrical grid safety.

- **Line 2: A Chief Risk Officer and Office for Risk Management.** The Chief Risk Officer (CRO) would be the single point of accountability for ensuring effective management of extreme risks across Government.

- **Line 3: An independent National Extreme Risk Institute.** This would provide an audit and advisory function to the CRO.

Funding: An MVP of this model would cost £8.26 million. This cost would include (i) funding for the three lines of defence (£4.12 million annually), and (ii) a £4.14 million dedicated pot of funding for the CRO to implement the initial changes required.

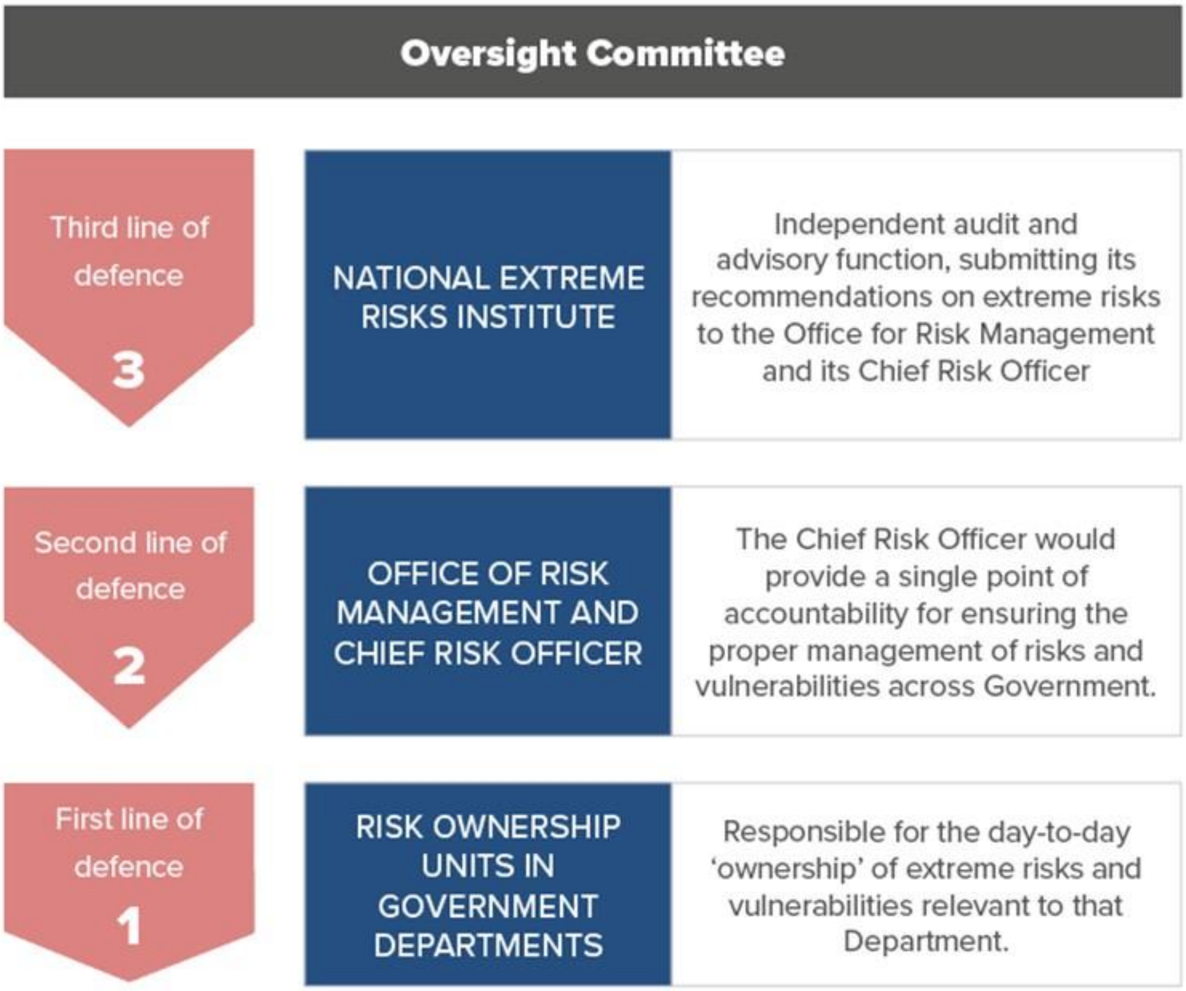
Personnel decision: to appoint a new Chief Risk Officer. The CRO must have substantial clout and independence to succeed. This proposed structure should also take into account the wider, longer-term review of crisis response structures across CCS and NSS.

Why: There is roughly a *one in six chance* of existential catastrophe in the next 100 years from extreme risks such as pandemics, extreme climate change scenarios, nuclear conflicts, and the creation of an unaligned general artificial intelligence.

[1] This 'three lines of defence' structure is a familiar method used in the private sector, including in finance.

As COVID-19 demonstrated, the UK's approach to planning for extreme risks is currently insufficient in a range of key areas. Better identification and management of extreme risks has a high expected value of preventing future disasters and consequential economic damage.

The new Integrated Review highlights the need for "low-probability, catastrophic-impact threats" 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." The next step is to set out exactly how the Government can embed extreme risks into its resilience planning, its upcoming AI strategy and biosecurity review.



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.

Table of Contents

List of Tables

Annex A - Evidence Base: Overview of the Challenges and Recommendations

There is a substantial chance of an existential catastrophe in the next 100 years.
Existential catastrophes would destroy the UK's present and the future, affecting both present citizens and potential future ones. They therefore have uniquely high stakes because the UK would by definition be unable to recover from one single such disaster.[2]

Extreme risks which cause these catastrophes—including pandemics, extreme climate change scenarios, nuclear conflicts, and the creation of unaligned general artificial intelligence—are not low in their probability. In my book, *The Precipice,* I calculate the probability of such an event in the next century at one in six. See **Annex B** for a breakdown of my probabilities across different extreme risks.

As COVID-19 demonstrated, the UK's approach to planning for extreme risks is insufficient in key areas.
Extreme risks are currently being given insufficient attention, and may even be going unidentified, due to current shortcomings in the UK's approach to risk planning (see **Annex C** for details).

The UK does not simply need to improve its foresight tools. It also needs to ensure (i) better incentives for decision makers to act on these tools, (ii) greater skills and expertise in areas of emerging risk like AI and biosecurity, (iii) more robust electrical grid safety infrastructure, (iv) more research into extreme risks, and (v) strong international leadership in all of these areas.[3] [4]

To analyse the changes required on an ongoing basis, and to drive them through the system, we recommend a new 'three lines of defence' risk management structure, which is best practice in industry.

1. The first line of defence would be Government departments themselves, bolstered by eight newly created Risk Ownership Units sitting within them.
The Units would contain between two and six civil servants, and would be responsible for the day-to-day management of extreme risks relevant to that Department.[5]

[2] https://theprecipice.com/faq#existential-risk

[3] Arguably the most serious shortcoming is that the National Security Risk Assessment does not sufficiently explore high uncertainty risks or emerging risks and focuses too heavily on recent events. This led the UK to be well prepared for an influenza pandemic, but not for a coronavirus (as evidenced by the fact that we did not plan for a lockdown, which is now a key pillar of our strategy to deal with COVID-19). See Annex C for further details of this and other shortcomings in the UK's existing extreme-risk management process.

[4] COVID-19 is not the first time we have failed to identify an extreme risk, despite significant evidence of its existence. The risk from volcanic ash was only added to the UK's National Risk Assessment in 2012 after the 2010 and 2011 Icelandic eruptions, despite the availability of significant historical evidence from the 1700s to suggest that such an event of this kind was highly probable.

[5] The Units will also ensure that a culture of risk ownership is championed throughout their whole departments, and that departments in particular become much more attuned to the need to take extreme risk seriously and mitigate it.

The eight proposed Risk Ownership Units

1) **Artificial Intelligence Risk Unit (in BEIS)** - focusing on areas such as AI safety research, increasing AI foresight and progress tracking, and bringing more AI technical expertise into Government.
2) **Biological Security Risk Unit (in DHSC, or as an extension of the new National Institute for Health Protection)** - focusing on ensuring the biological security of the UK, going beyond naturally occurring pandemics and into areas such as countering the threat of biological weapons and developing effective defences to biological threats.
3) **Extreme Climate Change Risk Unit (in BEIS or DEFRA)** - focusing on ensuring that UK climate policy focuses sufficiently on mitigating the 'tail' scenario (approx. 5% probability) of more than a six-degree rise in global temperatures.
4) **Defence and Cybersecurity Risk Unit (in MoD)** - focusing on areas such as reviewing the UK's definition of Lethal Autonomous Weapons to that used by most other nations, ensuring AI systems are not incorporated into NC3 (nuclear command, control, communications), and running frequent scenario exercises relating to extreme risk events.
5) **Electrical Grid Risk Unit (in BEIS, or Cabinet Office Civil Contingencies Secretariat)** - focusing on boosting the resilience of the UK's electrical grid against extreme terrestrial and solar storms, man-made electromagnetic pulses and malicious digital intrusions.
6) **Extreme Risks Research Unit (in UKRI)** - focusing on producing and commissioning research in critical areas relating to AI and biosecurity risks and improved forecasting techniques.
7) **Extreme Risks Management Unit (in Cabinet Office Civil Contingencies Secretariat or Treasury)** - ensuring that the UK's risk management processes take proper account of extreme risks, and ensuring that officials are incentivised to prioritise the long term in their spending decisions through (for instance) amendments to the Green Book.
8) **International Extreme Risk Management Unit (in FCDO)** - focusing on the UK playing a global leadership role in the management of extreme risks, for example by increasing the capacity of the International Atomic Energy Agency to verify that nations are complying with safeguarding agreements, and leading calls for the creation of a new Treaty on the Risks to the Future of Humanity.

These Units must be completely embedded in their departments and be seen as part of those departments, rather than extensions of the second line of defence (see immediately below).

2. The second line of defence would be a Chief Risk Officer, supported by an Office for Risk Management.

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 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.
- Having a training function to ensure that best practice for risk management is transferred across Government.

Clout: the CRO needs to be very senior, and enjoy strong and vocal support both politically and from the Cabinet Secretary. They must be motivated primarily by the challenge of countering extreme risks, rather than short-term emergencies or career progression. The CRO must be strong enough to drive the changes required throughout the system, and to hold departments to account for managing their risks. The remit of the role must be the full breadth of the risk portfolio across Government.

Independence: The CRO needs a significant degree of independence from politics, and their team would ideally be established as a non-departmental public body.[6]

3. The third line of defence would be a National Extreme Risks Institute, sitting outside Government as a legally independent entity capable of raising its own funds.

The Institute would have an *independent audit and advisory function* and would submit its recommendations to the Office for Risk Management and its CRO.

It would be staffed by a team of 13, comprising (i) academic and technical experts in the field of extreme risks, (ii) former civil servants with experience inside Government to ensure that the recommendations it produces for the CRO are sufficiently concrete and actionable, and (iii) operational and support roles.

Better identification and management of extreme risks has an expected value of saving millions of lives and £billions, and we now have an ideal moment to make these changes.

[6] Chief Risk Officers in the private sector derive much of their authority from having a dual report to both the CEO and to the (usually independent) chair of a board-level risk committee. We therefore recommend putting in place an Oversight Committee chaired by the Head of the Institute, so that the CRO has an independent reporting line to this chair as well as to the Cabinet Secretary.

The combination of the COVID-19 response, the upcoming publication of the Integrated Review and the ongoing civil service reforms means that the Government has an excellent opportunity to make these changes now and become a world leader in this field.

The UK is already an academic world leader in the field of extreme risks. By implementing the changes we recommend, the UK Government would become a world leader too. It would have put in place the single most robust system for extreme risk management in the world.

The total cost of this proposal is £8.26 million annually.

This cost would include funding for the three lines of defence (£4.12 million annually), in addition to an annual dedicated pot of funding for the CRO to implement the initial changes required, which would total £4.14 million annually.

The estimate is based on costed recommendations that have been prepared by extreme risks experts at Oxford University's Future of Humanity Institute and Cambridge University's Centre for the Study of Existential Risk. This can only be an estimate, since the CRO will ultimately determine the extreme risk recommendations to be implemented, based on the recommendations provided by the Institute.

Annex B - Breakdown of the Chance of Existential Risk in the Next 100 Years

(Source: *The Precipice*, Toby Ord, Bloomsbury 2020)

Existential catastrophe via	Chance within next 100 years
Natural Risk	
Asteroid or comet impact	~ 1 in 1,000,000
Supervolcanic eruption	~ 1 in 10,000
Stellar explosion	~ 1 in 1,000,000,000
Total natural risk	~ 1 in 10,000
Anthropogenic risk	
Nuclear war	~ 1 in 1,000
Climate change	~ 1 in 1,000
Other environmental damage	~ 1 in 1,000
'Naturally' arising pandemics	~ 1 in 10,000
Engineered pandemics	~ 1 in 30
Unaligned artificial intelligence	~ 1 in 10
Unforeseen anthropogenic risks	~ 1 in 30
Other anthropogenic risks	~ 1 in 50
Total anthropogenic risk	~ 1 in 6
Total existential risk	~ 1 in 6

Table 1: Rough estimates for the existential risk from different threats.

Annex C - Current shortcomings in the UK's Approach to Risk Planning

(Excerpts from a research paper by Sam Hilton and Caroline Baylon, Research Affiliates at the University of Cambridge's Centre for the Study of Existential Risk.)

Summary of key areas for improvement

There are areas for improvement with the National Security Risk Assessment (NSRA):

- The NSRA does not sufficiently explore high-uncertainty risks (risks where estimating the likelihood is difficult). This is due to the exclusion of low-probability risks and emerging risks, and too great a focus on recent events.
- The NSRA categorises and compares risks in a potentially misleading manner, with descriptions of risks being based on what is considered reasonable to plan for.
- The NSRA process could benefit from greater use of external expertise.
- In the light of COVID-19, it is notable that the NSRA focused too much on influenza rather than other diseases. For example, the most recent National Risk Register claimed that "emerging infectious diseases" (which would include COVID-19) could lead to "up to 100 fatalities".

There is also scope for improving the UK's risk planning:

- There is no set process, body of expertise or oversight mechanism in place to ensure that departmental risk plans are adequate.
- In the light of COVID-19, 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 response strategies to COVID-19.

The UK has good risk management processes by international standards, yet the issues with the NSRA are sufficiently serious that *major risks to the UK may be going unidentified.* We hope the government will recognise the importance and urgency of addressing this.

General risk management challenges that confront any government

1. In our modern, interconnected world, many of the risks we face are global, such as the 2007-08 financial crisis or COVID-19. Global risks need global management. For example, improving biosecurity in other countries reduces the pandemic risks to the UK. International cooperation is therefore key.

2. Risk preparation increases after disasters occur, but can abate over time. For example, financial regulations are often brought in after a financial crisis but then reduced prior to the next financial crisis[7]. Protecting budgets, creating oversight mechanisms or making long-term commitments would help address this.

[7] IMF (2018). Regulatory Cycles: Revisiting the Political Economy of Financial Crises, WP/18/8, January 2018

3. There is a tendency to "prepare to fight the last war". Planners tend to assume that the future will have many of the same features as the past, yet future risks often differ significantly from past risks. This is a known issue in defence and risk management and was raised by civil servants we interviewed. Managing this requires being able to prepare for and handle situations of high uncertainty.

This tendency to prepare to fight the last war affected how well-prepared states were for the COVID-19 pandemic. An influenza pandemic has topped lists of UK concerns since swine flu in 2009, and the UK prepared for influenza rather than a coronavirus (or for a pandemic more broadly)[8], as we discuss below. Meanwhile, countries that had experienced outbreaks of SARS (a coronavirus) in the early 2000s had better plans to handle COVID-19 [9] [10] [11].

If the UK government's response to COVID-19 is just to better prepare for pandemics, or even just to better prepare for zoonotic pandemics or coronavirus pandemics, then the UK would be making this same mistake again. The next catastrophe could well be something else: a global food shortage, a solar storm, a nuclear incident, an attack on critical infrastructure, or an unexpected societal consequence of an emerging technology.

The international risk management landscape

Internationally, government risk management is poor. COVID-19 has highlighted a fact that was already known: that governments do not sufficiently prepare for disasters. For example, the 2019 Global Health Security Index[12] found that the UK was one of the most well-prepared countries for a pandemic, but that every country had significant weaknesses.

[8] Professor Van-Tam (2020) DQ1008 Oral evidence: UK Science, Research and Technology Capability and Influence in Global Disease Outbreaks

[9] The Guardian (2020). Experience of Sars a key factor in countries' response to coronavirus

[10] Axios (2020). SARS made Hong Kong and Singapore ready for coronavirus

[11] Fortune (2020). SARS taught Taiwan how to contain the coronavirus outbreak

[12] ghsindex.org (2019). 2019 Global Health Security Index. The Global Health Security Index was an assessment of global health security capabilities produced by Johns Hopkins, the Nuclear Threat Initiative and the Economist Intelligence Unit.

Annex D - Costings for the Three Lines of Defence

Summary of costs

Total Costs: £4.12 million annually:

- First line: Risks Ownership Units x 8: £2.08 million annually
- Second line: CRO and Office for Risk Management: £999k annually
- Third line: National Extreme Risks Institute: £1.02 million annually

First Line Costs

Total: £2.08 million annually

1. Artificial Intelligence Risk Unit (in BEIS)

Potential projects (see Annex E for details):

- Creating an AI Observatory to improve foresight and progress tracking in AI research
- Bringing more technical AI expertise into government through scheme equivalent to TechCongress, and creating new AI roles in key departments
- Creating a pool of machine learning-relevant compute to provide free for socially beneficial applications and AI safety, security and alignment research (NB would need to be funded separately)

Employee	Annual Salary	Staff Costs (NICS and pension)	Office Costs	Total Cost for Employee
Grade 6 - Head of Unit	£58,800	£8,167	£26,794	£93,753
Grade 7 - coordination lead with Office for Risk Management	£52,075	£7,083	£23,663	£82,821
Grade 7 - coordination lead with wider Department	£52,075	£7,083	£23,663	£82,821

Table 2: Artificial Intelligence Risk Unit costs.

Total annual operating cost for Unit: = £260k.

2. Biological Security Unit (in DHSC, or as an extension of the new National Institute for Health Protection)

Potential projects (see Annex E for details of the first two):

- Establishing a Biosecurity Liaison Officer to improve coordination between the biosciences and security communities
- Pushing for the screening of DNA synthesis for dangerous pathogens, and regulation of DNA synthesis machine screening
- Developing effective defences to biological threats, helping bring horizon technologies (especially pathogen-blind diagnostics) to technical readiness
- Promoting responsible biotechnology development across the world
- Developing talent and collaboration across the UK biosecurity community

Employee	Annual Salary	Staff Costs (NICS and pension)	Office Costs	Total Cost for Employee
Grade 6 - Head of Unit	£62,404	£8,747	£28,460	£99,611
Grade 7 - coordination lead with Office for Risk Management	£62,404	£8,747	£28,460	£99,611
Grade 7 - coordination lead with wider Department	£49,529	£6,673	£22,480	£78,682
Grade 7 - coordination with wider bioscience stakeholders	£49,529	£6,673	£22,480	£78,682
Grade 7 - key project (e.g. DNA screening)	£49,529	£6,673	£22,480	£78,682
Grade 7 - key project (e.g. horizon technologies)	£49,529	£6,673	£22,480	£78,682

Table 3: Biological Security Unit costs

Total annual operating cost for Unit: £514k

3. Extreme Climate Change Unit (in BEIS or DEFRA)

Potential project: Ensuring that UK climate policy focuses sufficiently on mitigating the 'tail' scenario (approx. 5% probability) of a more than six-degree rise in global temperatures.

Employee	Annual Salary	Staff Costs (NICS and pension)	Office Costs	Total Cost for Employee
Grade 6 - Head of Unit	£58,800	£8,167	£26,794	£93,753
Grade 7 - coordination lead with Office for Risk Management	£52,075	£7,083	£23,663	£82,821
Grade 7 - coordination lead with wider Department	£52,075	£7,083	£23,663	£82,821

Table 4: Extreme Climate Change Unit costs.

Total annual operating cost for Unit: £259k.

4. Defence and Cybersecurity (in MoD)

Potential projects (see Annex E for details):

- Ensuring that the UK Government does not incorporate AI systems into NC3 (nuclear command, control, communications), and leads on establishing this norm internationally
- Establishing a new Defence Software Safety Authority as a sub-agency of the Defence Safety Authority
- Creating an independent red team to conduct frequent scenario exercises
- Setting up throughout-lifetime stress-testing of computer and AI system safety and security
- Running more AI cyber security guidance and training
- Updating the UK's definition of Lethal Autonomous Weapons to that used by most other nations.

Employee	Annual Salary	Staff Costs (NICS and pension)	Office Costs	Total Cost for Employee
Grade 6 - Head of Unit	£63,500	£8,925	£28,970	£101,395
Grade 7 - coordination lead with Office for Risk Management	£53,500	£7,313	£24,325	£85,138
Grade 7 - coordination lead with wider Department	£53,500	£7,313	£24,325	£85,138

Table 5: Defence and Cybersecurity Unit costs.

Total annual operating cost for Unit: £272k.

5. Electrical Grid Safety Unit (in BEIS, or Cabinet Office Civil Contingencies Secretariat)

Potential projects (see Annex E for details):

- Conducting a comprehensive evaluation of the specific actions required to increase the resiliency of the grid against the likely cascading impact from both natural threats (terrestrial storms, solar storms) and manmade threats (cyber, physical attack, and electromagnetic pulses).

Employee	Annual Salary	Staff Costs (NICS and pension)	Office Costs	Total Cost for Employee
Grade 6 - Head of Unit	£64,500	£9,086	£29,434	£103,020
Grade 7 - coordination lead with Office for Risk Management and wider Department	£49,700	£6,700	£22,560	£78,960

Table 6: Electrical Grid Safety Unit costs.

Total annual operating cost for Unit: £182k.

6. Extreme Risks Research Unit (in UKRI)

Potential projects (see Annex E for details):

- More research into emerging technologies in areas such as AI safety, biosecurity and forecasting accuracy (NB would need to be funded separately).

Employee	Annual Salary	Staff Costs (NICS and pension)	Office Costs	Total Cost for Employee
Grade 6 - Head of Unit	£62,404	£8,747	£28,460	£99,611
Grade 7 - Key research project 1 (e.g. AI safety research)	£49,700	£6,700	£22,560	£78,960
Grade 7 - Key research project 2 (e.g. improved forecasting techniques)	£49,700	£6,700	£22,560	£78,960

Table 7: Extreme Risks Research Unit costs.

Total annual operating cost for Unit: £248k.

7. Extreme Risk Management Unit

(in Cabinet Office, Civil Contingencies Secretariat, or Treasury)

Potential projects (see Annex E for details):

- Revising the Green Book's discount rate and ensuring the Treasury adopts key recommendations on intergenerational fairness
- Reforming the National Security Risk Assessment and National Risk Register reform as per the recommended changes set out in Annex E.

Employee	Annual Salary	Staff Costs (NICS and pension)	Office Costs	Total Cost for Employee
Grade 6 - Head of Unit	£64,500	£9,086	£29,434	£103,020
Grade 7 - coordination lead with Office for Risk Management and wider Department	£49,700	£6,700	£22,560	£78,960

Table 8: Extreme Risk Management Unit costs.

Total annual operating cost for Unit: £182k.

8. International Extreme Risk Management Unit (in FCDO)

Potential projects (see Annex E for details):

- Exploring how the UK can play a global leadership role in the management of extreme risks, for example by increasing the capacity of the International Atomic Energy Agency to verify that nations are complying with safeguarding agreements, and leading calls for the creation of a new Treaty on the Risks to the Future of Humanity.

Employee	Annual Salary	Staff Costs (NICS and pension)	Office Costs	Total Cost for Employee
Grade 6 - Head of Unit	£59,500	£8,280	£27,112	£94,892
Grade 7 - coordination lead with Office for Risk Management and wider Department	£48,500	£6,507	£22,002	£77,009

Table 9: International Extreme Risk Management Unit costs.

Total annual operating cost for Unit: £172k.

Second Line (CRO and Office for Risk Management) Costs

Total: £999k annually.

Employee	Annual Salary	Staff Costs (NICS and pension)	Office Costs	Total Cost for Employee
CRO	£120,000	£18,033	£55,213	£193,246
Chief of Staff - Grade 6	£64,500	£9,086	£29,434	£103,020
Head of External Affairs - Grade 6	£64,500	£9,086	£29,434	£103,020
HR Business Partner - Grade 7	£49,700	£6,700	£22,560	£78,960
AI Policy, Grade 7 (coordinating between CRO and relevant Unit)	£49,700	£6,700	£22,560	£78,960
Biosecurity Policy, Grade 7 (coordinating between CRO and relevant Unit)	£49,700	£6,700	£22,560	£78,960
Climate Policy, Grade 7 (coordinating between CRO and relevant Unit)	£49,700	£6,700	£22,560	£78,960
Defence Policy, Grade 7 (coordinating between CRO and relevant Unit)	£49,700	£6,700	£22,560	£78,960
International and Electrical Grid, Grade 7 (coordinating between CRO and relevant Unit)	£49,700	£6,700	£22,560	£78,960
Private Secretary to CRO, HEO	£40,000	£5,137	£18,054	£63,191

Operations and Learning & Development Manager, HEO	£40,000	£5,137	£18,054	£63,191

Table 10: CRO and Office for Risk Management costs.

Total annual operating cost: £999k.

Third Line (Independent Institute) Costs

Total: £1.02 million annually.

Employee	Annual Salary	Staff Costs (NICS and pension)	Office Costs	Total Cost for Employee
Institute Director	£80,000	£11,585	£36,634	£128,219
Research Manager	£59,500	£8,280	£27,112	£94,892
Head of Fundraising	£49,700	£6,700	£22,560	£78,960
HR and Operations Manager	£40,000	£5,137	£18,054	£63,191
AI expert (providing recommendations)	£49,700	£6,700	£22,560	£78,960
Biosecurity expert (providing recommendations)	£49,700	£6,700	£22,560	£78,960
Climate expert (providing recommendations)	£49,700	£6,700	£22,560	£78,960
Defence expert (providing recommendations)	£49,700	£6,700	£22,560	£78,960
Electrical Grid expert (providing recommendations)	£49,700	£6,700	£22,560	£78,960
Government Management of Extreme Risks expert (providing recommendations)	£49,700	£6,700	£22,560	£78,960
Head of Comms	£49,700	£6,700	£22,560	£78,960
Executive Assistant for the Institute Director	£32,000	£3,179	£14,071	£49,250
Administrative Assistant	£32,000	£3,179	£14,071	£49,250

Table 11: Independent Institute costs.

Total annual operating cost: £1.02 million.

Annex E - Costed Recommendations from Extreme Risk Experts

Total annual cost estimate for these recommendations: £4.14 million.

1. Artificial Intelligence

Estimated initial cost of these recommendations (beyond cost of relevant Risk Ownership Unit in Annex D): £2.19 million annually.

Task: Create a pool of machine learning-relevant compute to provide free for socially beneficial applications and AI safety, security and alignment research

Access to huge amounts of AI computational resources ("compute")—for instance, computing clusters of machine learning-optimised computer chips—is critically important for both socially beneficial AI applications, AI safety and security R&D, and for maintaining UK scientific and economic leadership.

Most recent machine-learning breakthroughs and expected advances in this area are reliant on large compute budgets, beyond the current reach of academia and civil society. This has led to research being skewed towards short-term aims (for example, maximising ad click-through) rather than socially beneficial applications or AI safety, security and alignment.

We recommend creating a 'compute fund' to provide free or subsidised machine-learning relevant compute to select researchers working on socially beneficial AI applications or AI safety, security and alignment. This could include:

- Beneficial AI applications (e.g. medical research and diagnostics, and AI for the Sustainable Development Goals)
- AI safety research (e.g. interpretability, interruptibility, and other topics covered in the AI subsection of Section Four below)
- Providing open-source alternatives to commercial AI systems
- Increasing scrutiny of commercial models, including funding replication efforts
- Leveraging AI to test AI: deploying adaptive, automated tests to explore potential failure modes

Such a fund could also help support the UK's AI ecosystem. The compute disparity has led to leading researchers leaving academia for industry, reducing the number of academics available to train the next generation of PhDs, worsening a bottleneck for the UK's AI ecosystem.

Making the playing field more even would incentivise researchers to stay in academia to train the next generation. This could be part of a wider industrial strategy to promote and strengthen the UK's hardware and compute ecosystem, working in concert with policies which channel government support towards this sector, emphasise data localisation, increase the cybersecurity of domestic compute infrastructure, and monitor foreign direct investment.

The compute fund could contribute to this strategy by being encouraged to purchase its compute from data centres based in the UK. This would boost the UK's AI ecosystem and cement its status as one of the strongest in the world.

Estimated cost: Not available at this stage. This would need to be costed in detail separately by the relevant Risk Ownership Unit, once up and running.

Task: Improve foresight and progress tracking in AI research

AI capabilities, and their potential applications in society, are growing fast. In order to avoid falling behind and taking an overly reactive approach, we recommend that the UK Government funds or establishes its own capacity to anticipate and monitor AI progress and its implications for society.

This can form the basis of informing policy and regulation that the Government may want to develop in order to manage these societal implications, and particularly mitigate risks of increasingly widely deployed AI applications in critical areas. This could complement, and work closely alongside, initiatives like the OECD AI Observatory and Stanford's AI Index initiative.

Tracking should focus on establishing metrics and mechanisms to assess:

- The impacts of AI and automation domestically and internationally, including public sector, commercial and criminal use of AI and automation
- The positive and negative impacts of AI on the economy, the likelihood of global catastrophic risks, critical infrastructure failure, and achieving the Sustainable Development Goals
- AI talent (PhDs, patents, top papers, brain drain etc.)
- Access to computational resources
- Rate of adoption of AI technologies

Once costed by the relevant Risk Ownership Unit, we also recommend funding:

- A new body (housed perhaps at the Alan Turing Institute, but with close links to Government) which would synthesise existing research and establish metrics and mechanisms to assess progress in AI, its applications and impacts on society.
- Research projects in AI foresight and progress tracking that could be awarded by this new body.

Cost: We recommend budgeting £95k per head for six Grade 6 AI experts to work in the new body full-time, including office costs. This totals £570k (including office costs). An initial fund for the research projects would need to be explored and costed separately by the relevant Risk Ownership Unit, once up and running.

Estimated initial cost: £570k annually.

Task: Bring more technical AI expertise into government through a scheme equivalent to TechCongress and create new AI roles in key departments

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, which is currently sorely lacking.

There are various mechanisms that the UK Government could investigate to bridge this gap in technical expertise, which include:

- Setting up a TechCongress-equivalent scheme (potentially as part of the Cabinet Office Open Innovation Team) aimed at enabling the UK Government to recruit and gain access to AI expertise, both technical and non-technical (in fields like AI governance and ethics). The scheme could place experts in Parliament, but could also embed them within the Civil Service.
- Creating specific roles in, for instance, the MoD, ICO and BEIS. These roles would be targeted at experts in AI, machine learning, and cyber security, and their focus would be on assuring the safety and security of AI systems that are deployed in specific sectors, particularly those that serve critical functions to society (e.g. critical infrastructure, law enforcement, finance and defence).
- Setting up a fund that agencies can apply for to cover salaries of additional technical experts.
- Providing funding for existing civil servants to develop training and expertise in AI or machine learning. The Treasury currently provides scholarships for civil servants to study economics; an equivalent scheme should be devised for AI.

Canada's Vector Institute is an interesting example to draw on. It is focused on retaining and developing AI talent in Canada.

Cost: There are multiple ways to fund this recommendation. Our preferred model would be to hire two full-time Grade 7s to run the scheme (£164k annually), reporting to a Grade 6 (this would take 20% of their time, which is £19k annually).

To attract top talent into the most crucial newly created AI roles, we would want to budget for ten full-time Grade 7s (£938k, including office costs) and six full-time Grade 6s (£498k, including office costs) across Government.

The additional secondments and training where needed would need to be costed separately by the relevant Risk Ownership Unit, to be signed off by the CRO and the relevant Permanent Secretary (N.B. universities may be able to fund some of this work, for instance through PhD secondments into government).

Estimated initial cost: £1.62 million annually.

2. Biosecurity

Estimated initial cost of these recommendations (beyond cost of relevant Risk Ownership Unit in Annex D): £318k annually.

Create a biosecurity liaison officer to improve coordination between the biosciences and security communities.

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. Edward You currently holds such a role in the United States.

Cost: the Liaison Officer would be a Grade 3 role (£193k, including office costs). They would also need a budget of £125k per annum for convening the biosciences and security communities and commissioning papers.

Estimated initial cost: £318k annually

Task: Ensure that all DNA synthesis is screened for dangerous pathogens, and regulate DNA synthesis machines

Unless active controls are present, gene synthesis machines can provide a way for individuals to get their hands on dangerous or novel pathogens. Gene synthesis companies should be required to adhere to biosecurity guidelines, such as those released by the Secure DNA Project, for screening DNA orders for dangerous pathogens.

These guidelines go beyond the most commonly used International Gene Synthesis Consortium protocol to reflect rapid advancements in the field and current technological capabilities. 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 discreetly push for domestic and international regulation in this area.

Further, the UK should consider regulating DNA synthesis machines and encourage other countries to do likewise. The Government should also analyse how to restructure the sector so that in future, DNA synthesis is available as a service only from a handful of service providers worldwide, and that DNA synthesis machines cannot be purchased without a license.

We would also suggest that the UK introduces licensing requirements for DIY biotech labs. This would enable safe scientific innovation and security screening in a community over which there is currently little oversight.

Estimated initial cost: zero: a policy change to be explored by the relevant Risk Ownership Unit, once up and running.

Task: Pioneer clinical metagenomics in the NHS

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 could routinely send in a sample from any case in the UK that they cannot diagnose with standard techniques.

A central lab could perform the metagenomic sequencing and respond with the closest matches for approximately £100 per sample. This would be extremely helpful for both regular diagnoses and for novel pathogens. In the case of COVID-19, metagenomics would have immediately shown that the closest match was SARS, but that it was sufficiently different to be a novel SARS-like pathogen.

While the rollout of this technology has been suggested for US hospitals, it has yet to be taken up or adopted by any nation globally. The NHS provides an excellent launchpad to pioneer and develop this capability, and the UK possesses world-leading expertise. In the next five years, the UK should be deploying metagenomic diagnostics nationwide.

Estimated cost: Not available at this stage. This would need to be costed in detail separately by the relevant Risk Ownership Unit, once up and running.

3. Extreme Climate Change

Estimated initial cost of these recommendations (beyond cost of relevant Risk Ownership Unit in Annex D): zero

Task: Ensuring that UK climate policy focuses sufficiently on mitigating the 'tail' scenario (approx. 5% probability) of a more-than-six-degree rise in global temperatures.

If current policies remain in place, global temperature rises will probably exceed 3°C. However, there is significant uncertainty about how the climate will respond to emissions. Based on mainstream estimates of climate sensitivity, there is roughly a one-in-twenty chance of an increase in temperatures exceeding six degrees, which would result in significantly worse outcomes.

In order to ensure that these worst-case scenarios are avoided, we need to explore interventions that are effective under those circumstances.

The most plausible strategy to minimise this risk is to prioritise interventions that reduce global 'carbon intensity'—either through increasing energy efficiency (energy / GDP) or reducing the carbon intensity of energy (carbon / energy).

One idea we recommend exploring is enacting a carbon fee and dividend scheme on transport and heating fuels. This scheme would place a tax ('fee') on emissions of carbon dioxide, and then redistribute the revenue on a flat per-person basis (the 'dividend') or, alternatively, compensate by increasing the tax-free personal allowance.

An economy-wide instrument would be ideal, but we recognise there is a commitment to continued Emissions Trading Scheme (ETS) for current ETS sectors post-Brexit, and that the most feasible approach is therefore likely to be a fee on transport and heating fuels.

For sectors not covered by ETS, a carbon fee would provide an effective and cost-efficient price signal to decarbonise. Such an approach has been endorsed by Policy Exchange, and over 3,500 economists, including 27 Nobel Laureates—the largest ever public statement of economists.

Of course, new taxes in a difficult economic climate following a public health emergency would normally be a difficult sell. But support can be created by returning the revenue generated as a carbon dividend, and by paying out this dividend in advance of putting the fee in place. The dividend can also be "paid forward" as part of a stimulus, and there is even a "double dividend" because the carbon price also makes the binding UK climate target cheaper to reach for society as a whole.

By being the first major economy to endorse the "fee and dividend" model, the UK would continue a proud tradition of climate policy leadership ahead of COP26, and this idea could be a flagship part of the UK's 'Build Back Better' agenda post-COVID-19.

For further details, see this Policy Exchange paper and this Climate Leadership Council paper.

Estimated initial cost: zero—a policy change to be explored by the relevant Risk Ownership Unit once up and running

4. Defence and Cyber Security

Estimated initial cost of these recommendations (beyond cost of relevant Risk Ownership Unit in Annex D): £1.63 million annually.

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

As evidenced by the history of nuclear missiles, 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 MoD investigates the process that the UK would need to undertake to make a credible commitment that it will not incorporate AI systems into NC3 before then making this commitment.

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

Estimated initial cost: zero—a policy change to be explored by the relevant Risk Ownership Unit, once up and running.

Task: Establish a new Defence Software Safety Authority as a sub-agency of the Defence Safety Authority

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).

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 of those algorithms becomes ever more important to avoid accidents that could harm servicepeople or citizens, or lead to inadvertent escalation.

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 could 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:

1. Improving systemic risk assessment in defence procurement. Systemic risk assessment should include a range of questions we can advise on, including:

- If the technology in question was first developed in the private sector, are there adversarial threats that the system is unlikely to be designed to resist?
- If all systems of this type fail at the same time, what would be the effect on national defence?
- If a black-box system or an upstream supply chain phase is compromised by an adversary, what is the worst thing they could do?

2. Ensure clear lines of responsibility so that senior officials are held responsible for errors caused in the defence procurement chain.

Advocates for new and experimental systems and contracts must be held personally accountable (and know that they are accountable) throughout the lifetime of the systems procured, in addition to the accountability of operators, commanders and developers.

The Defence Safety Authority could oversee the regulation of experimental AI systems in defence and ensure clear lines of responsibility.

Cost: Two Grade 6s (£202k annually, including office costs) and Four Grade 7s (£336k annually, including office costs).

Estimated initial cost: £538k annually.

Task: Create an independent cyber-security red team to conduct frequent scenario exercises

We would recommend setting aside funding to maintain an independent and persistent 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 cyberattack on UK infrastructure
- The non-availability of one or more major cloud providers in the UK for an extended period of time
- Cut-off from the internet for an extended period of time

Funding for six experts would be sufficient. This would help ensure that the most important scenario exercises are conducted frequently, and that clear lessons learned are 'owned' by senior policy makers.

Cloud Down from Lloyd's of London is an example of a scenario exercise that could be run. This guidance from the Belfer Center on how to run a cyber war game may also be useful. Finally, the Centre for Doctoral Training in Cyber Security at Royal Holloway, University of London, has supervised at least one thesis paper on cyber wargames.

Cost: Two Grade 6s (£202k annually, including office costs) and Four Grade 7s (£336k annually, including office costs), plus a contractor budget of £250k annually.

Estimated initial cost: £788k annually.

Task: Set up throughout-lifetime stress-testing of computer and AI system safety and security

This stress-testing should be done during development, testing, training, early deployment, at regular intervals, and before retirement of relevant systems. We recommend having dedicated personnel who work to expose software and hardware vulnerabilities and design adversarial environments.

We also recommend making adversarial testing and red-teaming part of military exercises (whilst making sure to avoid misinterpretation of test actions).

Cost: Two Grade 6s (£101k each annually, including office costs).

Estimated initial cost: £202k annually.

Task: Run more AI cyber security guidance and training

Software development and deployment tools now include an array of security-related capabilities, including testing, fuzzing (the use of machine learning and similar techniques to find vulnerabilities in an application or system) and anomaly detection.

The National Cyber Security Centre should provide guidance and training on cyber security of systems using AI and machine learning.

Cost: 50% of one Grade 6's time (£50.5k annually, including office costs) should cover the costs to develop guidance before embedding this into existing training programmes run in the new College of National Security, along with the same amount of time for a Grade 6 National Cyber Security Centre employee's time delivering in-person training.

Estimated initial cost: £101k annually.

Task: Update the Ministry of Defence's definition of "Lethal Autonomous Weapons Systems"

Within the wide set of defence and defence-adjacent 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 MoD's definition of "Lethal Autonomous Weapons Systems" is quite different from that used by many other nations. It is idiosyncratic, as 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".

This is a very high bar to reach—almost human-level intelligence—and is so high as to be almost meaningless in this context. No system currently under research or development would be capable of meeting this definition. This is out of step with the definitions used by most other governments, which limits the UK in its ability to consider and protect against foreseeable risks associated with these systems, and to set international standards for this emerging technology.

In particular, the UK should work towards making sure there is a clear, internationally accepted definition of "autonomous weapon" so that dialogue can move forward. The Integrated Security, Defence and Foreign Policy Review provides an excellent opportunity to update this definition.

Estimated initial cost: zero—a policy change to be explored by the relevant Risk Ownership Unit, once up and running.

5. Electrical Grid Safety

Estimated initial cost of these recommendations (beyond cost of relevant Risk Ownership Unit in Annex D): zero.

Task: Increase the resilience of the UK's electrical grid against extreme terrestrial and solar storms, man-made electromagnetic pulses and malicious digital intrusions

The UK's century-old electrical grid is alarmingly fragile, and vulnerable to a myriad of threats that could result in sustained outages. The grid must be proactively secured against this dynamic threat landscape.

If the electrical 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.

The proposed funding would be devoted to conducting a comprehensive evaluation of the specific actions required to increase the resiliency of the grid against the likely cascading impact from both natural threats (terrestrial storms, solar storms) and manmade threats (cyber attacks, physical attacks, and electromagnetic pulses).

This effort should produce specific policies, procedures and technological solutions, together with implementation timelines and an estimate of required resources. It should include plans of action 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.
- Increase 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.

See this recent case study from the United States, which helped bring about the President's Executive Order on EMPs, and the National Defense Authorization Act for Fiscal Year 2020.

Estimated cost: Not available at this stage. Costs beyond the initial exploratory work by the Risk Ownership Unit would need to be costed in detail separately by that Unit, once up and running.

6. Increased Research Funding for Extreme Risks

Fund high-priority biosecurity, AI and other high-impact R&D projects

Estimated initial cost of these recommendations (beyond cost of relevant Risk Ownership Unit in Annex D): zero.

Task: High-priority biosecurity R&D projects

Specific research ideas:

- Enable ubiquitous pathogen agnostic detection in all clinical environments. This can be accomplished by:
    - Developing on-chip microfluidic devices that automate sample preparation for nucleic acid diagnostics (e.g. sequencing, PCR, and CRISPR Dx)
    - Integrating reagent-free sample preparation techniques into the on-chip microfluidics
    - Increasing funding for UK-based genetic sequencing companies to further decrease the raw sequencing costs
    - Developing bioinformatic pipelines and technical interfaces that require no expertise to run or analyse the output
- Environmental pathogen biosurveillance:

    - Programmes to bring this to fruition should focus on automation of sample preparation, reagent-free diagnostics, with a specific focus on DNA sequencing, and CRISPR diagnostics
    - Next-generation metagenomics (pathogen-agnostic infectious disease metagenomics); see Recommendation 1B above for more information
    - More robust forms of sequencing that can handle complex analytes
    - Development and miniaturisation of sample acquisition and preparation technologies
- Safe synthetic biology (e.g. programmes with goals similar to DARPA's Safe Genes)
- Non-pharmaceutical medical countermeasure R&D for large or fast-moving pandemics
- Non-invasive pathogen-agnostic infection detection methods
- Rapid scale-up of therapeutics, such as monoclonal antibodies
- Pre-discovered and rapidly adapted broad-spectrum small-molecule antivirals
- Safe Bio Spaces, with a focus on BSL-4 security labs to make them safer
- Technologies for securing physical spaces and preventing transmission in high-human-traffic-flow environments (e.g. safe airports, planes and trains)
- RNA vaccine platforms developed to the level of viral families for rapid adaptation, scale up, and distribution for novel threats
- Transmission-suppression technologies for the built environment (sterilisation, self-disinfecting, and neutralizing technologies)
- Developing better tools for DNA synthesis screening (the technical aims should include accuracy under 200 base pairs and prediction of sequence/pathogen from oligonucleotides, with the stretch goal being prediction of pathogenicity in novel pathogens from sequence data alone)
- Systems engineering to better integrate processes for continual biosecurity
- Innovative and improved face masks, ventilation / air filtration, UV sterilization, food sterilization, etc.
- Comprehensive, constant, real-time global bioinformatic pathogen surveillance
- Bioinformatic and physical genomic and microbial forensics technologies.

Task: High priority AI safety R&D projects

Promoting technical AI safety research is critically important not only due to the negative externalities of unsafe systems, but because it will bolster the UK's competitiveness as the EU advances its Trustworthy AI legislative agenda.

We therefore recommend the UK funds technical AI safety research. This would ideally be done through the new UK ARPA, but it could also be done via the Alan Turing Institute, or through the newly proposed autonomous systems research hub at Southampton University.

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

1. Alignment: Most AI systems today are trained to optimise a well-defined objective (e.g. reward or loss function). This works well in some research settings where the intended goal is very simple (e.g. Atari games, Go, and some robotics tasks), but for many real-world tasks that humans care about, the intended goal or behaviour is too complex to be specified directly. 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—e.g. the oversimplified goal of "maximise the amount of money in this bank account" could lead a system to commit crimes. If we could instead learn complex objectives, we could apply techniques like reinforcement learning to a much broader range of tasks without incurring these risks. Can we design training procedures and objectives that will cause AI systems to learn what we want them to do?

2. Robustness: Most training procedures optimise a model or policy to perform well on a particular training distribution (data set). However, once an AI system is deployed, it is likely to encounter situations that are outside the training distribution or that are adversarially generated in order to manipulate the system's behaviour, and it may perform arbitrarily poorly on these inputs. 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?

3. Interpretability: Trained models are often extremely large, complex, and opaque. If a models' internal workings could be inspected and interpreted, or if we can develop tools to visualise or analyse the dynamics of a learned model, we might be able to better understand how models work, which changes to inputs would result in changed outputs, how the model's decision depends on its training and data, 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. The ability to interpret a system's decision-making process may also help significantly with validation or supervision; for example, if a learned reward function is interpretable, we may be able to tell whether or not it will motivate desirable behaviour, and a human supervisor may be able to better supervise an interpretable agent by inspecting its decision-making process.

4. Assurance of deep learning systems: It is not enough for models to be robust and have the right reward function. In particularly high stakes situations—defence applications, AI systems integrated into the power grid, and the like—we also need assurance that this is the case. However, traditional Testing & Evaluation, Verification & Validation methods for gaining such high assurance typically cannot be applied to deep learning systems. New methods need to be developed.

Further details about these potential research areas are available in this Open Philanthropy article, in the descriptions of the US DARPA XAI, and in Towards Trustworthy AI.

Task: Neglected but high-impact areas (natural risks, meat alternatives to reduce zoonotic disease risk, and accuracy of long-term forecasts)

Further research could also be commissioned in the following three areas.

1. Supporting meat alternatives to reduce zoonotic disease risk and carbon emissions. COVID-19 has its origins in animals eaten for food. Zoonotic diseases - including avian and swine influenza - pose serious risks of sparking the next major pandemic. Industrial animal agriculture provides an ideal breeding ground for zoonotic diseases, as well as driving antimicrobial resistance and increasing greenhouse gas emissions.

In light of growing global demand for meat, there is an urgent need to replace conventional meat and animal protein with plant-based, fermentation-derived and cultivated alternatives.

The UK Government should provide large-scale public funding for research into plant-based and cultivated meat and commit to make post-Brexit Britain a world-leader in this emerging sector. This also offers major commercial opportunities for UK PLC.

2. Improving the accuracy of long-term forecasts. 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 Philip Tetlock.

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 matched reality and how they can improve. Such techniques could be used to improve how the UK predicts the probabilities of future disasters. This could in part be similar to the work done in the US by IARPA's intelligence community prediction market.

3. Risks into neglected natural risks. The UK should commission research into the following low probability, highly destructive risks which are currently significantly under-researched:

- Asteroids and comets
  - Research the deflection of 1 km+ asteroids and comets, perhaps restricted to methods that couldn't be weaponised, such as those that don't lead to accurate changes in trajectory
  - Bring short-period comets into the same risk framework as near-Earth asteroids
  - Improve our understanding of the risks from long-period comets
  - Improve our modelling of impact winter scenarios, especially for 1–10 km asteroids. Work with experts in climate modelling and nuclear winter modelling to see what modern models say.
- Supervolcanic eruptions
  - Find all the places where supervolcanic eruptions have occurred in the past
  - Improve the very rough existing estimates on how frequent these eruptions are, especially for the largest eruptions
  - Improve our modelling of volcanic winter scenarios to see what sizes of eruption could pose a plausible threat to humanity
  - Liaise with leading figures in the asteroid community to learn lessons from them in their modelling and management.

Funding: We recommend allocating annual funding for a research fund across these research areas, to be awarded by the Extreme Risks Research Unit and signed off by the CRO.

Estimated cost: Not available at this stage. This would need to be costed in detail separately by the relevant Risk Ownership Unit, once up and running.

7. Improved Extreme Risk Management

Estimated initial cost of these recommendations (beyond cost of relevant Risk Ownership Unit in Annex D): zero.

Task: Revise the Green Book's discount rate, and ensure the Treasury adopts key recommendations on intergenerational fairness

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. This IfG paper 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 the Treasury:

- Revises the Green Book and the discount rate. The Green Book should have more detail on how to account for second-order effects. The discount 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).
- Adopts the suggestions from The House of Lords' Intergenerational Fairness and Provision Select Committee, including publishing long term statistical trends and forecasts, adopting Intergenerational Impact Assessments, and introducing a long-term fiscal rule that ensures that spending is maintained at a reasonable level for the whole of the Government balance sheet.

Estimated initial cost: zero—a policy change to be explored by the relevant Risk Ownership Unit, once up and running

8. International Risk Regulation

Estimated initial cost of these recommendations (beyond cost of relevant Risk Ownership Unit in Annex D): zero.

Task: Lead international calls for a new Treaty on the Risks to the Future of Humanity

Guglielmo Verdirame QC has called for a new Treaty on Risks to the Future of Humanity, with a series of UN Security Council resolutions to place this new framework on the strongest legal footing. Verdirame argues that the post-coronavirus settlement coming out of any international investigation into China's actions should have a larger purpose than simply addressing China's role in the coronavirus disaster.

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.

Verdirame argues that a new 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.

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.

Estimated initial cost: zero—a policy change to be explored by the relevant Risk Ownership Unit, once up and running.