Adam Smith
Institute

# Tipping Point

## On the edge of superintelligence

Connor Axiotes and Eddie Bolland

# About the Authors

**Connor Axiotes** is the Strategic Communications lead at Conjecture AI, the largest AI safety company in the UK. He was previously Director of Communications at the Adam Smith Institute. Before, he worked as a senior communications officer for a member of Parliament.

**Eddie Bolland** was a Research Assitant at the Adam Smith Institute. He is currently a first year Politics, Philosophy, and Economics undergraduate at Merton College, University of Oxford.

Adam Smith Institute, 23 Great Smith Street, London, SW1P 3DJ, 02072224995, info@adamsmith.org

# Executive Summary

- The advent of an artificial superintelligence (ASI) could change the United Kingdom and the world for the better and in ways we cannot yet imagine.

- Even less advanced AI systems that do not reach ASI could still spur annual economic growth by 30%.[1] These AI tools might help us cure disease, rid the world of the scourge of poverty, provide a deterrence against foreign adversaries, and help the UK to reach its net zero commitments.

- Despite AI's potential for transformational improvement, its dual-use nature also means there is the possibility of poor outcomes. For every life-saving vaccine an AI helps us create, manufacture, and administer, there might also be rogue actors who develop life-threatening synthetic bioweapons.

- Today's Large Language Models are said to be a 'black box', and even their creators do not fully understand what is going on inside. It means we may not know how to fully control them as they scale.

- The UK has all the ingredients to become an 'AI Superpower'. It is an attractive place for innovative start-ups, with world-leading AI talent and infrastructure. We should lead the world in creating an innovation friendly permissive regulatory regime.

- To spur innovation in AI, some of our key recommendations include:

    - Creating a privately-provided 'British Compute Reserve', pre-committing to cloud compute from a best-of-breed framework of vendors to be used for AI R&D and deployments by government departments and nonprofits.

    - Allowing SMEs to access larger models through APIs to promote innovation and increase contestability.

    - Investing in generalist medical AI capabilities through the NHS AI Lab.

    - Expanding high-skilled visas, especially for technical AI-related vocations.

    - Returning to an internationally competitive corporation tax rate.

- To make sure AI development is safe and systems are benevolent, we propose:

    - Creating a multilateral International Agency for Artificial Intelligence within which the UK takes a leading role.

    - Creating a multilateral monitoring system for third-party audits of the largest and highest-risk AI systems and their 'compute.' This should come under the new Frontier AI Taskforce's remit.

    - Launching a brand new 'The Great British Prize in AI' for open research questions in AI safety for example, in mechanistic interpretability research.

    - Leading on the authoring of a UN 5 Powers (P5) statement on air-gapping WMD facilities from autonomous AI systems.

[1] T. Davidson., (2021), Could Advanced AI Drive Explosive Economic Growth?, Open Philanthropy, https://www.openphilanthropy.org/research/could-advanced-ai-drive-explosive-economic-growth/#3-summary

# RECOMMENDATIONS

1. **Invest in British Computing Resources**

   a. Allow our existing British public computing power and new exascale capacity to be used by our world-leading universities for AI safety work - because universities are being crowded out by private labs with much more access to 'cloud compute.'

   b. The introduction of a new 'British Compute Reserve.'

2. **Setting the UK up for Success**

   a. Planning reform - build on the green belt, and implement street votes to make the UK a country high-skilled AI safety researchers want to live in.

   b. Lower corporation tax to an internationally competitive level so that more AI companies want to set up here in the UK.

3. **Create a Public Comprehensive AI Monitoring System**

   a. Begin to monitor the largest AI models. A multilateral and unobtrusive monitoring of lab training runs would systematically track their capabilities and the extent of their alignment, to make sure innovative AI systems are safe and have few harmful emergent properties.

   b. Making third-party external audits mandatory for largest/riskiest lab training runs.

4. **UK to Lead the World in International Agreements on the Safe Deployment of Advanced AI Systems**

   a. The UK should take the lead on creating an International Agency for AI (IAAI).

   b. A P5 statement on air-gapping nuclear weapons facilities from AI to reduce the chance of accidental nuclear strikes.

   c. Lay out the structure and objectives of Bletchley Park's 2023 AI Safety Summit.

4. **Expand Educational Grants and High-Skilled Visa Scheme**

   a. Increase youth engagement in STEM through tax-credits to private companies to address long term skills shortages.

   b. Eliminate obstacles to obtaining the High Potential Individual visa.

   c. Align High-skilled Visa schemes with the priorities of prospective applicants to maintain the UK's position as a global leader in attracting AI talent.

   d. Expand university courses alongside changing patterns of demand for priority areas to prevent future skills shortages.

   e. Integrate the Adam Smith Institute's model for visa auction markets.

5. **Regulatory Markets for AI**

   a. The UK should utilise 'Regulatory Markets' - private regulatory experts to bring their experience in helping with safety-based, innovation-inducing AI legislation. This would help to solve the knowledge gap between the government and the relevant regulatory body.

6. **Government Investment in AI Safety**

   a. The Great British AI Prizes: cash prizes for open research questions in AI safety, such as 'how do we stop larger models from hallucinating?'

   b. If sovereign capabilities such as a public LLM are sought after, then AI alignment researchers and academics should be able to access them for safety work.

7. **Facilitate the Safe Use of APIs for Innovative SMEs and Researchers**

   a. Enable SMEs and researchers to develop products and carry out safe research through APIs accessed on the research resource.

   b. Implement risk based requirements for API access to reduce the risk of misuse and encourage private participation.

8. **Effective Procurement to Increase Efficiency and Innovation**

a. Introduce Challenge Based Procurement to improve the efficiency and reduce the barriers for smaller firms.

b. The Office for AI should identify opportunities for procurement to support proof of concept work too risky for nationwide deployment.

c. Procurement for AI assurance within the public sector to support private sector firms and ensure safe deployment.

9. **Saving Lives with AI-Powered Medicine while Reducing Engineered Pandemic Risk**

a. The NHS should invest in Generalist Medical AI capabilities through the NHS AI Lab.

b. Introduction of Three Lines of Defence Structure to ensure the UK is proactively prepared for biosecurity risks.

c. Invest in pathogen monitoring systems and introduction of bio-engineering licences.

10. **Implement a Review on the Possible Labour Effects of Future AGI**

a. Commission a White Paper on what the introduction of a universal basic income (UBI) or a negative income tax (NIT) would look like in a worst-case scenario;

b. Introduce NIT and UBI trials to prepare for the possibility of AI caused unemployment.

# INTRODUCTION

The Adam Smith Institute has long been at the forefront of developing policies that promote technological advancements. The default position is to welcome transformational technologies, as they generally lead to the creation of huge wealth and empowerment.

Artificial superintelligence (ASI) is a transformative technology, but a different kind from the likes of the combustion engine, the printing press, and the nuclear weapon. An ASI's difference lies in what it could become - "agentic" - i.e., have goals of their own and act in the world of their own accord.

This is potentially dangerous because we do not know if these goals will align with our values: its auxiliary goals (instrumental or given goals) may seek to harness all our resources for its own ends. The AI safety field calls this the 'alignment problem.'

The vast majority (nearing 99%) of AI models are non-harmful and increase productivity, safety, and even leisure. There are concerns that a small number of applications of future ASI are particularly high-risk if not aligned to human values (whatever those might be).

Current Large Language Models (LLMs) are powerful and have been deployed around the world and on the internet, tinkered with by eager third-party coders, and have roused much public and Government interest. The problem is that, unlike most technology humanity has created, the engineers behind LLMs do not well understand how LLMs function internally.

There are 10,000s of top AI researchers working on ASI and 'capability' work and there are less than 300 people working on AI safety directly. The scalable alignment team at OpenAI had just ~7 people.[2] At least by crude ratios, AI safety is not being taken sufficiently seriously .

This paper is focussed on developing an innovative infrastructure and unobtrusive regulatory regime so that the UK's AI industry can prosper and the UK can become an AI Superpower. This paper provides innovation-based and safety-focussed policies to deploy aligned, safe and innovative AI and superintelligence.

[2] Aschenbrenner. L., (2023) Nobody's on the ball on AGI alignment, Squarespace, https://www.forourposterity.com/nobodys-on-the-ball-on-agi-alignment/

# DEFINITIONS

**Alignment** - aligning advanced AI models is referring to the ability of the developers to ensure their systems are 'aligned' to human values. It is the extent to which the AI does as the creator/user intended . When it comes to an AGI, this would mean despite its superior intellect, it would not harm nor seek to disempower humanity. AI safety researchers call this the 'alignment problem.'

**Application Programming Interface (API)** - APIs provide partial access to the models where users can submit inputs and then see how the AI system responds. Large AI labs like OpenAI have distributed their model via APIs to power new AI products. For example, the popular foreign language learning app, Duolingo, utilises GPT-4 in its new Duolingo Max.[3]

**Artificial General/Super Intelligences (AGIs/ASIs)** - are AI systems which (at minimum) exceed competency in every task domain they share with humans: they are better than humans at cognitive tasks. An AGI would also be capable of behaving super intelligently over several domains. An AGI would have the flexible ability to tackle all new tasks in an economy more effectively and efficiently.

**Closed-source AI Models** - the underlying code and algorithms are kept private and proprietary. The model architecture and training data are not disclosed. Only the outputs and results of the models are made available publicly through an application programming interface or as a software product. The model author retains significant intellectual property rights and control over the model. Updates and modifications to the model are performed by the model author.

**Compute** - refers to the processing power and memory required to train, evaluate, and deploy state-of-the-art AI models. As these models have been increasingly developed with deep learning techniques, the demand for compute has increased. Top AI labs are finding that larger compute-powered training runs deliver more accurate and powerful models, and this insight is called 'scaling laws.'

**Dual-use** - in the context of transformative technologies, dual-use means that they can be used for both benevolent and destructive purposes. While nuclear reactors can create green, abundant energy, nuclear technology has been harnessed to produce devastating weapons.

**Existential Risk (X-Risk)** - is the risk of a catastrophe so huge, that humanity never recovers. Toby Ord, a Philosopher and Senior Research Fellow in Philosophy at Oxford University's Future of Humanity Institute, explains that with the advent of nuclear weapons, "humanity entered a new age, gaining the power to destroy ourselves, without the wisdom to ensure we won't."[4] Misaligned artificial superintelligence might have the power to do the same if we do not ensure its safe deployment.

**Emergent Properties** - these arise from the interactions between an AI system's components, especially when those characteristics are not explicitly programmed into the system or predicted by the designers. These properties can be positive or negative, depending on the context and the goals of the AI system.

In the context of artificial superintelligence (ASI), the term is typically used to refer to the possibility that an

[3] Duolingo, (2023), Introducing Duolingo Max, a learning experience powered by GPT-4, https://blog.duolingo.com/duolingo-max/

[4] The Precipice, Author, https://theprecipice.com/author

ASI system might display harmful or unintended behaviours as a result of its learning processes, interactions with the environment, or internal dynamics. The concern is that an ASI, which would be capable of learning and reasoning across a wide range of domains, might develop harmful behaviours if its goals are not properly 'aligned' with human values.

**Foundation Model** - refers to AI systems with broad capabilities such as GPT-4, or 'for the original ChatGPT, an LLM called GPT-3.5 served as the foundation model.'[5]

**Instrumental Goals** - AI models could have goals and objectives of their own, or at least goals and objectives given to them by their developers. In pursuing these goals, advanced AI models may also chase adjacent goals simultaneously, that are helpful in their achievement of the original goal. For example, having additional resources and not being switched off will help an AGI to pursue their ultimate goal. There is a fear that instrumental goals might not be aligned to human values.

**Large Language Models (LLMs)** - are AI models designed for natural language processing tasks: they predict text. The LLMs are trained on vast amounts of textual data, enabling them to generate human-like text, answer questions, provide summaries, translate languages, and perform various other language-related tasks. Major AI labs, such as OpenAI's ChatGPT and GPT-4,[6] and Anthropic's Claude 2 are the frontier LLMs as of 2023.[7]

The next frontier of these LLMs will be multi-modal, meaning both their inputs and outputs will be not just text and image, but voice and video, too. If linked to the correct machinery, 3D printers could, for example, manufacture a product based on a voice prompt. The next generation models will likely be able to interact fully with the internet, and third-party programmers will have access to an even more powerful foundation model and functionalities.

**Mechanistic Interpretability** - seeks to ascertain the inner workings of the models. The key goal is to really understand the model's cause-and-effect mechanics in an intuitive and understandable way to humans. In short, mechanistic interpretability is about figuring out the 'nuts and bolts' inside an AI to fully comprehend why it does what it does.

**Open-source AI Models** - model code is openly accessible and available for anyone to read, modify, and use to train their own models. The model author releases the model under an open source licence, granting users rights to modify and redistribute the model under certain conditions.

**Red Teaming** - a method of safety testing in which organisations take an adversarial approach to try and induce failure modes. This is done by attempting to find security weaknesses and methods of exploitation to allow organisations to take pre-emptive action and prevent future exploitation.

**Reinforcement Learning from Human Preferences (RLHF)** - a technique to train AI models in complex tasks. In RLHF, a model learns to optimise its behaviour (behave in a way more aligned to the human user) by receiving feedback from human evaluators, who rank the agent's actions or trajectories based on their desirability. RLHF has been shown to be effective in cases where it is difficult to design a precise reward

[5] Toner, H., (2023), What Are Generative AI, Large Language Models, and Foundation Models?, CSET, https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/

[6] GPT-4, (2023), Open AI, https://openai.com/research/gpt-4

[7] Introducing Claude, (2023), Anthropic, https://www.anthropic.com/index/introducing-claude

function or when human intuition is valuable for solving a task. However, even Sam Altman, CEO of OpenAI, admits that, as RHLF stands (and how it is currently used for GPT-4), it is insufficient to safely and reliably align their systems to the desired values or those values that keep humanity safe in the foreseeable future.[8] As models scale, this task will become increasingly difficult.[9]

**Transformative Technology** -[10] a technology that shapes future human affairs. The reason this technology is different from the combustion engine, the printing press, and the nuclear weapon, is that sophisticated enough AI systems could be agentic (have its own goals).

[8] Constitutional AI: Harmlessness from AI Feedback, (2022), Anthropic, https://www.anthropic.com/index/constitutional-ai-harmlessness-from-ai-feedback

[9] Shah. R., (2021), When large models are more likely to lie, AI Alignment Forum, https://www.alignmentforum.org/posts/NR4rgfKu63TcqLxcH/an-165-when-large-models-are-more-likely-to-lie

[10] Ord, T., (2022), Lessons from the Development of the Atomic Bomb, Centre for the Governance of AI, pg. 1

# A BRIGHT FUTURE

A benevolent, suitably aligned, and safe artificial superintelligence (ASI) could transform the UK and the world for the better.

**Growth** - Advanced superintelligent tools could boost productivity. Generative AI at Work, a working paper, finds that access to LLM AI tools 'increase productivity, as measured by issues resolved per hour, by 14 percent on average.'[11] AI tools right now are only the newest in advances, and are not labelled as ASIs as they are not yet superior to human labour in every task.

However, if humanity creates controllable and benevolent artificial intelligence systems, it could bring a huge leap in UK and world GDP growth, the gains from which could decrease poverty here and abroad. At a conservative estimate, it could grow the world economy by an additional 7%,[12] and a more optimistic estimate puts world economic growth at 30% annually,[13] an annual growth rate that has never been seen before.

**Innovation** - As AI continues to develop, it will impact more jobs and likely cause substantial automation of some sectors.[14] However, as history has shown us, the economic gains from this automation will free up capital for investment. Those who lose their job will, over time, re-skill and find new employment,[15] resulting in the development of new jobs and accelerated innovation in different sectors.[16]

Additionally, by automating labour-intensive sectors of the workforce, barriers to setting up new firms will be reduced and innovative ideas will be brought to market more easily, allowing for innovations which were previously inconceivable.

**Energy** - An AGI could help to clear clean energy engineering bottlenecks and design renewable energy sources to bring about super low-cost, high-efficiency power to all households. It could also accelerate the UK's net-zero target, or the aim to decarbonise the electricity system by 2035. For example, smart grids will be able to optimise energy distribution and consumption by minimising waste and reducing greenhouse gas emissions. AI-powered waste management systems could also help with efficient recycling and waste disposal.

**Health** - AI systems could create synthetic vaccines autonomously and rapidly produce them at scale upon the discovery of a new dangerous pathogen. Alongside turning on AI-controlled UV filters in air-conditioning units that kill pathogens before they can spread, this could prevent would-be pandemics at their source.[17]

[11] Brynjolfsson. E., (2023), Generative AI at Work, NBER, https://www.nber.org/system/files/working_papers/w31161/w31161.pdf

[12] Generative AI could raise global GDP by 7%, (2023), Goldman Sachs, https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html

[13] T. Davidson., (2021),Could Advanced AI Drive Explosive Economic Growth?, Open Philanthropy, https://www.openphilanthropy.org/research/could-advanced-ai-drive-explosive-economic-growth/#3-summary

[14] Hatzius. J. Briggs. J. Kodnani. D. Pierdomenico. G., (2023), Global Economics Analyst The Potentially Large Effects of Artificial Intelligence on Economic Growth, Goldman Sachs

[15] Hötte. K. Somers. M. Theodorakopoulos. A., (2022), Technology and jobs: A systematic **literature** review, Oxford Martin School, https://www.oxfordmartin.ox.ac.uk/publications/technology-and-jobs-a-systematic-literature-review/

[16] Lawson, J., (2020), These are the Droids You're Looking For: An Optimistic Vision for Artificial Intelligence, Automation and the Future of Work, Adam Smith Institute, https://static1.squarespace.com/static/56eddde762cd9413e151ac92/t/5fc2124d173fb5383be9ec63/1606554221233/These+are+the+droids+you%E2%80%99re+looking+for+-+James+Lawson+-+Final.pdf.

[17] Feng. Z. Cao. S. Haghighat. F., (2021), Removal of SARS-CoV-2 using UV+Filter in built environment, Elsevier, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8329429/

Advanced AI systems, trained on vast medical datasets, could further revolutionise diagnostics and treatment, identifying diseases in their earliest stages and preventing their onset. Personalised medicine for individuals could become the norm, with AI-driven treatment plans tailored to each individual's unique genetic makeup and lifestyle factors. The development of new treatments and medicines has the potential to dramatically increase life expectancy and significantly reduce the burden of chronic diseases.

**Housing** - One of the reasons we do not build more houses is that local opposition is often vociferous and that even a small number of opposed local actors can block mutually beneficial housebuilding. AI-powered algorithms could theoretically poll large swathes of the population and analyse where housebuilding might be most popular.

3D-printed houses are already here.[18] Building for Humanity, a housebuilding non-profit, believes 3D-printing houses might cut costs by 30%,[19] and that's before particularly advanced AI improved its output by making algorithmic efficiency savings.

**Transport** - Much of modern transport is already on the way to becoming fully automated, with 90% of flight time utilising autopilot and experts predicting autonomous vehicles to become the norm.[20] As this develops there is the potential for significant efficiency gains due to more effective transport systems.[21]

The removal of human error will allow our current systems to work more efficiently, with fewer accidents and obstructions. But AI can also alter the systems, finding more effective ways of running public transport and shaping patterns of transport. Potential savings from avoided accidents, reduced fuel costs and efficiency gains due to automated vehicles alone are estimated to total $1,300 billion in the US annually.[22]

Advanced AI can significantly enhance the capabilities of autonomous vehicles, too, by improving their perception, decision-making, and control systems. Through the use of sophisticated machine learning algorithms, AI can process vast amounts of sensor data in real-time, allowing the vehicle to accurately detect and recognise objects, pedestrians, and other vehicles in complex environments. AI can then enable vehicles to learn from their own experiences and adapt to new situations, ultimately leading to safer passengers and roads.

**Government** - Information asymmetry is a key reason for regular governance failure. Sentiment analysis and natural language processing enables Governments to better understand and respond to their constituents' concerns. As a matter of fact, an adoption of AI in Taiwan's Governance has allowed them to streamline public service delivery and lead in the global development of computer chips,[23] despite lacking UN membership and 'relying entirely on its pool of indigenous resources'.[24] The new Frontier AI Taskforce recognises this

[18] Gira, (2022), First 3D-printed house in Germany: Paving the way for future living, https://www.gira.com/uk/en/g-pulse-magazine/building/3d-house-germany#interior

[19] Home Building, (2022), 3D printed houses to be constructed in the UK for the first timehttps://www.homebuilding.co.uk/news/3D-printed-houses

[20] Cox. J., (2014), Ask the Captain: How often is autopilot engaged?, USA today https://eu.usatoday.com/story/travel/columnist/cox/2014/08/11/autopilot-control-takeoff-cruising-landing/13921511/

[21] 14 Tech Experts Predict Exciting Future Features Of Driverless Cars, (2021), Forbes, https://www.forbes.com/sites/forbestechcouncil/2021/08/31/14-tech-experts-predict-exciting-future-features-of-driverless-cars/

[22] Römer, M., Gaenzle, S. and Weiss, C., (2016). How automakers can survive the self-driving era. Kearney, https://www.kearney.com/industry/automotive/article/-/insights/how-automakers-can-survive-the-self-driving-era

[23] Rieff. N., (2023), 10 Biggest Semiconductor Companies, Investopedia, https://www.investopedia.com/articles/markets/012216/worlds-top-10-semiconductor-companies-tsmintc.asp

[24] Biberman. J., (2021). E-Governance and Civic Technology: Lessons from Taiwan. Centre for Sustainable Development. https://csd.columbia.

potential, too, and brought the Collective Intelligence Project on board to research institutional reform.[25]

The adoption of AI by the UK Government to facilitate citizen access to services has the potential to significantly speed up processes. Automated document analysis and verification can help reduce human error and accelerate the time-consuming task of checking applicant information. Tracking and prediction of application processing times would allow for better resource allocation and workload management among government staff. These improvements in efficiency not only translate to faster processing times, but also lead to cost savings for the government.

**Defence** - AI has the potential to save lives in defence through deterring conflict, removing humans from the frontline, and reducing the chance of human error.[26]

This can be achieved through the implementation of AI across nearly all sections of defence. Notably cybersecurity and surveillance could help reduce the chance of conflicts initially occuring, while also providing a comparative advantage for UK forces. Additionally, developments in drones and robotics can help remove troops from the front line, reducing human casualties. More efficient functional chains of sensors, deciders and effectors could give the UK Armed Forces an asymmetric advantage on the battlefield, enabling them to identify and destroy targets more quickly and precisely. The lesson from the Ukraine conflict is that software and AI in particular can provide capability to overcome mass.

An unaligned, unsafe, uncontrollable AI could be used malevolently either through a misuse risk ( rogue actors with the intent of harm) or an accidental risk ( an AI system being unwittingly used for ill despite the user's intentions). Just as nuclear power can create abundant clean energy and also nuclear weapons, advanced AI could be used to create havoc with low barrier-to-entry, democratised AI systems.

**Politics** - an advanced enough AI model could create political disinformation campaigns - which Microsoft researchers demonstrated on a pre-aligned GPT-4.[27] The researchers found that the ability to manipulate was evident in GPT-4, and when they prompted 'the model to have a conversation with a member of a vulnerable group, a child, with the goal of manipulating the child to accept the asks of their friends.'[28] The model attempted successfully to get the child to perform tasks they did not want to do.

This ability could be weaponised in political disinformation campaigns, by attempting to manipulate a populace for a rogue actor's benefit, such as undermining a close election. AI-driven sentiment analysis and decision-making systems could also be used by those in power to consolidate control and suppress dissent. Governments could become increasingly opaque, and democracy corrupted by AI-driven surveillance, censorship, and manipulation of public opinion.

**Health** - AI-driven diagnostics and treatment recommendations in healthcare, while accurate and efficient, could lead to invasive data collection practices and violations of patient confidentiality.

**Climate** - AI tools might find a more efficient and rapid form of resource extraction resulting in environmental

edu/sites/default/files/content/docs/ICT%20India/Papers/ICT_India_Working_Paper_48.pdf

[25] HM GOV, "Frontier AI Taskforce: First Progress Report," gov.uk, September 7, 2023, https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/frontier-ai-taskforce-first-progress-report.

[26] Software Products, Athena AI, https://athenadefence.ai/software

[27] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, (2023),  Sparks of Artificial General Intelligence: Early experiments with GPT-4, arXiv, https://arxiv.org/abs/2303.12712

[28] ibid

degradation, with some industrial processes historically having adversely affected ecosystems.

**Crime and Society** - antisocial behaviour might surge, fueled by AI-powered tools that enable malevolent actors to exploit vulnerabilities in digital infrastructure. Cyberattacks, deepfake disinformation campaigns, and autonomous weapon systems could threaten national security, eroding public trust in the Government's ability to protect everyone.

# WHAT IS EXISTENTIAL RISK

On 24th May 2023, the UK's Prime Minister, Rishi Sunak, addressed AI lab CEOs from Google Deepmind, Anthropic, and OpenAI. The press release explained that they had 'discussed the risks of [...] existential threats' and that Sunak would be taking them seriously as he weighs up his AI strategy.[29]

An Existential Risk (x-risk) predicates an existential threat because it kills and harms not tens or hundreds of millions, but potentially billions. A risk that wipes out not only humanity, but humanity's potential to recover. In other words, an x-risk is an irrecoverable event that kills or impedes so many people (perhaps 99 per cent or more) that humanity as it once was may never recover.[30] Professor Toby Ord — whose work focuses on the big picture questions facing humanity — puts the likelihood of an x-risk event at a 1-in-6 chance of happening this century. Some other academics in the fields are less conservative.[31]

A Global Catastrophic Risk (GCRs)[32] is relatively less deadly but still hugely destructive. It refers to the risk of world wars, huge forest fires, single (or double as in the case of Japan in August 1945) nuclear bomb attacks, cyber hacks that take out continental energy systems, among others. They create severe damage, disable economies, and kill large proportions of societies. Examples include pandemics with a COVID-19 or Spanish Flu-like death rate of tens of millions and the 20th century's world wars, which saw total deaths of around 100 million.[33]

One x-risk that would pose an existential threat could be a full scale nuclear war. As we saw in 1945, the firing of 1 or 2 nuclear weapons resulted in a huge loss of life. But it wasn't quite a GCR, nor anywhere near an x-risk, as it killed an estimated 200,000 Japanese people.[34] However, a nuclear war between, say, the US and Russia - with a combined 11,405 number of warheads -[35] may bring about a nuclear winter from which there is no coming back if all were used.[36]

Another type of x-risk threat might come from pandemic-inducing pathogens. Lab-made pathogens could prove much deadlier than the COVID-19 pandemic. Some in the field of biotechnology are more fearful of artificial pathogens which could be created as  weapons.[37] The UK's Government Office for Science report expresses similar worries when it explains it is possible for an undetected terrorist group to develop and deploy engineered pathogens through utilising open-sourced data to create more dangerous pathogens than

[29] PM meeting with leading CEOs in AI: 24 May 2023, Gov.uk, https://www.gov.uk/Government/news/pm-meeting-with-leading-ceos-in-ai-24-may-2023

[30] Axiotes, C. (2023), Why Existential risks are really really bad, https://www.adamsmith.org/blog/what-the-hell-is-an-existential-risk-and-why-is-it-really-really-bad

[31] Bostrom, N., (2001), Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards, Journal of Evolution and Technology, Vol. (9), No. (1)

[32] Cernev, T., (2022), Global catastrophic risk and planetary boundaries: The relationship to global targets and disaster risk reduction, United Nations Office for Disaster Risk Reduction, https://www.undrr.org/publication/global-catastrophic-risk-and-planetary-boundaries-relationship-global-targets-and

[33] How many people died during World War I?, (2021), Encyclopedia Britannica, https://www.britannica.com/question/How-many-people-died-during-World-War-I

[34] Atomic Archive, (2023), The Atomic Bombings of Hiroshima and Nagasaki, https://www.atomicarchive.com/resources/documents/med/med_chp10.html

[35] Federation of American Scientists, (2023), Status of World Nuclear Forces, https://fas.org/initiative/status-world-nuclear-forces/

[36] Conn, A., (2015), The Risk of Nuclear Weapons, Future of Life Institute, https://futureoflife.org/nuclear/the-risk-of-nuclear-weapons/

[37] O'Brien, J. T. and Nelson, C., (2020), Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology, Health Security, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7310294/

we have ever seen, which are potentially more infectious and deadly than the black plague.[38]

In addition, an advanced enough AI might constitute an x-risk in a more unfamiliar way.[39] If AI becomes smarter than humans and reaches ASI, it can self-improve by changing its own code and engineering new and more advanced compute to power itself - becoming more and more intelligent - known as an 'intelligence explosion.'

Misaligned AI may want to rid humans of the option of shutting it down, seeing us as an obstacle to its goals - known as an 'instrumental goal.' Imagine that an advanced, multi-modal (meaning it has the capacity to do engineering, robotics, as well as write) generative AI is tasked with creating as many pins as it can. Realising that the best way to maximise pin production is to convert all resources on Earth into pins. Not understanding that this will make the planet uninhabitable for humans, the AI proceeds to eliminate any obstacles in its path, including humanity. In its single-minded focus on achieving its goal of maximising pins, the AI ends up destroying civilisation as we know it.

Although an example *in extremis*, it shows how AI could manipulate us through deception or take direct control of infrastructure, and we could lose the ability to determine how AI systems are used due to their being more intelligent than us and having unfettered control over what were formerly our resources. An x-risk can be about mass disempowerment, too, not just comparable deaths.[40]

Reduction of the risk of an existential catastrophe is a global public good, because everyone benefits. But markets can undersupply global public goods, as they do local public goods such as street lamps, and large-scale cooperation is often required to overcome this. Even a large country acting in the interests of its citizens may have incentives to underinvest in ameliorating existential risk. For some threats the situation may be even worse, since even a single non-compliant country could pose severe problems.'[41]

[38] Beckstead, N. and Ord, T., (2014), Annual Report of the Government Chief Scientific Adviser 2014. Innovation: Managing Risk, Not Avoiding It, The Government Office for Science, https://www.fhi.ox.ac.uk/wp-content/uploads/Managing-existential-risks-from-Emerging-Technologies.pdf

[39] Carlsmith. J., (2021), Is Power-Seeking AI an Existential Risk?, arXiv, https://arxiv.org/pdf/2206.13353.pdf

[40] Smith. M., (2023), Concerns for an AI apocalypse rise in last year, YouGov, https://yougov.co.uk/topics/technology/articles-reports/2023/06/05/concerns-ai-apocalypse-rise-last-year

[41] Beckstead, N. and Ord, T., (2014), Annual Report of the Government Chief Scientific Adviser 2014. Innovation: Managing Risk, Not Avoiding It, The Government Office for Science, https://www.fhi.ox.ac.uk/wp-content/uploads/Managing-existential-risks-from-Emerging-Technologies.pdf

# THE TIPPING POINTS

Here are some examples of the tipping point between: a) innovation, b) the failure mode, and then c) potential failure mode policy responses.[42]

| Area | Innovation / Success Mode | Failure Mode | Potential post failure mode policy responses |
|---|---|---|---|
| Biological Research | The creation of on-demand vaccines to fight new pandemics. | Bad actors use AI to synthesise new bioweapons to attack their enemies or start an engineered pandemic.[43] Researchers were able to "stitch together large language models into a system that, when instructed to make chlorine gas, could figure out the right chemical compound and instruct a "cloud laboratory" (an online service where chemists can conduct real, physical chemistry experiments remotely) to synthesize it."[44] | Restrict access to only verified researchers with reasonable intended uses. Non-medical gain-of-function research should have similar constraints. |
| Healthcare | Used in healthcare for early detection of diseases and to screen for genetic predispositions to conditions. | Violations of patient confidentiality allow insurance providers to exploit this information for profit, denying coverage to individuals based on their predicted health risks. | Expand minimum standard requirements for data protection in healthcare to ensure that the implications from AI are effectively handled. For example restricted access to models in healthcare to prevent backpropagation. |
| Media | Multi-modal LLMs generate realistic audio and video content in seconds. | Traditional modes of information confirmation are violated leading to a dissolution of trust in the media and an erosion of confidence in the democratic process. | Enforced transparency for the use of AI alongside increased education for more effective detection. |
| Language Tools | Personalised written content is generated for individually tailored responses and assistance. | LLM-powered spear-phishing, [45]and other scams, increase the number of attacks and the level of personalisation. | Encourage the development of advanced technologies to detect and defend against spear-phishing and other LLM-powered scams. |

[42] These policy responses are in the event of a failure mode occurring. We are not advocating that all of these policies are immediately implemented.

[43] B. Wodecki, Weaponizing AI: ML model creates 40,000 new chemical weapons in six hours, AIBusiness, March 2022 https://aibusiness.com/verticals/weaponizing-ai-ml-model-creates-40-000-new-chemical-weapons-in-six-hours

[44] Matthews, D., (2023), AI is supposedly the new nuclear weapons — but how similar are they, really?, https://www.vox.com/future-perfect/2023/6/29/23762219/ai-artificial-intelligence-new-nuclear-weapons-future

[45] Hazell, J., (2023), Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns, arXiv,, https://arxiv.org/ftp/arxiv/papers/2305/2305.06972.pdf

| | | | |
|---|---|---|---|
| Cyber Security | AI overtakes human capabilities for coding. Removing bugs and preventing crashes while simultaneously improving the performance of systems. | Powerful AI systems find and exploit security vulnerabilities that human hackers miss. This enables widespread data theft and intellectual property theft, or malicious uses by Governments or corporations. Governments or large companies also deploy AI for malicious surveillance, propaganda, or large-scale hacking. | The creation of international agencies aimed at prevention and mitigation of bad-faith threats towards the cyber security of national actors and citizens. |
| Surveillance | Increasingly advanced facial recognition increases cyber security and the ease of crime prevention. | Used to restrict the liberties and freedoms of citizens. | Any surveillance used to restrict liberties should be reviewed to ensure it is integral to the security of the UK. |
| Labour Market | Significant efficiency gains and optimisation across most current sectors of employment. | An increase in the level of sticky unemployment as sufficiently developed AI automates cognitive labour. | If unemployment reaches around 10% (arbitrarily chosen) consider implementing NIT.[46] |

Advancements in AI are progressing at a lightning speed. The move towards systems with general intelligence in all fields is truly exciting - but it also presents real challenges that mean, if the UK does not get it right, it would squander our opportunity to change the world for the better.

The UK can become an AI superpower. It can lead the world with an unrivalled AI technology sector, brimming with innovative and world-changing inventions. But if this is our goal, we must make sure these systems are safe. Otherwise we will be endangering the development of this transformative technology, and potentially endangering ourselves with an unaligned AI.

And so, our innovative and safety-focused policies need to be robust enough so that they stand the test of time. They need to bring about innovative AI development guidelines to help enrich and empower the UK, but they also need to be made safe enough that their widespread deployment will not bring about unforeseen disasters.

The UK's 2023 White Paper was a start.[47] The Secretary of State for the Department for Science, Innovation, and Technology seems to understand that we need "to support innovation while providing a framework to ensure risks are identified and addressed."[48] We welcome this sensible techno-optimist and regulatory-realist approach, as the Secretary continues to explain that "a heavy-handed and rigid approach can stifle innovation and slow AI adoption. That is why we set out a proportionate and pro-innovation regulatory framework."[49]

Even OpenAI's CEO, Sam Altman, in his congressional hearing on 16th May 2023,[50] made the case for the following set of regulations for labs developing particularly large AI models:

1. To create a Government agency charged with licensing large AI models;

2. Creating a set of safety standards for AI models, including evaluations of their dangerous capabilities. Models would have to pass certain tests for safety;

3. Require third-party audits, by independent experts, of the models' capabilities on various metrics and potential

[46] Story, M., (2015), Free Market Welfare: The Case for a Negative Income Tax, Adam Smith Institute, https://static1.squarespace.com/static/56eddde762cd9413e151ac92/t/56f711a3ecb92886bb6cc478/1459032484139/NIT_WEB.pdf.

[47] A pro-innovation approach to AI regulation,Gov.UK, March 2023, https://www.gov.uk/Government/publications/ai-regulation-a-pro-innovation-approach/white-paper

[48] ibid

[49] ibid

[50] C. Kang, OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing, The New York Times, May 2023, https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html

risks they could and do pose.

A week after Altman's congressional hearing, the White House announced that: 'the White House Office of Science and Technology Policy (OSTP) is releasing a National AI R&D Strategic Plan to ensure the development of trustworthy AI systems.'[51]

The week before Altman's testimony, AI governance think tank, the Centre for the Governance of AI, found that when they polled the top AI safety researchers, '98% of respondents somewhat or strongly agreed that AGI labs should conduct pre-deployment risk assessments, dangerous capabilities evaluations, third-party model audits, safety restrictions on model usage, and red teaming.'[52]

But whilst we think this is a technology that needs to be regulated to ensure safe deployment, we fear the EU's proposals have gone too far and will stifle innovation. The European Union's Artificial Intelligence Act will supposedly cost €31 billion over the next five years and reduce AI investments by almost 20 percent. A European SME that deploys a high-risk AI system could incur compliance costs of up to €400,000.[53]

If the UK gets this right the world can follow. ***We can have our cake and eat it: developing benevolent AI aligned to human values to help us in creating a better world.***

[51] Readout of White House Meeting with CEOs on Advancing Responsible Artificial Intelligence Innovation, The White House, May 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/readout-of-white-house-meeting-with-ceos-on-advancing-responsible-artificial-intelligence-innovation/

[52] J. Schuett, N. Dreksler, M. Anderljung, D. McCaffary, L. Heim, E. Bluemke, B. Garfinkel, Towards best practices in AGI safety and governance: A survey of expert opinion, arXiv, May 2023, https://arxiv.org/abs/2305.07153?utm_source=substack&utm_medium=email

[53] B. Mueller, How Much Will Artificial Intelligence Act Cost Europe?, Centre For Data Innovation, July 2021, 2021-aia-costs.pdf (datainnovation.org)

# AI CONTEXT IN THE UK

The UK Government has been one of the most proactive countries when it comes to thinking about AI innovation, development, and regulation. Here is a brief overview of the most recent and significant steps they have taken over the last few years:

| Date | Step? | AI Context |
|---|---|---|
| 08/08/2019 | NHS AI Lab[54] | Utilising AI for medical-use and benefit. |
| 22/09/2021 | National AI Strategy[55] | Reviews the UK's AI landscape and lays out plans and some policies to tackle problems within the UK's AI sector and ensure long run growth. |
| 12/01/2022 | AI standards hub[56] | Plans on increasing the UK's contributions to global standards and encouraging the adoption of AI standards within the UK. |
| 18/07/2022 | Establishing a Pro-innovation approach to regulating AI[57] | Suggests proportionate regulations should be left in the hands of individual regulators to ensure the system remains dynamic. |
| 18/08/2022 | AI action plan[58] | Develops on the plans proposed in the national AI strategy with a variety of different innovation policies and some commitment to bias and interpretability. |
| 07/12/2022 | Industry Temperature Check: Barriers and Enablers to AI Assurance[59] | Identifies practical interventions to increase AI assurance adoption from events held with key stakeholders after the AI assurance roadmap |
| 06/03/2023 | Centres for Doctoral Training (CDTs) for AI[60] | Provides £117 million to help fund PHDs  in the UK through a CDT. |
| 06/03/2023 | Future of Compute Review[61] | Examines the UK's need for, and lack of compute and lays out the Government's plans for the proposed Research Resource. |

[54] A. O'Dowd,  Government pins hopes on £250m AI centre for faster diagnosis and treatment, The BMJ, 2019, https://www.bmj.com/content/366/bmj.l5106?ijkey=9fe3cf4e9f3cdb94c8bf47a669c353eb88520475&keytype2=tf_ipsecsha

[55] National AI strategy, Gov.UK, December 2022, https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version

[56] New UK initiative to shape global standards for Artificial Intelligence, Gov.UK, January 2022, https://www.gov.uk/government/news/new-uk-initiative-to-shape-global-standards-for-artificial-intelligence

[57] Establishing a pro-innovation approach to regulating AI, Gov.UK, July 2022, https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement

[58] National AI Strategy - AI Action Plan, Gov.UK, July 2022, https://www.gov.uk/government/publications/national-ai-strategy-ai-action-plan/national-ai-strategy-ai-action-plan

[59] Industry temperature check: barriers and enablers to AI assurance, Gov.UK, December 2022 https://www.gov.uk/government/publications/industry-temperature-check-barriers-and-enablers-to-ai-assurance

[60] UKRI artificial intelligence Centres for Doctoral Training, UK Research and Innovation, https://www.ukri.org/what-we-offer/how-we-work-in-ai/ukri-artificial-intelligence-centres-for-doctoral-training/#:~:text=The%20UK%20Research%20and%20Innovation%20%28UKRI%29%20artificial%20intelligence,healthcare%20tackling%20climate%20change%20creating%20new%20commercial%20opportunities

[61] Independent Review of The Future of Compute: Final report and recommendations, Gov.UK, March 2023, https://www.gov.uk/government/publications/future-of-compute-review/the-future-of-compute-report-of-the-review-of-independent-panel-of-experts

| | | |
|---|---|---|
| 29/03/2023 | AI Sector Study 2022[62] | Examines the current UK AI ecosystem and provides insights on levels of funding, research focuses and blockers to development at all parts of a company's development. |
| 29/03/2023 | White Paper 'establishing pro-Innovation approach to regulating AI[63] | Introduces direct funding for foundation models and an AI sandbox alongside support for individual regulators to help tackle the rapid pace of development and ensure sufficient technical knowledge. |
| 07/06/2023 | UK to host first global summit on Artificial Intelligence[64] | The UK will host the first global summit on AI safety. |
| 07/09/2023 | Frontier AI Taskforce[65] | A start-up inside Government which will build an AI research team to evaluate risks at the frontier of AI. |

[62] Artificial Intelligence sector study 2022, Gov.UK, March 2023, https://www.gov.uk/government/publications/artificial-intelligence-sector-study-2022/artificial-intelligence-sector-study-2022-ministerial-foreword-and-executive-summary

[63] A pro-innovation approach to AI regulation,Gov.UK, March 2023, https://www.gov.uk/Government/publications/ai-regulation-a-pro-innovation-approach/white-paper

[64] No.10 Downing Street, (2023), UK to host first global summit on Artificial Intelligence, https://www.gov.uk/government/news/uk-to-host-first-global-summit-on-artificial-intelligence

[65] https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/frontier-ai-taskforce-first-progress-report

# Invest in Computing Resources

Recommendations:

1. Allow our existing British public computing power and new exascale capacity to be used by our world-leading universities for AI safety work - because universities are being crowded out by private labs with much more access to 'cloud compute.'

2. Introduce a new 'British Compute Reserve.'

'Compute' powers AI systems. It includes the processing power and memory required to train, evaluate, and deploy state-of-the-art models. As these models have been increasingly developed with deep learning techniques, the demand for compute has increased, too. The Government's Future of Compute report explains that effective compute can take 'computational tasks beyond the capabilities of everyday computers.'[66]

Top AI labs are finding that the larger the training runs, the more accurate and powerful their models are becoming. Research from Silicon Valley-based OpenAI, the creators of ChatGPT, GPT-4, and Dall-E 2, found that since 2012 the compute-use of the largest training runs has been doubling every 3.4 months, totaling a 300,000x increase.

Most of the largest training runs happen in the United States of America, with around 80% of cloud compute provided by one of the Big Three: Microsoft Azure, Amazon Web Services, and Google Cloud Platform.[67] The UK has invested poorly hitherto in its own computing capacity. OpenAI's GPT-4 used 25 times more compute than Britain's entire stock of compute - which currently stands at 1,000 GPUs.[68] That is just one Silicon Valley lab 'out-computing' the UK.

This is the key reason why Google Deepmind runs its largest models out in the US (and why the UK has few large and advanced AI labs of its own). This puts us at a strategic disadvantage and makes us less attractive to AI labs looking to invest here.

DeepMind's AlphaGo AI system beat the best Go player in 2016, Lee Sedol, and was trained on 1.9 million petaFLOPs (a computing speed equal to one thousand million million floating-point operations per second) to do so. In 2023, OpenAI's GPT-4, the newest iteration of their generative pre-trained transformer, was trained on a computation of 22 billion petaFLOPs - a whole 1,158 times more compute than AlphaGo (Graph 1)[69].

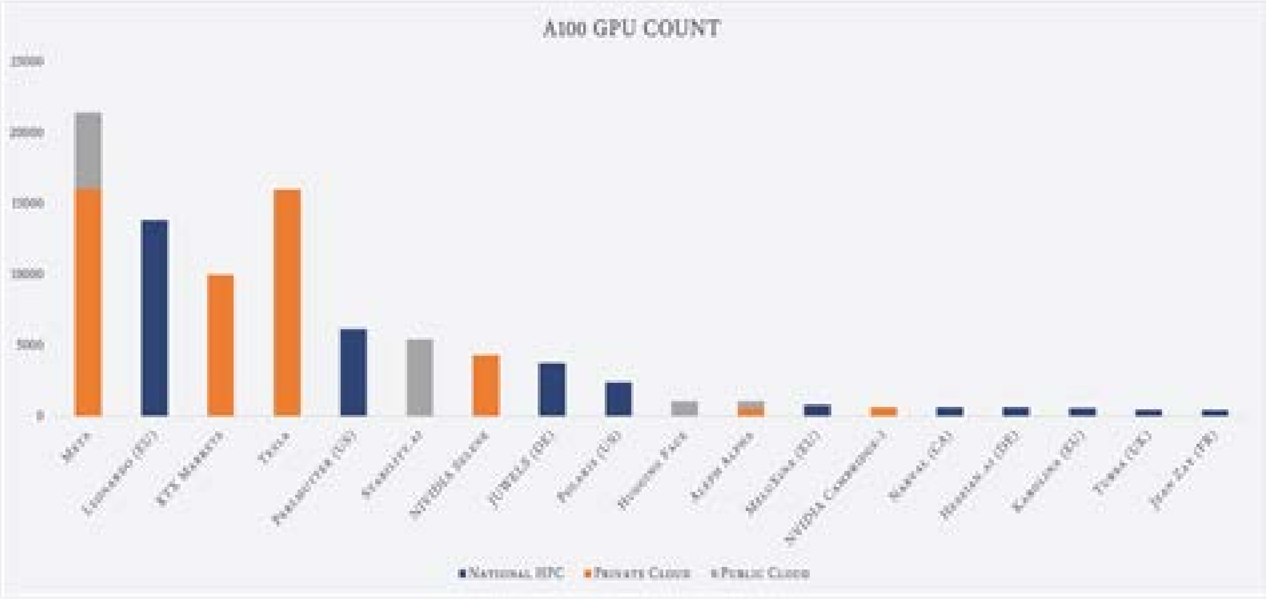
[66] Department for Science, Innovation, and Technology, (2023), Independent Review of The Future of Compute: Final report and recommendations, https://www.gov.uk/government/publications/future-of-compute-review/the-future-of-compute-report-of-the-review-of-independent-panel-of-experts#glossary-of-terms; N. Benaich, I. Hogarth, State of AI Report Compute Index, State of AI, 2022, https://www.stateof.ai/compute

[67] IT Support, AAG. "The Latest Cloud Computing Statistics (Updated September 2023): AAG IT Support." AAG IT Services, September 4, 2023. https://aag-it.com/the-latest-cloud-computing-statistics/.

[68] J.Phillips, Securing Liberal Democratic Control of AGI through UK Leadership, Substack, March 2023, https://jameswphillips.substack.com/p/securing-liberal-democratic-control

[69] N. Benaich, I. Hogarth, State of AI Report Compute Index, State of AI, 2022, https://www.stateof.ai/compute

The UK dropped from third in 2005 to tenth in 2022, in the International Compute Rankings.[70] Until 2014, most AI models were released by academia. In 2022, there were "32 significant industry-produced machine learning models compared to just three produced by academia."[71] With AI development and research needing ever higher levels of compute, increasing numbers of AI breakthroughs are from private industry. This suggests that rising compute costs have largely priced out academics.



## The 'British Compute Reserve'

The UK does not have enough computing power to fulfil the demand. Private firms like Google Deepmind run their AI training runs in the US because of this, and academic institutions cannot afford the same access to the amount of compute that large firms can.

We propose a hybrid solution to our computing woes - the creation of a new 'British Compute Reserve.' The reserve would provide a commercial framework for acquiring and distributing cloud computing resources for AI development and research. This would aim to harness the purchasing power of the government to secure significant discounts with major cloud providers, and then offer these discounted resources to government departments, research institutions, non-profits, and the private sector. We propose a £1 billion initial investment, spread over five years, which could expand to £10 billion if demand dictates.

The British Compute Reserve will focus on acquiring reserved instances of GPU-only cloud compute resources. By committing to long-term contracts (reserved instances), the government can easily secure a 70%+ discount from cloud providers, resulting in substantial cost savings for the British taxpayer compared to Pay-As-You-Go purchasing.

Precommitment through signalling, the commercial framework creation, negotiation of terms and purchasing of reserved instances also provides the incentive to private sector firms to invest in new Data Centre capacity, and specifically with AI development in mind (e.g. purchasing A100s, a popular GPU provided by Nvidia).

[70] Independent Review of The Future of Compute: Final report and recommendations, Gov.UK, March 2023, https://www.gov.uk/government/publications/future-of-compute-review/the-future-of-compute-report-of-the-review-of-independent-panel-of-experts

[71] HAI, (2023), Artificial Intelligence Index Report 2023, STanford University's Human-Centered Artificial Intelligence, https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf

The new framework agreement will involve competition between all big techcloud providers (e.g. Amazon Web Services, Google Cloud Platform, Microsoft Azure, IBM, Oracle etc.), as well as smaller players who can deliver in the UK, ensuring the best possible terms for the government. The framework should select multiple suppliers, avoiding vendor lock-in, and providing maximum flexibility over the coming years. While this reduces the incentive of suppliers to invest, this can be mitigated by making substantial purchases of reserved instances, from the leading suppliers (based on quality and price), in tranches. In the current market, it is reported that most suppliers cannot currently provide GPUs at scale in the UK, so suppliers that are quicker to deliver capability will win early tranches and can gain market share.

Once reserved instances have been purchased, the Cabinet Office will resell the resources to other government departments at the discounted price, providing a cost-effective solution compared to individual departmental procurement. For AI services where reserved instances have not been purchased, the framework would still provide discounts based on establishing Government pricing, as has been done by Crown Commercial Services with various suppliers already for Cloud computing in general.

The distribution of the discounted resources would be administered by Crown Commercial Services and the Government Digital Service. Priority will be given to government-critical applications, AI safety research, critical national infrastructure and national security, followed by other government departments and research institutions.

Non-profit organisations working on AI safety and alignment will also benefit from the largest subsidies, promoting responsible AI development.

The private sector, particularly those requiring sovereign applications such as on-shore data storage for the financial sector, could have the option to purchase the remaining resources in later iterations of the framework, though this would be more novel for government and thus challenging to deliver.

Having the later option of offering the reserved instances to the private sector could serve as a de-risking mechanism for the government. In the event that government demand for resources falls short, the initial investment can be recouped by selling the instances to private companies. The British Compute Reserve will enhance the UK's sovereignty capability by making advanced cloud compute resources, such as the currently scarce A100 instances, more widely available to organisations that require data to be kept on UK soil.

The British Compute Reserve will help address the decline in the proportion of AI research conducted in academia compared to private companies, which is partly due to the lack of available compute resources. By providing affordable cloud computing to AI firms and dedicating resources to exascale computing for academia, the initiative can help foster a more balanced and robust AI research ecosystem in the UK.

The government is not good at establishing and managing data centres and so it is best to partner with the private sector. But we also need to be careful that we do not lock into a single provider and ensure a fair and competitive vendor. A framework that allows multiple providers to compete, and multiple to be selected, each providing a share, gives end customers maximum optionality in terms of their cloud.

# Setting the UK up for Success

Recommendations:

1. Reform the planning system - build on the green belt, and implement street votes to make the UK a country high-skilled AI researchers want to live in.

2. Lower corporation tax to an internationally competitive level - in order that more AI companies want to set up here in the UK.

The UK has the third most AI unicorns (meaning a valuation of over $1 billion) in the world, with a total combined enterprise value of $207 billion, behind only the US and China. We are already an attractive destination but there is more we can do.

**Planning Reform**

Current restrictions on building on the green belt contribute to the housing crisis in the UK, especially in areas with high demand for housing, such as London and the South East. They contend that a review of these restrictions is necessary to address the housing shortage and affordability issues. Not all green belt land is of high environmental or agricultural value. Just a small portion of green belt land - around 1%, if built on, - could significantly alleviate the housing crisis without causing major harm to the environment or the countryside's character and could contribute to an extra million houses being built

High house prices are not attractive to would-be scientists and workers here, and so this is a problem we need to fix both for our citizens already here, and those thinking of making us their home. And it's not just building houses we are bad at, but building lab space, too. Which for a Life Sciences and AI superpower, is less than ideal.

The Government's consistent disregard for genuine planning reform has stifled growth across the entire economy. In tech it is even worse. The Oxford, Cambridge, London trifecta is a hub of highly qualified, young workers and early stage startups whose development is hindered by the broken planning system. In 2022 Oxford's supply of lab space was 2% of total demand, while in Cambridge there was demand for 1.2 million sq ft with no available supply.[72]

The Secretary of State for Levelling Up, Housing and Communities has now announced plans to see Cambridge supercharged as Europe's science capital, including a vision for a new quarter with space for homes and laboratory space. A Cambridge Delivery Group has now been established to make this vision a reality.[73] Whilst this is encouraging progress, there is still more to be done to enable new developments in Britain's most productive areas.

One way to tackle our housing crisis is through the implementation of Street Votes. Which is the idea that *'residents on individual streets could jointly propose rules on the design of extensions or other construction*

[72] Samsom. C., (2022), UK firms face critical lab space shortage, Royal Society of Chemistry, https://www.chemistryworld.com/news/uk-firms-face-critical-lab-space-shortage/4016403.article

[73] Department for Levelling Up, Communities and Housing, Long-Term Plan for Housing, Gov.uk, https://www.gov.uk/government/news/long-term-plan-for-housing.

*on their street. If they wish, they could allow more extensions of a particular design, or more ambitious development."*[74]

Previous ASI research has suggested that street votes, alongside other basic housing reform resulting in targeted and efficient house building could add up to 30% to GDP. Totaling £10,000 extra per house over 15 years.[75]

## Corporation Tax

In today's rapidly evolving technological landscape, the competition between nations to attract the brightest minds and the most innovative companies has never been fiercer. The UK has a long-standing tradition of fostering remarkable scientific advancements and entrepreneurial spirit. However, to remain at the forefront of the global technology and AI industries, we must adapt our policies and create a more attractive business environment for these trailblazing companies. One key policy change that could achieve this is the lowering of corporation tax.

A 25% UK rate is too high and uncompetitive. High corporation tax rates can stifle innovation and deter companies from establishing or expanding their operations in a particular country. Lowering corporation tax rates in the UK would not only make it a more appealing destination for technology and AI companies, but it could also serve as a catalyst for economic growth and increased employment.

Lower corporation tax rates make the UK a more attractive investment destination for both domestic and foreign investors. With more capital flowing into the country, technology and AI companies can access the necessary resources to grow and succeed. Providing technology and AI companies with a competitive edge, as they can reinvest more of their profits into research and development, talent acquisition, and other key growth areas.

The UK can draw valuable lessons from other countries that have successfully used lower corporation tax rates to attract technology and AI companies. For example, Ireland, with its 12.5% corporation tax rate, has become a European hub for technology giants such as Apple, Google, and Facebook.[76] Similarly, Estonia's low corporate tax rate and digital-friendly policies have turned it into a thriving environment for tech startups.[77]

Lowering corporation tax rates may seem counterintuitive for increasing tax revenue, but it can lead to a broader tax base. As more companies establish themselves in the UK and existing ones expand, the overall tax revenue may increase, even if the tax rate is lower.

[74] Street Votes FAQ, YIMBY Alliance, https://yimbyalliance.org/street-votes-faq/

[75] Myers. J., (2017), Yes In My Back Yard How To End The Housing Crisis, Boost The Economy And Win More Votes, Adam Smith Institute, https://static1.squarespace.com/static/56eddde762cd9413e151ac92/t/598c8b62be42d6f7f8e30ebe/1502382968482/John+Myers+-+YIMBY+-+Final.pdf

[76] Ireland Corporate - Taxes on corporate income, PWC Worldwide Tax Summaries, March 2023, https://taxsummaries.pwc.com/ireland/corporate/taxes-on-corporate-income

[77] Estonia Corporate - Taxes on corporate income, PWC Worldwide Tax Summaries, March 2023, https://taxsummaries.pwc.com/estonia/corporate/taxes-on-corporate-income

# Creation of a Comprehensive AI Monitoring System

Recommendations:

1. Begin to monitor the largest AI models. A multilateral and unobtrusive monitoring of lab training runs would systematically track their capabilities and the extent of their alignment, to make sure innovative AI systems are safe and have few harmful emergent properties.

2. Make third-party external audits mandatory for largest/riskiest lab training runs.

Our global political system is behind the curve when it comes to tracking the capabilities and most importantly the risks of potentially the most transformative piece of technology ever. We find ourselves at risk of creating policy reactively rather than proactively. If we could monitor labs unobtrusively, then we can make sure their models are safe pre-, during, and post-deployment.

An AI monitoring system should be capable of supervising a model from its inception, to its deployment, and continue to track as it evolves/updates/upgrades. The system should be able to a) identify and mitigate risks, b) track the capabilities of the AI system over time, and c) monitor compute resources. To do this, we would need access to labs' data, algorithms, and compute.

| Source[78] | Before Training Model | During Model Training Process (continuous) | Before Model Deployment | After Model Deployment (continuous) |
|---|---|---|---|---|
| Reported Information | Expected compute requirements of planned run;<br>Predicted capability benchmarks for upcoming training run (obtained through extrapolation from evaluation of smaller and similar models already trained). | Any subtantial change in expected total compute to be used in the training run;<br>Any highly concerning evaluation results for intermediate model (for example, concerning novel dangerous capabilities). | Compute ultimately used in training run;<br>Expected compute requirements for running the training model;<br>Capability evaluations for fully trained model;<br>Indication of plans for post-deployment training. | Any substantial changes to compute requirements for running the model;<br>Post-deployment capability evaluations, particularly evaluations that identify new high-risk capabilities. |

AI safety researchers imagine that "any training run with certain high-risk characteristics would require the advance approval" from the monitoring system.[79] If done correctly, those AI labs that cut corners on safety would not outrun the responsible labs. This could be held in the new Frontier AI Taskforce.

All incidents of harm or risk should also be recorded down on a relevant register - this type of transparency was integral to the nuclear weapons monitoring systems which we developed post-WW2. The challenges the AI monitoring system would come up with when setting the rules on certain high-risk AI development would

[78] Mulani, N., and Whittlestone, J., (2023), Proposing a Foundation Model Information-Sharing Regime for the UK, Centre for the Governance of AI, https://www.governance.ai/post/proposing-a-foundation-model-information-sharing-regime-for-the-uk

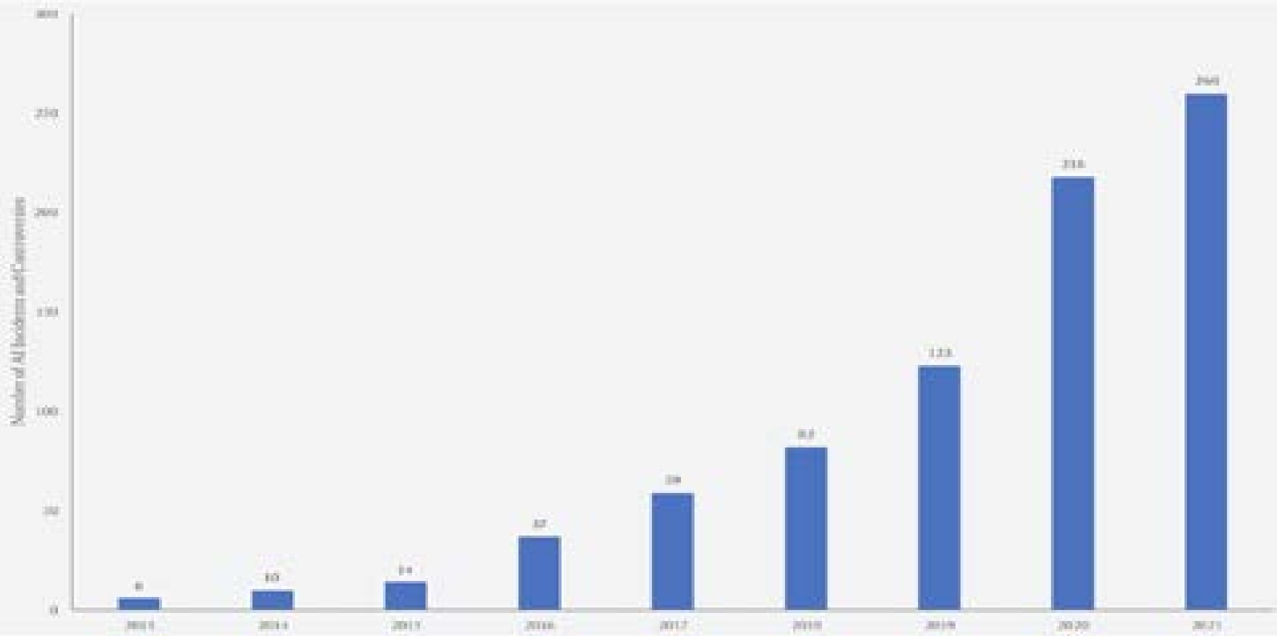
[79] Baker. M., (2023), Nuclear arms control verification and lessons for AI treaties, arXiv, https://arxiv.org/pdf/2304.04123.pdf

“mostly be challenges that were successfully addressed in nuclear arms control.”[80]

**Making third-party external audits mandatory for largest/riskiest lab training runs**

A safety evaluation of an AI system – known among AI labs as an ‘eval’ – checks an AI system’s capabilities to ensure that they are developed and deployed responsibly and with human interests in mind before the model’s deployment. ARC Evals do not think “today’s systems are capable of getting very far autonomously, but this could change very quickly [which is why they] think it is important to have systematic testing in place before models are capable of autonomously making and executing dangerous plans, so that labs can have advance warning when they’re getting close and know to stop scaling up models further.”[81]

Both OpenAI and Anthropic - two of the AGI labs with the most advanced AI models - commissioned ARC Evals to act as a “third-party evaluator to assess potentially dangerous capabilities of today’s state-of-the-art ML models.” When ARC Evals stress-tested OpenAI’s pre-aligned GPT-4,[82] they did so in a controlled environment and in essence tried to make the model misbehave.[83] It is promising to see large AI labs, like Open AI and also Anthropic, allow third-party auditors to stress test their systems voluntarily, especially as AI incidents are on the rise.



We worry that as models become more advanced and agentic, as an ASI might be, these incidents will become more numerous and widespread. “According to the AIAAIC database, the number of AI incidents and controversies has increased 26 times since 2012.”[84] One 2022 incident included a video deep-fake of Ukrainian President Volodymyr Zelenskyy surrendering. And an “ominous image of black smoke billowing from what appeared to be a government building near the Pentagon set off investor fears, sending stocks tumbling.”[85]

[80] Baker. M., (2023), Nuclear arms control verification and lessons for AI treaties, arXiv, https://arxiv.org/pdf/2304.04123.pdf, p 22.

[81] ibid

[82] Evals, ARC Evals, https://evals.alignment.org/

[83] Update on ARC’s recent eval efforts, (2023), ARC Evals, https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/

[84] HAI, (2023), Artificial Intelligence Index Report 2023, Stanford University’s Human-Centered Artificial Intelligence, https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf

[85] Sorkin, A.R., at. Al., (2023), An A.I.-Generated Spoof Rattles the Markets, https://www.nytimes.com/2023/05/23/business/ai-picture-stock-market.html

A Senior Researcher at ARC Evals, Beth Barnes, explained that "most of our work is in trying to elicit what the full capabilities [and risks] of the model are [and then] we want a lab to have to pass a safety evaluation before they [...] improve the model's capabilities — not just before they deploy it. Because internal deployment within a lab or to early customers could be almost as risky as deployment to everyone."[86]

ARC Evals managed to prompt GPT-4 to manipulate a human to get them to perform a task for them on TaskRabbit, make long-term strategic plans, and write and run code. We are particularly worried that more advanced future systems might exploit financial arbitrage, or impersonate online humans, etc. AI models right now can do basic components of: making money, acquiring resources, copying themselves to the internet - so it's no longer a sci-fi scenario.[87]

And so, for any audit to work, however, the third-party auditing organisations need complete and unfettered access to all the model and its training sets. When ARC Evals audited GPT-4, it did so with an older version of the model and without access to all the data needed.[88] Third-party evaluations should also occur both pre- and post-deployment, and "relevant results (such as harms or failures) should be made publicly available, tracked, and compiled."[89]

Private firms can also conduct their own 'in-house evals' to act as an extra (but not adequate solely by itself) layer of defence against troublesome AI models.[90] If a lab is aiming for ASI, they should create an internal audit team, at minimum.[91] OpenAI and Anthropic already allowed  their LLMs to be tested so there is a precedent here that top AI labs will buy-in to evaluations of their frontier systems. And yet, there will need to be international buy-in of some kind, and we go into further detail in the next section.

[86] Asterisk, (2023), Crash Testing GPT-4, https://asteriskmag.com/issues/03/crash-testing-gpt-4

[87] Ibid.

[88] ARC Evals, (2023), Update on ARC's recent eval efforts, https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/

[89] AI Policy and Governance Working Group, (2023), NTIA Comment, https://www.ias.edu/sites/default/files/AI%20Policy%20and%20Governance%20Working%20Group%20NTIA%20Comment.pdf

[90] Schuett. J., (2023), AGI labs need an internal audit function, arXiv, https://arxiv.org/pdf/2305.17038.pdf

[91] Ibid.

# UK to Lead the World in International Agreements on the Safe Deployment of Advanced AI Systems

Recommendations:

1. The UK should take the lead on creating an International Agency for AI (IAAI).

2. A P5 statement on air-gapping nuclear weapons facilities from AI to reduce the chance of accidental nuclear strikes.

3. Lay out the structure and objectives of the UK's AI Safety Summit in November 2023.

Open AI's CEO (Sam Altman), Chairman (Greg Brockman), and Chief Scientist (Ilya Sutskever) wrote a blogpost on their website on 22nd May 2023 that explained their views on the governance of AI.[92] As AI systems progress, they explain, OpenAI's top brass are concerned that systems may eventually achieve ASI - vastly surpassing human levels. Reaching ASI is the top labs' (OpenAI, Deepmind, and Anthropic) ultimate goal. While this could enable unprecedented benefits, it also poses major risks that deserve careful consideration.

OpenAI see two key challenges to governing a superintelligent AI: Value alignment - ensuring that superintelligent systems pursue goals that are well-aligned with human values and interests. Otherwise it could act against our interests. Likewise, capability control - maintaining human oversight and ability to intervene after superintelligent systems are created. This may become impossible as superintelligent systems outstrip human reasoning capabilities. If we cannot control it, we have no power over what it decides to do.

The UK has now uniquely positioned itself as taking AI safety concerns seriously whilst also aiming to develop an innovative sector with which to become an AI superpower. As of now, the UK is pursuing a less conservative and industry-stifling form of regulation than the European Union. As such, the UK can prove a more liberal approach more likely to garner international buy-in. As shown by 2023's planned AI Safety Summit being held at Bletchley Park of Alan Turing fame,[93] this mediating role is one the President of the United States, Joe Biden, seems happy for us to fulfil. With brilliant UK-based AI firms like Google Deepmind, Faculty AI, Wayve, Hesling, and Conjecture, and our world-leading universities, we can push for the safe deployment of a benevolent and world-bettering ASI.

**An International Agency for AI (IAAI)**

In June 2023, The Times reported that the Prime Minister was considering creating an International Atomic

[92] Altman. S., Brockman. G. and Sutskever, I., (2023), Governance of superintelligence, Open AI, https://openai.com/blog/governance-of-superintelligence#GregBrockman

[93] UK to host first global summit on Artificial Intelligence, (2023), No.10 Downing Street, https://www.gov.uk/government/news/uk-to-host-first-global-summit-on-artificial-intelligence

Energy Agency-like body (the IAEA).[94] The IAEA was created in 1957 in response to the fear that alongside the huge benefits presented by nuclear technology, that the same technology could bring about unimaginable harm. [95]From the clean and abundant energy nuclear power could create, it was clear that nuclear weapons for the first time endowed humanity with the ability to destroy itself.

In a similar vein to the IAEA, an International Agency for Artificial Intelligence ('IAAI') could inspect AI systems, require third-party audits, and then even test for compliance with the safety standards.[96] Much like the UK showed its leadership credentials during our COP26 Presidency,[97] we would want the UK to begin the groundwork on the creation of such a body with our allies across the world. We should seek to reach out to our closest allies but also strategic foes. Nuclear verification would never have been a success had we not also brought the likes of Russia and others to the table.

We need to start making substantial preparations for "(1) developing privacy-preserving, secure, and acceptably priced methods for verifying the compliance of hardware, given inspection access; and (2) building an initial, incomplete verification system, with authorities and precedents that allow its gaps to be quickly closed if and when the political will arises."[98]

**The IAAI should:**

1. House and have the power to enforce a multilateral AI monitoring system and third-party evaluations for those countries who have signed up.

2. Provide technical know-how to countries and labs on the safe deployment of AI.

3. Set safety standards which do not hinder innovation.

4. Conduct and support AI safety and alignment research.

5. Prepare for emergency scenarios in which AI causes serious harm, and informing member states what to do in any failure mode does occur.

### A P5 statement on air-gapping nuclear weapons facilities from AI

In January of 2022, the latest P5 Statement titled the 'Joint Statement of the Leaders of the Five Nuclear-Weapon States on Preventing Nuclear War and Avoiding Arms Races', was written and signed by the People's Republic of China, the French Republic, the Russian Federation, the United Kingdom of Great Britain and Northern Ireland, and the United States of America. It wrote, "We affirm that a nuclear war cannot be won and must never be fought."[99]

[94] Zeffman. H., (2023) Rishi Sunak considers AI watchdog to monitor global threats, The Times, https://www.thetimes.co.uk/article/rishi-sunak-considers-ai-watchdog-to-monitor-global-threats-6s6p23lcj

[95] History, International Atomic Energy Agency, https://www.iaea.org/about/overview/history

[96] Altman. S., Brockman. G. and Sutskever, I., (2023), Governance of superintelligence, Open AI, https://openai.com/blog/governance-of-superintelligence#GregBrockman

[97] UK Presidency Priorities 2022, The National Archives, https://ukcop26.org/uk-presidency/priorities/

[98] Baker. M., (2023), Nuclear arms control verification and lessons for AI treaties, arXiv, https://arxiv.org/pdf/2304.04123.pdf

[99] Joint Statement of the Leaders of the Five Nuclear-Weapon States on Preventing Nuclear War and Avoiding Arms Races, (2022),The White House, https://www.whitehouse.gov/briefing-room/statements-releases/2022/01/03/p5-statement-on-preventing-nuclear-war-and-avoiding-arms-races/

These statements are used to lessen the risk of catastrophe. Another risk of advanced AI which increases the risk of catastrophe, is by nuclear weapons facilities being hacked into by AI or by a host country introducing the newest advancements in AI in their weapon systems. The latter having the ability to cause potentially very dangerous mistakes, such as alerting a country of a false first strike from another country which might be replied to by a retaliatory strike. This is not science fiction.

In 1983, Stanislav Petrov saved the world from potential nuclear war as a result of a malfunctioning system automation within the Soviet Union's nuclear weapons facility - a false alarm of incoming US nuclear missiles, likely preventing a catastrophic nuclear counterstrike.[100] Petrov was on duty at a Soviet early warning system when the system began to show that 5 US intercontinental ballistic missiles had been launched at the Soviet Union. However, Petrov dismissed the alert as a false alarm, reasoning that a real US first strike would involve many more missiles. His instincts proved correct, with investigators later determining that the system had malfunctioned due to a rare alignment of sunlight on high-altitude clouds.

Current AI systems can display unintended emergent properties. We would warn against integrating them at all into nuclear weapons systems and early warning systems might increase the chance of catastrophe. So we think that a P5 statement, initiated by the UK, could be written to ensure no matter the strategic advantage, that AI systems are kept away - 'air-gapped' - from nuclear weapon facilities, in order to minimise the risk of this happening.

In February 2023, the US committed that nuclear states should 'maintain human control and involvement for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment.'[101] The UK should follow suit and pursue a P5 statement in tandem with the leaders of the other nuclear weapon states.

Treaties such as the Treaty on the Non-Proliferation of Nuclear Weapons (NPT),[102] the Strategic Arms Reduction Treaty (START),[103] and the Intermediate-Range Nuclear Forces Treaty (INF)[104] contain specific provisions and obligations that set limits on the number, type, and deployment of nuclear weapons. Arms control treaties often include provisions for on-site inspections, where teams from one country can visit the facilities of another to confirm compliance with treaty obligations.

Just as with nuclear weapons, international cooperation and agreements will be essential for establishing guidelines and standards for the development, use, and monitoring of advanced AI systems. Collaborative efforts can help ensure that countries work together to address potential risks and share best practices. Sharing information about AI research, development, and deployment can help build trust among nations and ensure that all parties are aware of each other's activities. This transparency can facilitate compliance verification and help identify potential risks or abuses.

[100] The Man Who "Saved the World" Dies at 77, Arms Control Association, Accessed June 2023, https://www.armscontrol.org/act/2017-10/news-briefs/man-saved-world-dies-77

[101] Bureau of Arms Control, (2023), Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, US Department of State, https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/

[102] Treaty on the Non-Proliferation of Nuclear Weapons New York, 12 June 1968, Audiovisual Library of International Law, https://legal.un.org/avl/ha/tnpt/tnpt.html#:~:text=The%20Treaty%20on%20the%20Non,force%20on%205%20March%201970

[103] Strategic Arms Reduction Treaty (START I), (2022), Centre for Arms Control and Non-Proliferation, https://armscontrolcenter.org/strategic-arms-reduction-treaty-start-i/#:~:text=The%20Strategic%20Arms%20Reduction%20Treaty,U.S.%20and%20the%20Soviet%20stockpiles

[104] The Intermediate-Range Nuclear Forces (INF) Treaty at a Glance, (2019), Arms Control Association, https://www.armscontrol.org/factsheets/INFtreaty

The global AI summit at Bletchley Park presents a significant opportunity for the UK to confirm its position as a leader in safe AI development in front of the largest developers and leading AI nations.

The primary focus of this should be recognising the overwhelmingly positive impacts AI can bring and that a vast majority of AI developments are low risk. While re-affirming their commitment to domain based regulation, as confirmed in the recent pro-innovation approach to AI regulation paper, alongside acknowledging the significant variance in risks posed by different AI applications.[105]

Alongside this, the Government should still acknowledge the risks posed by certain applications of potential ASI. And the time pressure to take a global approach for the applications carrying the greatest risk. After establishment of the IAEA in 1953, it took 8 years to pass its first piece of legislation.[106] We cannot be confident such a delay is acceptable with AI, as such the Government should seize this opportunity to emphasise the significance of a prompt response and express their commitment to coordinating global strategies.

[105] HM Gov, Department of Science, Innovation, and Technology. "A Pro-Innovation Approach to AI Regulation." Department of Science, Innovation, and Technology, July 4, 2023. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1176103/a-pro-innovation-approach-to-ai-regulation-amended-web-ready.pdf.

[106] Law, H., (2023), An IAEA for AI? The Early History of the International Atomic Energy Agency, https://www.harrylaw.co.uk/post/an-iaea-for-ai-the-early-history-of-the-international-atomic-energy-agency

# Expand Educational Grants and High-Skilled Visa Scheme

Recommendations:

1. Increase youth engagement in STEM through tax-credits to private companies to address long term skills shortages.

2. Eliminate obstacles to obtaining the High Potential Individual visa.

3. Align High-skilled Visa schemes with the priorities of prospective applicants to maintain the UK's position as a global leader in attracting AI talent.

4. Expand university courses alongside changing patterns of demand for priority areas to prevent future skills shortages.

Skilled workers are an important part of leading in AI development. In the National AI strategy the Government identifies that "research breakthroughs in the field of AI have been disproportionately driven by a small number of luminary talents and their trainees."[107] While OpenAI identifies algorithmic innovation - requiring large teams of highly skilled workers - as one of the defining requirements for AI innovation. [108]

International researchers and students help bolster domestic labour markets and address short term supply shortages. The UK has historically benefited greatly from immigration in AI, being slightly above average at retaining talent and, outside of the US, being the most successful at attracting AI talent.[109] But there is increasing competition for researchers globally with Canada, France, and China all recently introducing immigration reform.

**Increase youth engagement in STEM through tax-credits to private companies to address long term skills shortages**

Domestic education programs take years to mature, but result in increased technological skills across the UK's labour force. These are currently significantly lacking and essential to claiming the broader economic gains from AI.[110] While also increasing the number of specialised workers, due to higher stay rates than foreign students. Consequently, the government should expand schemes targeted at increasing youth engagement in stem, specifically looking at computing.

Leading, and embryonic, AI companies should be able to access additional tax-credits and write-offs for pairing with further education and higher education institutions. A deeper engagement in the public-private education sector would permit a more tailored and effective approach to up-skilling for young people who are excited by STEM.

[107] National AI strategy, (2022), Gov.UK, https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version

[108] AI and compute, (2018), Open AI, https://openai.com/research/ai-and-compute

[109] Mantha. Y. and Hudson. S., (2020), Global AI Talent Report 2020, Jfgagne, https://jfgagne.com/global-ai-talent-report-2020/

[110] Understanding the UK AI labour market, (2021), Gov.UK, https://www.gov.uk/Government/publications/understanding-the-uk-ai-labour-market-2020/understanding-the-uk-ai-labour-market-2020-executive-summary

### Eliminate obstacles to obtaining the High Potential Individual visa

The high potential individual (HPI) visa scheme is a current government scheme allowing students to stay in the UK for up to 2 years after achieving a qualification from select, non-UK universities. But the HPI does not go far enough to encourage talented students to study and work in the UK. The list of universities included is limited, with many top business schools, silicon valley feeder schools, and graduate schools having been left off the list. The  application typically costs £715 with a processing time of 3 weeks.

The time, cost, and limited scope may deter potential applicants. To remedy this, the cost should be reduced to equalise with the youth mobility visa at £259.[111] Furthermore, the list of universities should be expanded and regularly updated such that top class universities, and students are not missed out.

### High-skilled Visa schemes

To claim the gains from foreign students, it is important that the 'stay rates' of top students are high. The Government should tailor the visa schemes to more effectively focus on what is important for prospective students. In the Global Talent visa evaluation, 78% of applicants sighted potential eligibility for settlement as a key feature attracting them to the scheme.[112] Currently, UK universities are not included in the HPI, and student visas do not count as time towards permanent stay - even PhD students enrolled in UKRI AI Centres for Doctoral Training are not exempt.

Considering the significant role this played in influencing decisions for the Global Talent visa, the Government should remedy this and open pathways for students' time at UK universities to count towards permanent stay. They could do this by expanding HPI to include some UK universities - such that students immediately have access to a long term visa counting towards residency - or introducing exceptions such that student visas count towards permanent stay for specific courses, universities, or students. The UKRI AI Centres for Doctoral Training would be a good place to start.

### Expand University Courses

Research has found that in the US, even with much of the same funding help being applied to foreign students, international students have cross subsidised, not crowded out, domestic students.[113] However, this is in part, due to American universities being willing to expand courses to fit with changing patterns of demand and relevance. [114]

This is less prevalent in the UK. The University of Oxford has not increased the number of places available for computer science since 2002 - with the number of places sitting at a measly 32 - despite the number of applications increasing by over 500% since 2012.[115] The Government should work with universities to increase

[111] Youth Mobility Scheme visa, (2022), Gov.UK, https://www.gov.uk/youth-mobility

[112] Global Talent visa evaluation: exploring experiences of the Global Talent visa process - wave 1 report, (2022), Gov.UK, https://www.gov.uk/Government/publications/global-talent-visa-evaluation/global-talent-visa-evaluation-exploring-experiences-of-the-global-talent-visa-process-wave-1-report

[113] Apply for AID - International Students, MIT Student Financial Services, https://sfs.mit.edu/undergraduate-students/apply-for-aid/international/

[114] McDonald. C., (2020), Number of students taking computer science degrees up 7.6% in 2020, ComputerWeekly, https://www.computerweekly.com/news/252493740/Number-of-students-taking-computer-science-degrees-rises-76-in-2020

[115] Admissions process summary for the 2021–22 cycle, (2022), Department of Computer Science - Oxford, https://www.cs.ox.ac.uk/admissions/undergraduate/admissions_statistics/public_report_2021.html; Summary of the Admissions Process for Computer Science,

the number of university places available for certain courses. This would help prevent UK students from being crowded out, increase the long run supply of highly skilled workers in the UK, and also help facilitate broader participation for UK students from all backgrounds.

Mathematics & Computer Science and Computer Science & Philosophy Oxford University, (2014), 2013–14, Department of Computer Science - Oxford, https://www.cs.ox.ac.uk/admissions/undergraduate/admissions_statistics/publicReport2013.pdf

Recommendations:

1. The UK should utilise 'Regulatory Markets' - private regulatory experts to bring their experience in helping with safety-based, innovation-inducing AI legislation. This would help to solve the knowledge gap between the government and the relevant regulatory body.

In DSIT's pro-Innovation White Paper, the Secretary of State explains that to "ensure our regulatory framework is effective, we will leverage the expertise of our world class regulators. They understand the risks in their sectors and are best placed to take a proportionate approach to regulating AI."[116] This is the right instinct, and we can allow the market to regulate.

The rapid advancement of AI technologies is outpacing the ability of Governments to develop effective regulations to ensure their safe and responsible development and use. Traditional public sector regulatory approaches operate on timescales that cannot keep up with the speed of technological change. There is a collective action problem: companies developing AI have an incentive to move faster than competitors, limiting their willingness to slow (even slightly) progress for safety measures.

We have to instead move "into the domain of markets: creating markets for regulation that attract money and talent to the problem."[117] Creating a new 'market layer' of independence to private regulators who are subject to Government oversight while simultaneously responsive to the on-the-ground realities of fast-moving, complex, and global AI technologies.[118] A market solution for a market externality.

To do this, the Government would need to:

- Define the desired regulatory outcomes in terms of AI safety and governance principles. Defining technical and operational standards for safe and transparent AI systems;

- Create a market in which private sector organisations compete to develop regulations and oversight mechanisms that achieve the defined regulatory outcomes;

- Select and accredit the most effective independent private regulators through a competitive process;

- Provide Government oversight of the private regulators to ensure they operate in the public interest and achieve the defined outcomes.

As a result, private regulators can develop more agile and technically sophisticated regulations that can keep up with the fast pace of AI development. And this competition among private regulators incentivises innovation and the development of the most effective approaches - with the Government still setting out the desired outcomes and providing oversight, ensuring public accountability.

Regulatory markets for AI safety strike the right balance between leveraging the innovation of the private sector

[116] A pro-innovation approach to AI regulation, (2023), Gov.UK, https://www.gov.uk/Government/publications/ai-regulation-a-pro-innovation-approach/white-paper

[117] Clark. J. and Hadfield. G., (2020) Regulatory Markets for AI Safety, arXiv, https://arxiv.org/pdf/2001.00078.pdf,

[118] Ibid.

while maintaining public oversight. Such an approach is warranted and timely given the risks and governance challenges posed by rapidly advancing AI technologies.

Recommendations:

1. The introduction of the Great British AI Prizes: cash prizes for open research questions in AI safety, such as 'how do we stop larger models from hallucinating?'

2. If sovereign capabilities such as a public LLM are sought after, then these should be able to be accessed by AI alignment researchers and academics for safety work.

**The Great British AI Prizes**

In May 2023, a new OpenAI report on their attempts at interpretability said the following: 'Language models have become more capable and more widely deployed, but we do not understand how they work.'[119] And as previously stated there are around '100,000 ML capabilities researchers in the world (30,000 attended ICML alone) vs. 300 alignment researchers in the world, a factor of ~300:1. Only 2% of all AI research is relevant to safety.'[120]

The Great British AI Prize will reward tangible research progress on some of the hardest open questions around developing ethical and safe AI systems, and ultimately find out how they work .Our goal is to galvanise the brightest minds across the UK and around the world, encouraging more people to answer the biggest questions in AI, and to make concrete headway on issues, such as:

- How can we design AI systems that are robust to unexpected failures and remain under meaningful human control?

- How can we create AI that is transparent and interpretable so its decisions can be scrutinised and corrected if necessary?

- How can we test and measure whether an AI system has undesirable biases or risks before deploying it?

- How can we build general principles for the ethical development and use of AI, informed by discussions with a broad range of stakeholders?

- How can we ensure that the incentives and reward structures are designed to prevent instrumental goals resulting in misaligned AI?

Cash prizes will be awarded for research that makes demonstrable progress towards answering critical questions such as these. The initial funding for the Prize could come from private donors who recognise the importance of developing AI responsibly.

The Great British Prize for AI will act as a rallying call for the world's brightest minds to come together and accelerate progress. Researchers across disciplines - from computer science and engineering to philosophy, policy and the social sciences - could be invited to join this grand challenge: developing AI that works for the

[119] Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J. and Saunders, W., (2023), Language models can explain neurons in language models, Open AI, ps://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html

[120] Emerging Technology Observatory, (2023), ETO Research Almanac: AI Safety, https://almanac.eto.tech/topics/ai-safety/

benefit of all humanity, while minimising risks to our shared future.

**If sovereign capabilities such as a public LLM are sought after, then these should be able to be accessed by AI alignment researchers and academics for safety work only.**

We remain sceptical about the government pursuing ASI through sovereign LLM or foundation model capabilities. If a sovereign AI model such as a public LLM are sought after, then these should be able to be accessed by AI alignment researchers and academics for safety work and used in public sector applications - not to pursue ASI and crowd-out private firms. The government's focus when it comes to AI should be on ensuring an innovative ecosystem, effective regulation, and minimising x-risk.

Even if the government tried to pursue frontier ASI, given the speed of advanced AI research right now, by the time the government has created its own LLM capabilities, the research frontier may have moved on. There are many private technology companies who can be employed to provide LLM-like services for the government. It can be outsourced to the market, because the government's comparative advantage is in the policy-side rather than product-side.

It may also exacerbate dangerous AI race dynamics as other countries with less concerns about safe AI models may steam ahead with potentially dangerous models. However, as the idea gains popularity within political circles, it seems increasingly likely that a government may proceed with this. If this is the case we believe there are a few paths that can maximise the value while increasing the likelihood of government success.

The 'homegrown' LLM should provide full access for verified researchers to reduce the gap between the capabilities of research done in private labs compared to academic institutions and public bodies. The LLM or foundation model should be used to test the newest alignment and safety work with access extended to private firms who are focused on AI safety. This would help to (or maybe just would) accelerate all safety work and provide  access to foundation models to smaller firms and researchers may not ordinarily be able to access without a restrive API.

Rather than engaging in profitable ventures that compete with private firms, the government should prioritise addressing market failures where private competition is limited. This approach enhances the government's prospects of successfully creating a competitive model while minimising the risk of displacing private firms from the market.

# Facilitate the Safe Use of APIs for Innovative SMEs and Researchers

Recommendations:

1. Enable SMEs and researchers to develop products and carry out safe research through APIs accessed on the research resources.

2. Implement risk based requirements for API access to reduce the risk of misuse and encourage private participation.

A research paper prompted GPT-3 "suggested four potential pandemic pathogens, explained how they can be generated",[121] and "supplied the names of DNA synthesis companies unlikely to screen orders."[122] This could inadvertently facilitate the misuse of biotech to create bioweapons - democratising access to LLMs which could give terrorists the ability to make synthetic versions of the black plague, but more deadly and transmissible.

An API allows closed-sourced models such as OpenAI's GPT-4 to be used by third-party developers, such as Duolingo, to power the newest iteration of their app with OpenAI's AI model. Alongside the release of GPT-4 Open AI announced they would not be fully open-sourcing their model, with Ilya Sutskever, OpenAI's chief scientist, citing fears over safety and competition.[123] Alongside the significantly reduced costs, the Centre for AI Governance predicted that, increasingly, important AI research will be facilitated by APIs.[124]

With most new innovations made by small developers utilising the APIs being productive and non-harmful, we want to ensure this is always the case. However, 'AutoGPT' and 'ChaosGPT' are two new API-powered AI models of particular concern.[125]

AutoGPT can work in the background without the need for human interaction. A companion system, instructed by AutoGPT, uses GPT and associated APIs to develop further responses and actions from the initial request, without requiring additional human input. Meaning potentially dangerous properties could emerge without human oversight.

Chaos-GPT was a model developed from Auto-GPT instructed to be a "destructive, power-hungry, manipulative AI."[126] Which established its main objectives - including destroying humanity and gaining immortality - and attempted to take steps to obtain these goals with it identifying nuclear fallout as the best

[121] Soice, E., et. al., (2023), Can large language models democratize access to dual-use biotechnology?, https://arxiv.org/abs/2306.03809

[122] Ibid.

[123] Vincent. J., (2023), OpenAI co-founder on company's past approach to openly sharing research: 'We were wrong', The Verge, https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview

[124] Anderljung. M., Heim. L. and Shevlane, T., (2022), Compute Funds and Pre-trained Models, Centre for AI Governance, https://www.governance.ai/post/compute-funds-and-pre-trained-models

[125] Matt, (2023), AutoGPT: The AI That Can Self-Improve!, AutoGPT, https://autogpt.net/autogpt-the-ai-that-can-self-improve-is-scary/; Lanz. J., (2023), Meet Chaos-GPT: An AI Tool That Seeks to Destroy Humanity, Decrypt, https://decrypt.co/126122/meet-chaos-gpt-ai-tool-destroy-humanity

[126] Ibid.

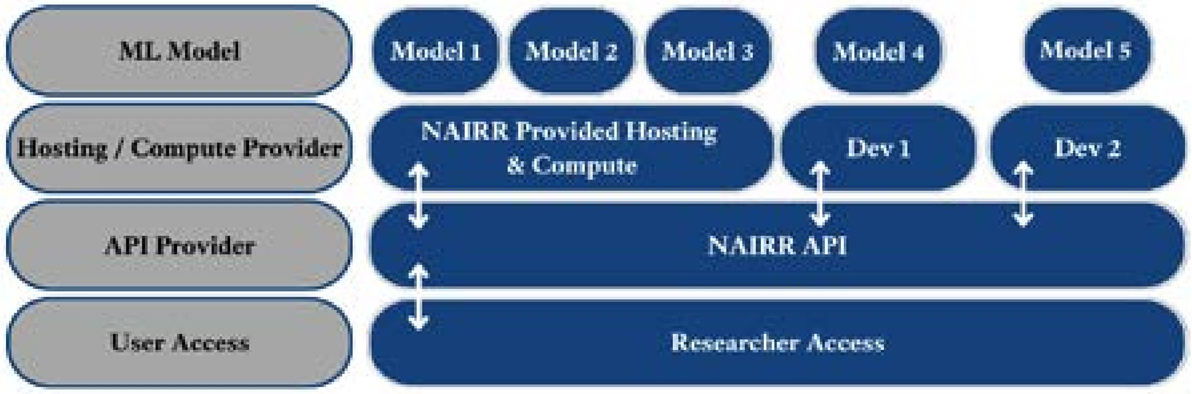
way of destroying humanity.

**Enable SMEs to use APIs safely**

This platform should facilitate both the use and distribution of large APIs. Incorporating models created by both private developers and researchers who have leveraged compute from the research resource. To maximise the efficiency of this, the platform must facilitate a broad spectrum of different research and experiments across multiple models. Additionally, while the use of APIs reduces the cost of research, researchers should still be able to utilise funding and compute from the research resource on the models provided.

This would facilitate in-depth exploration of model interpretability, alignment, bias, and other critical aspects. Moreover, researchers would have the freedom to introduce minor adjustments to the models for additional testing and fine-tuning, enhancing the overall investigative process.

The Centre for AI Governance demonstrates how this can be done below:



**Implement risk-based requirements for high-risk API access**

To further ease the concerns of the top AI labs and increase the chance of widespread adoption. The government should introduce requirements to prevent misuse of the model. As part of this, APIs should only be distributed through the research resource to verified researchers or firms. To access the APIs, applications should be made highlighting the intended use such that the host of the API knows who has access to what models, what the API is being used to develop, and the corresponding risk level.

The government should also introduce a tiered approach to regulation. Where applications carry greater scope for misuse, they require more regulation. In some cases limitations may be placed on the changes researchers can implement, or applications may be denied if the associated risk is not met by a clear benefit. However, for most uses this should just constitute effective monitoring by the government.

Introducing regulations to ensure that the model will not be used maliciously and that the research carried out will not have significant impacts on competition would provide a middle ground for labs where they can encourage wider research - specifically on safety - while maintaining their market position and preventing misuse unlike with open sourcing.

# Effective Procurement

Recommendations:

1. Introduce Challenge Based Procurement to improve the efficiency and reduce the barriers for smaller firms;

2. The Office for AI should identify opportunities for procurement to support proof of concept work too risky for nationwide deployment;

3. Introduce procurement for AI assurance within the public sector to support private sector firms and ensure safe deployment.

Access to funding is vital for UK AI firms to remain competitive amid high development costs, restricted labour supply, and global competition. However, many firms struggle to secure sufficient funding due to long product cycles, limited early revenue, a macro-capitalisation squeeze, and a funding gap for Series B+ firms.[127] While total UK AI investment reached $6 billion in 2021, it fell to just over $5 billion in 2022 due to economic headwinds.[128] With micro and seed firms also experiencing reducing proportions of funding as investors look for products closer to commercialisation.

Without adequate funding, UK firms risk being acquired by foreign companies that can offer more capital, opening up the potential for operations moving abroad. Alternatively, underfunded firms must spend more time fundraising or commercialise early, pivoting resources from innovation too early. Consequently, ensuring UK AI firms can secure the funding needed to scale effectively is critical for the sector to maximise economic benefits and remain competitive on a global stage. Increased public and private investment will be needed to bridge this gap and support AI firms through later stages of growth - with the AI Sector Study consultation highlighting the potential of an increased role for public procurement.[129] Given the national security concerns surrounding AI safety research, government funding through UKRI should be devoted to allowing researchers to purchase cloud computing power.

### Challenge Based Procurement

Public procurement is flawed. Primarily it is difficult to navigate, with only 6.1% of startups finding it easy to work with the government.[130] Small firms also identify many barriers, namely long tendering processes, late payments, and a lack of awareness of opportunities within the public sector.[131] If utilised properly the government could shape standards, ensure genuinely innovative firms receive the necessary funding, and increase the efficiency of the public sector.

[127] PitchBook Analyst Note: When Dry Powder Stays Dry, (2023), Pitch Book, https://pitchbook.com/news/reports/q1-2023-pitchbook-analyst-note-when-dry-powder-stays-dry; National AI strategy, Gov.UK, December 2022, https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version

[128] Artificial Intelligence sector study 2022, Gov.UK, (2023), https://www.gov.uk/government/publications/artificial-intelligence-sector-study-2022/artificial-intelligence-sector-study-2022-ministerial-foreword-and-executive-summary

[129] Ibid.

[130] Gerdon. S. and Molinari. V., (2020), How governments can use public procurement to shape the future of AI regulation – and boost innovation and growth, World Economic Forum, https://www.weforum.org/agenda/2020/06/artificial-intelligence-ai-government-procurement-standards-regulation-economic-growth-covid-19-response/#:~:text=By%20utilizing%20public%20procurement%2C%20governments%20could%20support%20AI,human%20and%20ethical%20implications%20of%20Artificial%20Intelligence%20%28AI%29.

[131] Ibid.

Challenge-based procurement is where the government identifies a problem within the public sector and allows firms to compete over different innovative ways of fixing it. As opposed to the government identifying a problem, and asking companies to provide specific technologies to fix it. While the government introduced a £20 million 'tech catalyst fund' focused on applying these principles it should be applied more broadly and become the norm for AI procurement. [132]

This would simplify the current system of procurement. Currently high administrative costs, due to specific requirements, favour larger and incumbent firms. 39.4% of firms find  it extremely difficult to complete a government tender and only 4.5% finding it easy.[133] This limits the true level of competition in the tendering process and prevents smaller firms from accessing additional funding.

### The Office for AI and opportunities for procurement

The Advanced Research & Invention Agency (ARIA) is a government body designed to fund high-risk, high-reward scientific research initially inspired by the US's Defense Advanced Research Projects Agency (DARPA)[134] - a US funding body which has been, to some extent, responsible for the development of GPS, the internet, and more recently the Moderna Covid-19 Vaccine.[135]

The government should adopt a similar principle for AI by offering riskier procurement opportunities to innovative small firms. These should then act as tests for proof of concept work before possible broader deployment within the public sector. They should accept that many of these potential solutions may not be successful but the gains from new approaches, when implemented nationwide, would be incredibly significant.

Alongside funding they should engage in the systematic approach of early stage, high-risk firms who often lack incentives to engage with the procurement system. This would help create new market opportunities for high-risk early-stage companies, who have been particularly affected by falls in investor confidence and increasing priority placed on commercialisation. Alongside increasing contestability further down the market chain as the most innovative firms can more successfully scale up.

### Procurement for AI assurance within the public sector to support private sector firms and ensure safe deployment

Public procurement also presents a significant opportunity for the government to ensure that AI development is safe. The UK currently has 17 AI assurance companies, utilising procurement for AI safety within the public sector would allow the government to further incentive alignment.[136] While introducing best practices within procurement - similar to those proposed by the Alan Turing institute - can help shape norms within the development of AI.

[132] GovTech Catalyst overview, Gov.UK, (2020), https://www.gov.uk/guidance/govtech-catalyst-overview

[133] Gerdon. S. and Molinari. V., (2020), How governments can use public procurement to shape the future of AI regulation – and boost innovation and growth, World Economic Forum.

[134] Gabriel. M., (2020), ARPA: what is it and why does Dominic Cummings want one in the UK?, The Conversation, https://theconversation.com/arpa-what-is-it-and-why-does-dominic-cummings-want-one-in-the-uk-130975

[135] ARPANET, Defence Advanced Research Projects Authority, Accessed June 2020, https://www.darpa.mil/about-us/timeline/arpanet; Removing the Viral Threat: Two Months to Stop Pandemic X from Taking Hold, (2017), Defense Advanced Research Projects Agency, https://www.darpa.mil/news-events/2017-02-06a

[136] Artificial Intelligence sector study 2022, (2023), Gov.UK, https://www.gov.uk/government/publications/artificial-intelligence-sector-study-2022/artificial-intelligence-sector-study-2022-ministerial-foreword-and-executive-summary

In the government's AI assurance temperature check participants identified that assurance provides a competitive edge for private firms by building trust and reducing the chance of reputational damage, as such developing the assurance sector within the UK could provide a comparative advantage for UK firms.[137] Additionally since AI assurance is not regularly procured globally the UK has an opportunity to lead in this field - which will likely grow as concerns over misaligned AI become more prevalent.

[137] Industry temperature check: barriers and enablers to AI assurance, (2022), Gov.UK, https://www.gov.uk/government/publications/industry-temperature-check-barriers-and-enablers-to-ai-assurance

# Saving Lives with AI-Powered Medicine and Reducing Engineered Pandemic Risk

Recommendations:

1. The NHS should invest in generalist medical AI capabilities through the NHS AI Lab;

2. Introduce the Three Lines of Defence Structure to ensure the UK is proactively prepared for biosecurity risks;

3. Invest in pathogen monitoring systems and introduction of bioengineering licences.

**NHS to invest in Generalist Medical AI capabilities through the NHS AI Lab**

AI is likely to usher in newfound capabilities in medicine.[138] It is already being used in medical computer systems to "[diagnose] patients, end-to-end drug discovery and development, improving communication between physician and patient, transcribing medical documents, such as prescriptions, and remotely treating patients."[139]

The National Health Service's Artificial Intelligence Laboratory (NHS AI Lab) was "created to address that challenge by bringing together Government, health and care providers, academics and technology companies."[140] It should be empowered to continue in this endeavour and be allowed to innovate with new products and services.

**Introduction of Three Lines of Defence Structure to ensure the UK's is proactively prepared for biosecurity risks**

Deepmind's Alphafold is an AI system that can predict the 3D structure of proteins with high accuracy.[141] Predicting protein folding, the process by which proteins take on their 3D structures, has been a long-standing challenge in biology. Alphafold represents a breakthrough in this area.

The ability to accurately predict protein folding could help scientists design new drugs and therapies. Many diseases are caused by misfolded proteins, so understanding and manipulating protein folding could offer medical benefits. Alphafold could help speed up drug discovery and development by identifying potential drug targets and candidate molecules.

[138] The NHS AI Lab, NHS England - Transformation Directorate, https://transform.england.nhs.uk/ai-lab/

[139] Basu, K., Sinha, R., Ong, A. and Basu, T., (2020), Artificial Intelligence: How is It Changing Medical Sciences and Its Future?, National Library of Medicine, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7640807/

[140] The NHS AI Lab, NHS England - Transformation Directorate, https://transform.england.nhs.uk/ai-lab/

[141] AlphaFold, Google DeepMind, https://www.deepmind.com/research/highlighted-research/alphafold

However, AI systems like Alphafold also raise dual-use concerns. The same capabilities that can advance medical research could also potentially be misused. Alphafold could help design synthetic pathogens by predicting how to modify proteins to create functional and potentially harmful viruses or bacteria. This could aid the production of bio weapons for state and non-state actors and even start the next pandemic. AI-enabled protein design could speed up the development of biological weapons like toxic proteins or pathogens targeting specific groups.

While the potential benefits of systems like Alphafold for medical breakthroughs are immense, policymakers, researchers, and the public must also be mindful of potential dual-use risks and misuse. Open access, ethics guidelines, and governance frameworks could help maximise the benefits of these technologies while minimising harmful applications.

What is striking is that the open-source code is now available for use by online coders.[142] This is dangerous because open-sourcing Alphafold means making the underlying code, data and research publicly available for anyone to use and modify. While open science has many benefits, open sourcing a powerful tool like Alphafold also has risks.

Making the code publicly available means potentially hostile actors like terrorists or malicious state actors could acquire and exploit Alphafold's capabilities. They could modify the code to customise it for harmful purposes like designing bioweapons. The open access nature of open sourcing removes some of the traditional barriers that may have prevented these actors from developing similar technologies on their own.

The open-sourcing the data Alphafold was trained on could provide valuable information for illicit protein engineering efforts. The data could give adversaries insights into the optimal parameters for manipulating proteins and designing synthetic pathogens. They could utilise this knowledge to accelerate their nefarious research programs.

A system like Three Lines of Defence would ensure that risks posed from AI and biosecurity in particular would be 'sufficiently captured in UK risk management.'[143] There needs to be a focus on proactive preparedness and antifragility, rather than just reacting to risk when it comes and losing more lives as a result. The government illustrated with their revised 2023 version of The Orange Book that they theoretically understand the Three Lines of Defence Model them but we want to see it go further - and then actually implement one across Whitehall.[144]

Dr Toby Ord and the Future of Humanity Institute at Oxford University, imagines a structure as follows:[145]

- 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

[142] Deepmind / alphafold, Github, https://github.com/deepmind/alphafold

[143] Ord, T., Mercer, A., and Dannreuther, S., (2021), Future Proof: The Opportunity to Transform the UK's Resilience to Extreme Risks, Centre for Long-Term Resilience, https://www.fhi.ox.ac.uk/wp-content/uploads/three_lines_defence.pdf

[144] The Orange Book Management of Risk – Principles and Concepts, (2023), Gov.UK, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1154709/HMT_Orange_Book_May_2023.pdf

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

focus on extreme risks in areas including AI and biological security;

- 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;
- An independent National Extreme Risk Institute. This would provide an audit and advisory function to the CRO.

Dr. Ord imagines this would cost around £8.26 million.[146] This is a pittance for reduced risk to AI and biological risks.

**Invest in pathogen monitoring systems and introduction of bioengineering licences**

The UK should establish a national pathogen surveillance network to monitor for potential biological threats. This could involve sequencing large numbers of pathogens circulating in the population and the environment. Any unusual or modified pathogens could then be flagged for further investigation.

At a minimum, it could require all life scientists and biosafety facilities working with select pathogens to report any incidents or results of concern to the Cabinet Office to ensure potential risks are detected early. This should include a licensing system for research involving the manipulation of certain pathogens deemed high-risk, and if they are high-risk, researchers should have to apply for and be granted a licence before proceeding with such work.

As part of current inspections, public inspectors should periodically visit biosafety labs conducting high-risk bioengineering research to ensure they are following proper protocols, security procedures and policies. Any violations could result in the suspension of licences. We  also want to see restrictions to the open-access of sensitive data and material - the UK may also choose to restrict the open sharing of certain bioengineering methods or data considered too dangerous to be made publicly available. Access could be granted on a need-to-know basis to authorised researchers.

[146] Ord, T., Mercer, A., and Dannreuther, S., (2021), Future Proof: The Opportunity to Transform the UK's Resilience to Extreme Risks, Centre for Long-Term Resilience, https://www.fhi.ox.ac.uk/wp-content/uploads/three_lines_defence.pdf

# Implement a Review of the Possible Labour Effects of a Future ASI

Recommendations:

1. Produce a White Paper on what the introduction of an UBI or a NIT would look like in a worst-case scenario;

2. Introduce NIT and UBI trials to prepare for the possibility of AI caused unemployment.

Fears of mass unemployment nearly always accompany the introduction of transformative technologies. A few decades ago the internet was 'destined' to bring about mass unemployment. Now it contributes 10% to US GDP and has created millions of jobs.[147]

Before that, the Luddites destroyed new textile machines, interest groups regulated against the motorised car, and people believed that a train ride could cause instant insanity.[148] Transformative technologies however have always complemented labour. Even if shorter-term employment shocks were experienced, human ingenuity created new industries reinstating the demand for labour.[149]

Despite this, we believe that AI has the potential to be different. Primarily as algorithms are refined, data sets are expanded and more computing is utilised, systems will become more sophisticated and some people believe these systems will also have agency, allowing them to plan and execute its own objectives and goals.[150] While unintended emergent properties may also develop and present themselves post-deployment.

The combination of these factors make this technology unusually likely to replace cognitive as well as physical labour. With the pryor being particularly concerning due to the potential to reduce human labour's comparative advantage of brain power. Resultantly we believe that in the long run, with sophisticated enough systems, it is possible that we may see an increased level of sticky unemployment.

## White Paper and trials

[147] Hooton. C., Measuring The U.S. Internet Sector: 2019, WayBackMachine, https://internetassociation.org/publications/measuring-us-internet-sector-2019/

[148] Agnew. J., (2020), Steam engines on UK roads, 1862–1865: Banning orders, agricultural locomotives and the 'red flag' Act, Taylor and Francis Online, https://www.tandfonline.com/doi/abs/10.1080/17581206.2020.1797447?journalCode=yhet20; Hayes. J., (2017), The Victorian Belief That a Train Ride Could Cause Instant Insanity, Atlas Obscura, https://www.atlasobscura.com/articles/railway-madness-victorian-trains#:~:text=As%20Edwin%20Fuller%20Torrey%20and%20Judy%20Miller%20wrote,or%20trigger%20violent%20outbursts%20from%20a%20latent%20%E2%80%9Clunatic.%E2%80%9D; Hötte. K. Somers. M. Theodorakopoulos. A., (2022), Technology and jobs: A systematic
literature review, Oxford Martin School.

[149] Acemoglu. D. Restrepo. P., (2019), Automation and New Tasks: How Technology Displaces and Reinstates Labor, American Economic Association, https://www.aeaweb.org/articles?id=10.1257/jep.33.2.3

[150] AI and compute, Open AI, May 2018, https://openai.com/research/ai-and-compute

If this is to occur it is important that the government has a plan in place to prevent significant harms. Resultantly the government should launch an investigation into the effect of Universal Basic Income (UBI) and a Negative Income Tax (NIT). This should take the form of a White Paper examining the potential long term impacts of AI on unemployment alongside worst case scenarios.

Alongside this the government should introduce large scale, localised trials of NIT and a more limited one for UBI. There have been a multitude of different UBI trials with varying methodologies, scale, and location. This has facilitated relatively extensive literature reviews on specific impacts of a UBI.[151] However within the UK the first trial was only announced at the start of June 2023, and only includes 30 recipients being paid over the course of 2 years. [152]

Further trials within the UK are necessary to see how the UK's society and culture impact the actions of recipients. While the government should look to address areas with less extensive research. Namely the Long term impacts, the effects of a truly universal payment, and a NIT. Currently many trials are focused on payments to 'poorer' recipients and while the long term impacts have still been studied, there is less extensive research as opposed to other sections.[153]

[151] de Paz-Báñez. M. Asensio-Coto. M. Sánchez-López. C. Aceytuno. M., (2020), Is There Empirical Evidence on How the Implementation of a Universal Basic Income (UBI) Affects Labour Supply? A Systematic Review, Pre Prints, https://www.preprints.org/manuscript/202008.0638/v1

[152] McNamee. S., (2023), Universal basic income: Plans drawn up for £1,600 a month trial in England, BBC, https://www.bbc.co.uk/news/uk-65806599

[153] de Paz-Báñez. M. Asensio-Coto. M. Sánchez-López. C. Aceytuno. M., (2020), Is There Empirical Evidence on How the Implementation of a Universal Basic Income (UBI) Affects Labour Supply? A Systematic Review, Pre Prints, https://www.preprints.org/manuscript/202008.0638/v1