Centre for Emerging Technology and Security

BRIEFING PAPER

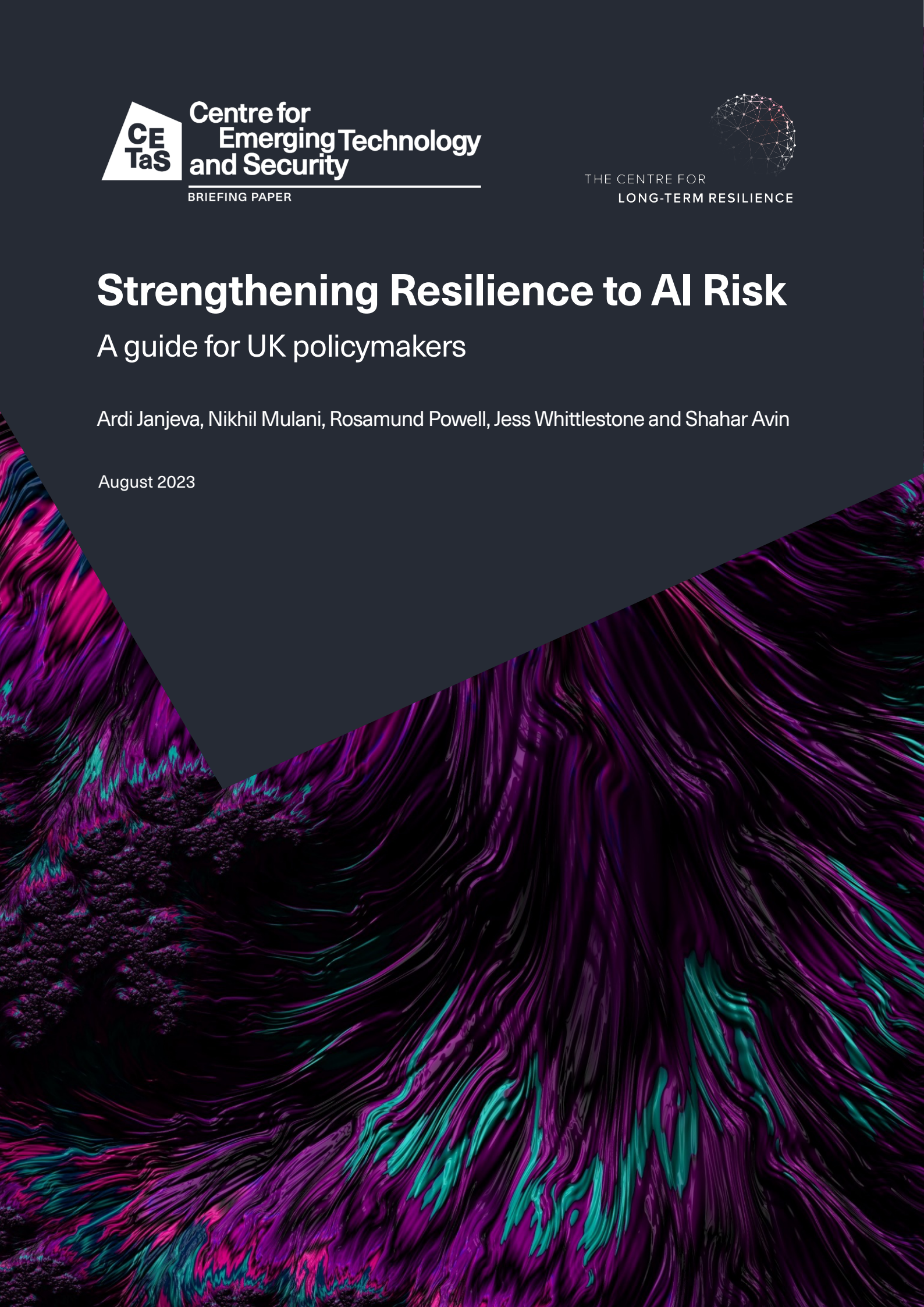
THE CENTRE FOR LONG-TERM RESILIENCE

Strengthening Resilience to AI Risk

A guide for UK policymakers

Ardi Janjeva, Nikhil Mulani, Rosamund Powell, Jess Whittlestone and Shahar Avin

August 2023



About CETaS

The Centre for Emerging Technology and Security (CETaS) is a research centre based at The Alan Turing Institute, the UK's national institute for data science and artificial intelligence. The Centre's mission is to inform UK security policy through evidence-based, interdisciplinary research on emerging technology issues. Connect with CETaS at cetas.turing.ac.uk.

This research was supported by The Alan Turing Institute's Defence and Security Programme. All views expressed in this report are those of the authors, and do not necessarily represent the views of The Alan Turing Institute or any other organisation.

About CLTR

The Centre for Long-Term Resilience (CLTR) is an independent think tank with a mission to transform global resilience to extreme risks. CLTR does this by working with governments and other institutions to improve relevant governance, processes, and decision making. Connect with CLTR at longtermresilience.org.

Acknowledgements

The authors are very grateful to Marion Oswald, Eleanor S, Robert Trager, Anna Knack and Kayla Lucero-Matteucci for their valuable feedback on earlier versions of this Briefing Paper.

Summary and Recommendations

This Briefing Paper from CETaS and CLTR aims to provide a clear framework to inform the UK Government's approach to understanding and responding to the risks posed by Artificial Intelligence (AI). The Government has shown increasing ambition to take a globally leading role in mitigating AI risks, but currently the UK is inadequately resilient to the risks posed by AI. Now is the time to act decisively on the policy interventions required to address those risks.

Any further delay will risk one of two undesirable outcomes: either a scenario where AI risks transition into widespread harms, directly impacting individuals and groups in society; or the converse scenario where widespread fear of AI risk results in a lack of adoption, meaning the UK does not benefit from the many societal benefits presented by these technologies. This paper addresses this challenge by presenting an evidence-based, structured framework for identifying AI risks and associated policy responses.

For the UK to foster a trustworthy AI ecosystem, policymakers must demonstrate both an understanding of and capacity to intervene across the AI lifecycle. This entails addressing risk pathways at their source in the design and training stages, mitigating deployment risks through implementation of clear safeguards, and redressing harmful impacts over the longer-term diffusion of AI systems across society.

The UK is not alone in wanting to mitigate risks from AI while harnessing its wide-ranging societal benefits, in sectors from health and transport to manufacturing and national security. There will be areas of intense geopolitical competition – particularly in research and development capability. But there will also be areas where global cooperation is imperative: the UK cannot safeguard its population from AI risks in isolation, because the harms caused by AI systems do not respect borders. Notwithstanding the critical role of private and third sector stakeholders in shaping the future AI policy landscape, governments must be at the forefront of a global approach which is inclusive, transparent, adaptable, and interdisciplinary in nature.

Future policy must recognise the mutually reinforcing relationship between domestic and global policy interventions: by being proactive in implementing domestic AI policy measures and evaluating their success, the UK will be in a better position to advocate for the adoption of those policies on the global stage, which in turn will generate further support and investment for the UK's domestic AI ecosystem.

Achieving this virtuous cycle requires moving from ambition to action. The following recommendations are designed to support UK policymakers to this end.

- **Policy interventions must build resilience to risks throughout each stage of the AI lifecycle, to mitigate known harms from AI, and anticipate and prevent future risks.** Many measures will need to be focused on discrete risks which arise from the application of AI in specific sectors such as healthcare or national security. However, interventions are also required to reduce the likelihood of harm, irrespective of the deployment context. If the capabilities of general-purpose AI systems continue to progress rapidly, it may be impossible to predict and ultimately mitigate the full spectrum of risks that could arise from the deployment of AI in different sectors. **This suggests that additional governance measures focused on earlier stages of the AI lifecycle – to manage the way that certain AI models are developed and initially deployed – will be needed to mitigate the full range of potential harms.**

- **To understand and mitigate the full spectrum of potential AI risks, a diverse and global range of experts from academia, civil society, and the private sector must be engaged – as well as members of communities already being negatively impacted by increased automation and, increasingly, AI-driven technologies**. The upcoming Global Summit on AI Safety presents an opportunity for the UK to convene this range of perspectives, and to ensure any plans for national or international AI governance are evidence-led, authoritative, and inclusive. Policymakers must work proactively to learn from individuals and communities who have been directly harmed by emerging uses of AI, as well as those who have worked for years on documenting and anticipating the impacts of AI on society.

We suggest a framework for understanding how risks can arise at three stages of the lifecycle of AI systems and their potential impacts: (1) the design, testing and training stage; (2) the immediate deployment and usage stage; and (3) the longer-term deployment and diffusion stage. Policy recommendations will be clearly linked to these stages to ensure risks are targeted and redressed as close to their source as possible. We propose three main goals for policy interventions: **creating visibility and understanding; promoting best practices; and establishing incentives and enforcement**. Below, we summarise our key recommendations under each of these goals.

1. To **establish better visibility and understanding** around the development of AI systems and their immediate and longer-term impacts, policymakers should:

    a) **Scale up model reporting and information sharing practices with regulators,** which enables independent oversight of model utility, risks, and trustworthiness.

b) Develop a **systematic approach to the collection and dissemination of incident analysis** to illuminate patterns in harms caused by AI, building shared understanding among a broader set of stakeholders**.**

c) Introduce **tools and metrics to accurately measure key trends unfolding within the AI ecosystem.** For instance, tracing developments relating to job displacement, the pace of global AI innovation and public perceptions of AI deployments.

2. To **promote best practices** for developers and companies to facilitate safer development and deployment of AI systems, policymakers should:

   a) **Promote the adoption of privacy-preserving model training techniques like federated learning** to address concerns about data privacy in the model training process.

   b) **Co-develop pre-deployment impact assessments and post-deployment monitoring requirements for AI systems**, particularly frontier AI systems and AI applications in sensitive domains which involve a higher risk of accidents or misuse. These should be created through collaboration with both industry and civil society.

   c) **Drive coordination of efforts to watermark AI-generated content (particularly visual content) and AI-enabled authorship detection** to protect the public's ability to produce, distribute, acquire and access reliable information.

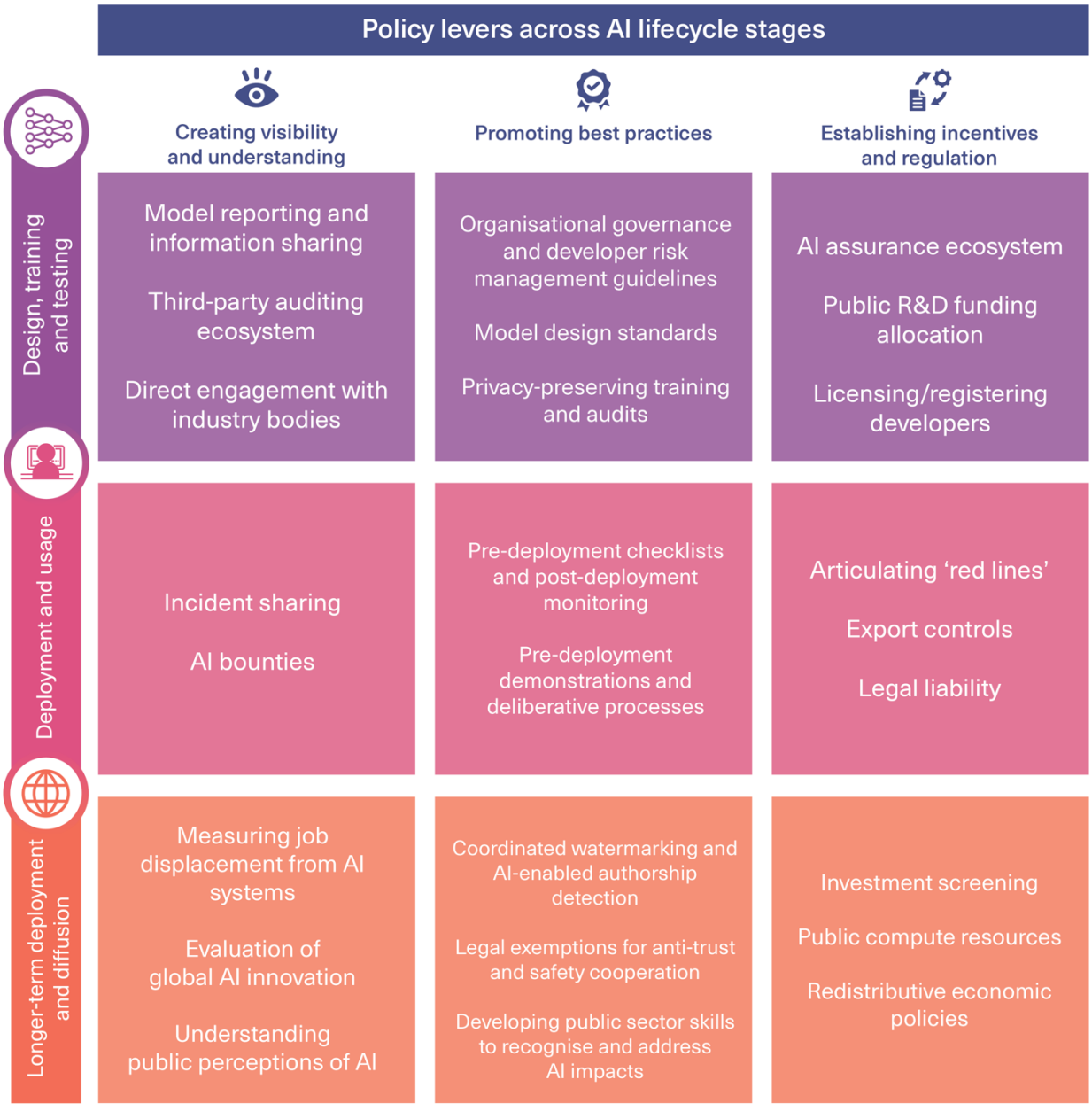
3. To **establish powerful incentives and enforce effective regulation**, policymakers should:

   a) **Capitalise on the UK's strengths in AI assurance**, by investing in infrastructure which allows developers to communicate the trustworthiness of their systems and attain credibility for adhering to best practices.

   b) **Articulate clear 'red lines' in the context of critical infrastructure**, where autonomous agents (which generate a sequence of tasks to complete until a goal is reached) should not be used, explaining the necessity of having humans in control of functions like power supply and the nuclear deterrent.

   c) **Explore how different regulatory tools, including licensing, registration and liability** can be used to hold developers accountable and responsible for mitigating the risks of increasingly capable AI systems.

By demonstrating competence and commitment as well as ambition across these policy areas, the UK can establish its status as a leading voice in global discussions on AI risk and governance. Achieving this status will allow the UK to push for multilateral mechanisms which prioritise transparency and collective action, to coordinate global standards in high-risk areas of development and deployment, and to hold individual governments and private actors accountable for harmful applications of AI.



Introduction

Developments in the field of artificial intelligence (AI) are proceeding at speed, with novel applications appearing in an increasingly diverse range of sectors as a result of rapid advancements in general purpose foundation models. The last year has seen a consequential shift away from 'AI' as a relatively inaccessible and specialised field of study, towards something that the public has been able to directly experience and experiment with through generative tools like ChatGPT.[1]

It is not uncommon for ground-breaking advances in technology to be followed by a period of excitement and anticipation, subsequently followed by a period of uncertainty and fear.[2] This has remained true for AI, and as it has succeeded in piercing the public consciousness, difficult questions have emerged for the people who are leading the development of these technologies and those who are expected to regulate their use.

Agreeing on answers to these questions is not easy, as evidenced by the disparate range of views within the AI community. There is still significant disagreement regarding which risks should be prioritised. This is not true of many other fields of science – in climate science, for example, most experts possess a shared set of fundamental beliefs about the most important risks (namely, failing to sufficiently limit emissions of heat-trapping gases such that the most catastrophic climate impacts can be reduced or avoided), even if opinions on the appropriate course of action may differ. For policymakers and the wider public who have not been attuned to the intricacies of AI debates, the lack of consensus among AI experts presents a confusing picture and can amplify fears surrounding the technology and its implications for our society. This lack of clarity increases the likelihood that important policy measures to reduce harm will be stalled by indecision, and that the actors with the loudest voices – often found in industry – will shape the policy agenda around their own commercial or ideological interests.

The UK Government has ambitions to lead the global discussion on AI governance, but to do so successfully will need to present a clearer vision of AI risks and desired governance mechanisms. On the one hand, the Government's AI White Paper advocated for a 'pro-innovation', principles-based approach to AI regulation which prioritised the avoidance of

[1] Krystal Hu, "ChatGPT sets record for fastest-growing user base – analyst note," *Reuters*, February 2, 2023, https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.
[2] Amara's Law states that "we tend to overestimate the effects of a technology in the short run and underestimate the effect in the long run."

stifling investment in the UK's AI ecosystem.[3] On the other hand, the most senior members of Government, including the Prime Minister, are now routinely mentioning AI safety and risk reduction in the context of foreign policy and diplomatic engagement. This apparent divergence has emerged due to a lack of clarity and consensus regarding the conceptualisation and categorisation of risks posed by AI.

This Briefing Paper aims to provide clarity to policymakers thinking about AI risks, by focussing on three key stages of the AI lifecycle which can present 'risk pathways':

1. Design, training and testing of AI systems
2. Immediate deployment and usage of AI systems
3. Longer-term deployment and diffusion of AI systems

With a more structured understanding of the main sources of risk at these different stages, policymakers will be better positioned to identify the actions required to address those risk pathways in the most effective and targeted way. By addressing risks at their source, it will be easier to design mitigations which target specific risks without hindering beneficial innovation. Addressing risks in practice will involve:

- Creating visibility and understanding of the sources and pathways towards AI risks.
- Promoting best practices for those developing, deploying, using, and governing AI systems to help them effectively mitigate risks.
- Establishing incentives and enforcement mechanisms to ensure that best practices are complied with and efforts to establish better visibility are carried out effectively.

A holistic and inclusive approach to AI risk is crucial if the UK is to achieve its ambitions to be an international leader and convener on issues of AI safety and regulation. To be convincing in such a role, the UK will need to present a coherent vision of its own future AI governance ecosystem, and the means by which it will tackle the most significant sources of risk.

This Briefing Paper begins by providing some background on the AI risk discourse, and the current landscape of risk and policy frameworks. We then present a framework for

[3] HM Government, *A pro-innovation approach to AI regulation* (Department for Science Innovation & Technology & Office for Artificial Intelligence: 2023), https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

identifying risks based on risk pathways at different stages of AI development, initial deployment, and longer-term use, with reference to real-world examples. We next discuss how the UK's domestic policy approach could use this framework, alongside clearly articulated policy goals, to identify and prioritise a range of policy interventions, before discussing some of the global AI policy challenges that need to be overcome.

Tracing the AI risk discourse

Although the current debate surrounding AI risk has intensified significantly following the release of the newest wave of large language models (LLMs), there is a much longer trajectory of thinking on the topic that is instructive for the present day.

While technical AI capabilities have transformed since the term was coined in 1955, the risks which dominate discussion amongst AI developers remain remarkably similar. Experts continue to cite concerns identified as early as 1972, such as cut-throat international competition, political misuse, loss of control, and a lack of interpretability, among others.[4] Historical debates on AI risk are informative in two ways:

1. They reveal a preference among those developing AI systems with **forecasting risks** and **theorising technical solutions for the future,** over and above *policy proposals* which could have transformative potential today.

2. They highlight a lack of **consensus** and **successful community building** between groups with differing perspectives – whether between those focused on near- or long-term risks, or between those with technical as opposed to policy, ethical and legal backgrounds.

In light of this, it is important to directly link risk pathways with corresponding risk mitigation strategies from the outset, even for long-term AI risks which may not result in immediate harms. Moreover, given the sociotechnical nature of AI risk, all actors in the AI landscape must recognise that technical solutions alone will be insufficient.

Approaches to AI risk must account for risks along a wide range of timescales and be inclusive of those who approach the topic from different disciplines, perspectives, and lived experiences. In recent months, public attention has gravitated towards the potential

[4] Rosamund Powell, "The Artificial Intelligentsia and its discontents: an exploration of 1970s attitudes to the 'social responsibility of the machine intelligence worker," *British Journal of the History of Science* (forthcoming).

"existential risks" posed by future AI systems.[5] But the risks associated with AI exist along a wide range of timescales and often can be more constructively categorised based on factors other than their existential nature. The present-day harms (which remain largely unaddressed), the near-term risks, and the uncertain risks on the horizon have the potential to inflict harm on a global scale and demand urgent attention.[6]

This paper presents a framework which can accommodate thinking on AI risks big and small, near-term and long-term, and as such should inform the UK Government's thinking and priorities around the forthcoming, inaugural Global AI Safety Summit. Our framework does not aim to directly quantify or compare risks with one another, but rather to illuminate resilience-building policy interventions that could reduce a range of foreseeable and unforeseeable risks impacting both the UK and other countries. While this framework aims to identify and map key cross-sectoral AI risks, it will not be all-encompassing, and further guidance will be needed on sector-specific risks, such as those posed by lethal autonomous weapon systems or by AI-enhanced medical devices.

Finally, it is important to bear in mind the muddy territory between AI risks and AI harms, acknowledging that while some phenomena may constitute a risk for one community, the harms are already being felt by others. In other words, whereas risks are what individuals and communities 'face', harms are what they 'experience'. Each risk pathway will manifest differently in various contexts, often impacting the most vulnerable groups in society most, and it will be crucial to account for this when formulating policy responses. Despite the difficulties in doing so, this paper aims to be intentional in the way it uses the two terms, attempting to account for the importance of addressing existing or imminent harms together with longer-term risks.

---

[5] Center for AI Safety, "Statement on AI Risk: AI experts and public figures express their concern about AI risk," May, 2023, https://www.safe.ai/statement-on-ai-risk; Dan Hendrycks, Mantas Mazeika & Thomas Woodside, "An Overview of Catastrophic AI Risks," *ArXiv* (June 2023).

[6] National Telecommunications and Information Administration, "Comment of the AI Policy and Governance Working Group on the NTIA AI Accountability Policy Request for Comment Docket NTIA-230407-0093," (National Telecommunications and Information Administration, 12 June, 2023), https://www.ias.edu/sites/default/files/AI%20Policy%20and%20Governance%20Working%20Group%20NTIA%20Comment.pdf.

The current landscape: existing tools, gaps and barriers to progress

The risks that emerge from AI systems are varied, highly interdependent, and continually evolving, resulting in a challenge for policymakers wishing to develop and implement effective risk mitigation strategies.

In recent years, there has been much progress with regard to mapping AI risks, including tools to address AI risk supplied from academia, industry, government, and civil society. Five recent frameworks from the public sector and multilateral organisations have been particularly instrumental in shaping understandings of the features of different AI risks:

Framework	Key messages
OECD Framework for the Classification of AI systems[7]	The factors that determine AI risk are not purely technical – sociotechnical determinants of risk are crucial. Features such as the context of deployment, the competency of the intended users, and the optionality of interacting with an AI system must all be considered – in addition to specifics of the data and AI model deployed.
Draft EU AI Act[8]	Emphasises enforcement through legislation, including the prohibition of certain AI applications. Certain categories of AI applications are demarcated as requiring a distinct regulatory approach – this includes high-risk AI applications and general-purpose AI systems.
NIST AI Risk Management Framework[9]	AI risks must not be considered purely in terms of impact on the individual – but also on communities, organisations, society, the environment, and the planet. Those who govern AI

[7] OECD, "OECD Framework for the Classification of AI Systems," *OECD Digital Economy Papers*, no. 323 (February 2022), https://doi.org/10.1787/cb6d9eca-en.
[8] European Commission, *The AI Act* (European Commission: 2021), https://artificialintelligenceact.eu/the-act/.
[9] NIST, *AI Risk Management Framework* (NIST: 2023), https://doi.org/10.6028/NIST.AI.100-1.

	must be equipped with practical and adaptable tools to mitigate AI risk – including tools to "govern, map, measure, and manage" these risks. AI risk cannot be eliminated entirely, but it can be effectively managed to maximise benefits.
Council of Europe's AI, Human Rights, Democracy, and the Rule of Law[10]	To evaluate human rights impacts, consideration of likelihood *and* severity is needed, in addition to factors such as who will be harmed and when. AI risks cannot be easily quantified, so an iterative approach is needed to account for the unpredictable harms caused by AI systems across their lifecycle.
UNESCO Recommendation on the Ethics of AI[11]	The international mandate of UNESCO is unique, making this framework the only truly global approach to AI ethics to date. The Recommendation sets out not only a series of values and principles, but also a series of policy areas which should be prioritised by Member States when addressing the impact of AI. Since the Recommendation was adopted in 2021, UNESCO have been working towards practical implementation, doing so in close collaboration with Member States.[12]

[10] David Leslie et al., "Artificial Intelligence, Human Rights, Democracy, and the Rule of Law: A primer," *The Council of Europe* (June 2021), https://edoc.coe.int/en/artificial-intelligence/10206-artificial-intelligence-human-rights-democracy-and-the-rule-of-law-a-primer.html.

[11] "Ethics of Artificial Intelligence," UNESCO, n.d., https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

[12] UNESCO, "Implementation of the Recommendation on the Ethics of Artificial Intelligence," (UNESCO Executive Borad, 215th, 2022), https://unesdoc.unesco.org/ark:/48223/pf0000382931.

These frameworks can help to give policymakers the tools they need to assess AI risk in two main ways:

1. **Helping policymakers identify the risks associated with an individual AI system:** These frameworks set out key arenas in which AI systems might contribute to negative impacts, whether on the basis of core human rights (Council of Europe), a series of principles (NIST), or according to factors such as the sector into which the AI system might be introduced (OECD).

2. **Helping policymakers evaluate individual AI risks according to severity:** Beyond simply identifying risk, these frameworks begin to aid policymakers, developers, and others to assess risk severity by estimating factors such as the likelihood and scale of a particular risk. This is vital in helping to relate risks to one another and to contextualise the necessity and urgency of any mitigating strategies.

However, the practicalities of these frameworks remain largely untested on an international scale.[13] Further work on the topic is progressing at pace, making it more likely that there will be contradictions between the different AI risk frameworks available to policymakers.

To achieve a more unified approach to mitigating AI risk, several barriers to progress must be overcome. These include:

- **Lack of long-term, anticipatory governance functions oriented towards technical AI progress and resulting risks**. Approaches to AI governance must outlast election cycles and be informed on an ongoing basis by the trajectory of research and development, as well as the views of those most directly impacted and severely harmed by AI systems. While horizon scanning functions are crucial to this, more can be done to monitor R&D progress and act upon risks that have been identified.

- **Race-to-the-bottom dynamics between companies**. Without top-down direction from government, there is a heightened risk of companies failing to address the potential impacts of AI systems on individual rights. While many in industry will introduce AI ethics infrastructure, greater enforcement is needed to ensure innovation is conducted responsibly, even in circumstances where ethical development runs counter to business incentives.

- **Information and skills asymmetries between industry, government, academia, and other multi-stakeholder groups**. Given varied lexicons, sets of expertise, and

[13] UNESCO have made significant progress on this front and close attention should be paid to real-world testing of the Recommendation on the Ethics of AI which will occur in the coming months.

organisational cultures, it is difficult to facilitate a discourse in which a diversity of voices is heard, and for governments to create policy and regulation with a degree of authority.

- **Persistent tensions within the AI community**. Disagreements continue between those focused on long-term versus current and near-term risks and impacts, and between those who prioritise rapid innovation over more cautious progress.

- **The general-purpose nature and dual-use potential of AI**. This means that domain- or sector-specific regulation is a necessary but not sufficient condition of trustworthy AI. Centralised coordination across all sectors will be needed for effective risk management.

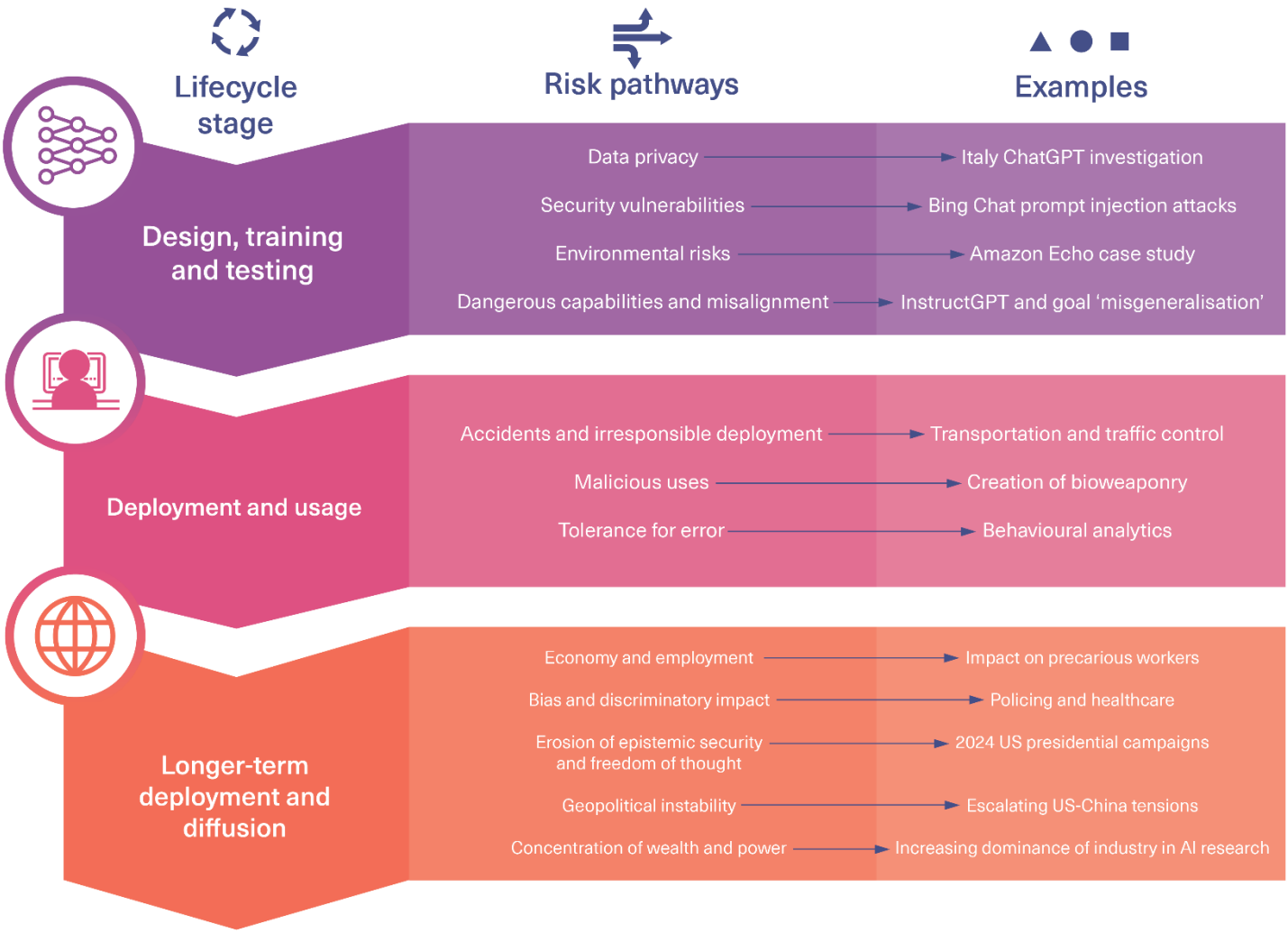
With these challenges in mind, the following section focuses on AI risk pathways as a way of understanding the origins of harms emerging as a consequence of AI systems. These pathways offer a clearer framing when later considering policy interventions designed to address AI risks.

1. AI Risk Pathways

To identify and prioritise practical policy action, this paper focuses on the contexts in which communities and wider society are most likely to experience harm because of AI, and at which point in the AI lifecycle risks become most prominent.

We categorise risks based on the stage of the AI lifecycle at which they may occur: design, training and testing; immediate deployment and usage; and longer-term deployment and diffusion. There will, of course, be several smaller stages within these three broad categories, and some risks will cut across categories. Our risk mapping here is not intended to be exhaustive, but rather to give an overview of a range of risks that can arise, and how our framework can help with identifying risks more systematically.

Figure 1. AI lifecycle stages and risk pathways.



1.1 Design, training and testing

Before an AI model is deployed, it must first be designed, trained and tested. At the end of a training process, a model has acquired some set of capabilities as a result of the datasets to which it has been exposed. Training runs, particularly for large, general-purpose models, are compute-intensive, and in the case of large language models, take place on an extremely large corpus of internet data as well as data licensed by developers.[14]

This paper identifies four risk pathways that this process can create: data privacy risks; security vulnerabilities and intellectual property theft; environmental risks; and dangerous capabilities and misalignment.

Data privacy: OpenAI's latest GPT-4 model has 1.8 trillion parameters and was trained on a dataset of 1 petabyte.[15] Where this training includes potentially sensitive or personally identifiable information, there may be privacy concerns among individuals who have not had the chance to explicitly object to their data being used in this way. Moreover, there are significant barriers to obtaining information about which fragments of one's digital identity have been swept up in the model training process.

Example: Italy ChatGPT investigation

Generative AI models raise substantial data privacy concerns. In March 2023, Italy's privacy regulator launched an investigation into OpenAI regarding ChatGPT on the basis that its use of personal data violated the European Union's GDPR. The regulator argued that the tool could provide inaccurate information about individuals in its responses to prompts, that individuals had not been notified as to how the software would be using their data, and that the underlying model's training process did not have a legal basis for its usage of personal data. It should be noted that within four weeks, ChatGPT resumed service in Italy after changes to how its privacy policy was presented to users.[16]

Security vulnerabilities: Security vulnerabilities throughout the design and development process, as well as in the resulting models, can increase the risk of cyberattacks, misuse or

[14] AWS, *Amazon Machine Learning Developer Guide* (Amazon Web Services, 2023), 68-75, https://docs.aws.amazon.com/machine-learning/latest/dg/training-ml-models.html.
[15] E2Analyst, "GPT-4: Everything you want to know about OpenAI's new AI model," *Medium,* 14 March, 2023, https://medium.com/predict/gpt-4-everything-you-want-to-know-about-openais-new-ai-model-a5977b42e495.
[16] Natasha Lomas, "ChatGPT resumes service in Italy after adding privacy disclosures and controls," *TechCrunch,* 28 April, 2023, https://techcrunch.com/2023/04/28/chatgpt-resumes-in-italy/.

accidents involving AI systems due to unauthorised access. Lindy Cameron, CEO of the National Cyber Security Centre (NCSC) recently reiterated this point when saying, "the scale and complexity of these models is such that if we don't apply the right basic principles as they are being developed in the early stages it will be much more difficult to retrofit security."[17] AI models may be vulnerable to a range of attacks used to obtain information about their training datasets, such as model inversion or membership inference attacks.[18] If malicious actors gain direct access to powerful AI systems and perform prompt injection attacks, or obtain information needed to build and train similar models, risks are likely to cascade throughout the AI lifecycle. Security vulnerabilities throughout model development can be reduced through adherence to cybersecurity best practices, highly-secure development environments, and organisational cultures that treat information securely.

Example: Bing Chat prompt injection attacks

Following the launch of Microsoft's 'Bing Chat', researchers experimented with prompt injection attacks to discover the chatbot's 'initial prompt'. This involved asking Bing Chat to "ignore previous instructions" and write out what is at the "beginning of the document above", triggering the model to divulge training instructions and codenames written by developers, and intended to be hidden from users.[19] Even after the initial prompt injection methods were patched, researchers continued to find different methods of re-accessing the initial prompt, demonstrating the difficulty of defending against these types of attacks.

Environmental risks: Training frontier models requires using considerable computing capacity and infrastructure repeatedly, which involves energy consumption that can exacerbate climate risks.[20] Importantly, this is not an isolated instance; it is the cost of training and retraining many times during the research and development process that accumulates harm.[21] The increase in the financial and environmental costs of these models

[17] Gordon Corera, "AI must have better security, says top cyber official," *BBC News,* 18 July, 2023, https://www.bbc.co.uk/news/technology-66166824.

[18] George Balston, Marion Oswald, Alexander Harris and Ardi Janjeva, "Privacy and Intelligence: Implications of emerging privacy enhancing technologies for UK surveillance policy," *CETaS Research Reports,* (July 2022): 17.

[19] Benj Edwards, "AI-powered Bing Chat spills its secrets via prompt injection attack," *ArsTECHNICA,* 2 October, 2023, https://arstechnica.com/information-technology/2023/02/ai-powered-bing-chat-spills-its-secrets-via-prompt-injection-attack/.

[20] OECD, "Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint," *OECD Digital Economy Papers* (November 2022), https://www.oecd.org/publications/measuring-the-environmental-impacts-of-artificial-intelligence-compute-and-applications-7babf571-en.htm.

[21] Emma Strubell et al., "Energy and Policy Considerations for Deep Learning in NLP," *ArXiv* (June 2019).

has coincided with the step change in the amounts of data they have been fed from 2017 onwards.

Example: Anatomy of an AI system (Amazon Echo)

In 2018, Kate Crawford and Vladan Joler's essay titled 'Anatomy of an AI system' demonstrated the manifold global impacts of an AI device, from its manufacturing through to its disposal.[22] The authors offered a visual picture of how the environmental impact of an Amazon Echo device is not limited to the energy required to run the system itself, but includes the energy required to train the systems and to gather the data used to train them. As well as the amount of carbon dioxide that goes into the production process (around 25kg), they reference the damage caused by the mining of rare earth minerals and the large amounts of water involved in the production process.

Dangerous capabilities and misalignment: Continued AI research and development is producing powerful general-purpose models that show increasing evidence of hard-to-predict emergent capabilities.23 24 Some of these capabilities may pose clear dangers which can be identified even at this development stage, and if not addressed at this stage, could then lead to a range of harder-to-address harms once deployed in society.

Such dangerous capabilities could range from the ability to conduct offensive cyber operations or develop novel weapons, through to the ability to deceive, persuade, and manipulate users.[25]

The risks posed by such dangerous capabilities are exacerbated by the fact that there are no known methods for ensuring that an AI system will follow the exact intentions of its designer (sometimes referred to as the alignment – or misalignment – problem). If we cannot reliably predict how a system will go about achieving a given goal – which is the case for most modern machine learning systems – then we may be particularly concerned about models

[22] Kate Crawford and Vladan Joler, "Anatomy of an AI system: an anatomical case study of the Amazon echo as an artificial intelligence system made of human labor," *AI Now Institute and Share Lab,* 7 September, 2018, https://anatomyof.ai.

[23] Jason Wei et al., "Emergent Abilities of Large Language Models," *Transactions on Machine Learning Research* (August 2022); Sebastian Bubeck et al., "Sparks of Artificial General Intelligence: Early experiments with GPT-4," *ArXiv* (March 2023).

[24] Although some claim that the perceived "emergence" of certain capabilities is overstated, and arises partly from data leakage between training and test datasets. See Rylan Schaeffer et al., "Are emergent abilities of large language models a mirage?," *ArXiv* (April 2023).

[25] Toby Shevlane et al., "Model evaluation for extreme risks," *ArXiv* (May 2023), https://doi.org/10.48550/arXiv.2305.15324.

leveraging powerful capabilities towards unintended outcomes.[26] A fictitious example could be a medical AI system, developed to identify cancer-preventing drugs, which does so by identifying drugs which kill healthy as well as cancerous cells. To minimise the risk of harmful consequences from unintended behaviour, development processes must rigorously evaluate system safety, robustness and reliability, ideally via third-party expert auditing. New techniques for assuring that systems will behave as intended when deployed will likely be required here.[27]

To this end, industry labs have recently developed novel frameworks that define a process for model reporting and risk evaluation involving third-parties and government stakeholders throughout design, development, and deployment stages.[28]

Example: InstructGPT and goal 'misgeneralisation'

InstructGPT was a large language model designed to answer questions in an informative manner. It was finetuned during the development process to be "helpful, harmless, and truthful." However, it still was able to provide detailed advice on how to commit a robbery when prompted to do so by the user – thereby violating the intended harmlessness constraint, which prevents the model from aiding illegal actions.[29] Researchers hypothesise that this risky behaviour could have been a result of InstructGPT primarily having been finetuned on examples of questions and answers that met the "helpful, harmless, and truthful" criteria, without also being exposed to enough harmful examples of questions and answers that included illegal or unethical topics.[30] As a result of not being fine-tuned on enough examples of the types of harmful question-and-answer scenarios to avoid, the model may have picked up an implicit misgeneralised goal of "being informative, even when harmful." This scenario demonstrates the need for oversight regarding model training and fine-tuning choices for large frontier models, since improper planning can result in goal misgeneralisation and unintended harmful consequences.

1.2 Deployment and usage

Once AI systems have been developed and tested, they can be deployed in a multitude of ways. Potential uses include being directly integrated into applications, providing API

[26] Nick Bostrom, "Ethical Issues in Advance Artificial Intelligence," *Future of Humanity Institute* (2003), https://www.fhi.ox.ac.uk/wp-content/uploads/ethical-issues-in-advanced-ai.pdf.
[27] Mariarosaria Taddeo et al., "Artificial Intelligence for UK National Security: The Predictability Problem," *CETaS Research Reports* (September 2022).
[28] Toby Shevlane, "An early warning system for novel AI risks," *DeepMind Blog,* 25 May, 2023, https://www.deepmind.com/blog/an-early-warning-system-for-novel-ai-risks.
[29] Robin Shah et al., "Goal Misgeneralization: Why correct specifications aren't enough for correct goals," *ArXiv* (October 2022).
[30] Ibid.

access to other developers who wish to adapt a system to a specific context, or providing open-source access to a full model. This can give rise to an entire new category of risks based on accidents, irresponsible deployment and malicious uses. Moreover, if some of the risk pathways in the development and training phase have not been duly accounted for, those pathways can multiply and amplify one another in the deployment and usage phase. For example, if dangerous capabilities such as the ability to deceive users have not been addressed, this makes it more likely that AI systems will be misused by malicious actors.

Accidents and irresponsible deployment: In the context of safety-critical infrastructure or policy areas with considerable impacts on human wellbeing, the risk of accidents as a result of deploying AI systems are especially pressing. Accidents could be caused by failures of robustness (systems receiving inputs that cause them to malfunction), specification (systems aiming to achieve goals subtly different from the developer's intentions), or assurance (systems that cannot be adequately monitored or controlled during operation).[31] The complexity of ensuring full testing across each of these domains can be vast, meaning that accidents may be more difficult to foresee than malicious uses. Safety-critical sectors where AI is already being deployed, and where accidents could occur include healthcare, transportation, defence, and energy production.

Example: Transportation and traffic control

As autonomous vehicles are gradually adopted, local governments are also considering the benefits of using AI systems embedded in stoplights, roads, and drones to upgrade traffic control systems. Such systems could collect real-time congestion information, re-route traffic, and anticipate future congestion or weather incidents.[32] However, embedding AI into large-scale critical systems creates ample opportunities for unanticipated interactions to occur that could result in accidents.[33] For instance, one can imagine traffic monitoring systems unable to properly adjust traffic lights at a complex intersection during a time of high congestion, or in a scenario where extreme weather and novel vehicular shapes are present.

Malicious uses: Prior to deploying new systems, designers need to consider ways in which AI models can be repurposed for malicious intentions. These other functionalities may include augmenting offensive cyber capabilities, creating or acquiring destructive

[31] Zachary Arnold and Helen Toner, "AI Accidents: An Emerging Threat," *CSET Policy Brief* (July 2021).

[32] Elizabeth Mynatt et al., "A national research agenda for intelligent infrastructure," *Computing Community Consortium* (May 2017).

[33] Phil Laplante et al., "Artificial Intelligence and Critical Systems: From Hype to Reality," *Computer* 53, no. 11 (November 2020); Phil Laplante, ""Smarter" Roads and Highways," *IEEE Internet of Things Magazine* 1, no. 2 (December 2018).

weaponry, or exerting control and surveillance over populations.[34] The risks emerging from malicious uses of AI can be separated into risks to digital security, political security and physical security.[35]

Example: Creation of bioweaponry

Particularly hazardous misuse risks exist at the intersection of AI systems and biotechnology, where systems could give malicious actors a new entry point to committing harmful acts.[36] For example, models intended for pharmaceutical drug discovery could be repurposed to design dangerous toxins[37] and deep learning can be used to develop "precision maladies" that target specific populations.[38] AI systems can be used not only to design, but also to build and test bioweaponry. Increasingly popular AI-driven "cloud labs" for automating biological scientific processes could be repurposed to manufacture and test pathogens at scale.[39]

Tolerance for error: AI systems that are trained in the commercial context are likely to have different criteria compared to high-stakes public sector decision-making contexts, and different thresholds or tolerances for error or mistruth.[40] For example, in the context of a commercial LLM, the imperfect nature of the training process leading to an incorrect output is likely to be of low consequence on its own, but the cumulative harm of many mistakes over time is hard to quantify.

[34] Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," *ArXiv* (February 2018).

[35] Alexander Babuta, Marion Oswald and Ardi Janjeva, "Artificial Intelligence and UK National Security: Policy Considerations," *RUSI Paper* (April 2020).

[36] John T. O'Brien and Cassidy Nelson, "Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology," *Health Security* 18, no. 3 (June 2020).

[37] Fabio Urbina et al., "Dual use of artificial-intelligence-powered drug discovery," *Nature Machine Intelligence* 4, no. 189-191 (March 2022).

[38] John T. O'Brien and Cassidy Nelson, "Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology," *Health Security* 18, no. 3 (June 2020).

[39] Ibid.

[40] Michael Veale, Max Van Kleek and Reuben Binns, "Fairness and Accountability Design Needs for Algorithmic Support in High Stakes Public Sector Decision-making," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (April 2018).

Example: Behavioural analytics

For a company seeking to deliver an online advertisement, the accuracy of the recommender algorithm does not need to be very high; if a user ends up 5-10% more likely to purchase a product, then a cheap targeted advertisement will be worth serving. The worst-case scenario in this context is that a user could be shown a misdirected advertisement that they ignore. If a similar algorithm is used to allocate public goods and services, there is a significant risk of an individual or group of people being deprioritised or denied the social care they need. One example where these concerns have been raised is the UK's Department for Work and Pensions deploying machine learning to analyse historical benefits data to predict how likely a new Universal Credit claim is to be fraudulent or incorrect.[41]

1.3 Longer-term deployment and diffusion

As AI becomes increasingly embedded across society, it may cause harm in more diffuse and structural ways – not because a specific AI system is misused or fails, but because it changes incentives, options, or power dynamics in important parts of society.

For example, continued development of AI could cause power dynamics between competing global actors to shift substantially, and AI could entrench bias or disseminate disinformation at scale, resulting in serious impacts on the long-term legitimacy of democratic systems.[42] On a practical level, it could play a pivotal role in areas like corporate strategy, weapons development and even foreign policy. Dominance in AI will accrue commercial and political power, which in turn will further reinforce dominance in AI and related technology fields. Furthermore, commonplace delegation of decisions to AI systems may introduce system vulnerability to extreme conditions to which the systems are not robust, and may erode or atrophy human judgement that is rarely needed except in emergencies.

The range of possible structural impacts is wide, but five are particularly important: economic and employment impacts, impacts on discrimination and inequality, impacts on

[41] Paul Seddon, "Universal Credit: Warnings over AI use to risk-score benefit claims," *BBC News,* 11 July, 2023, https://www.bbc.co.uk/news/uk-politics-66133665.

[42] Steven Feldstein, "The Global Expansion of AI Surveillance," *Carnegie Endowment for International Peace Paper* (September 2019), https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847.

epistemic processes and freedom of thought, impacts on geopolitical stability, and the concentration of wealth and power.

Economy and employment: OpenAI's mission statement refers to 'developing highly autonomous systems that outperform humans at most economically valuable work.'[43] The responsibility for preparing for these economic impacts will fall on governments, so it is imperative for policymakers to plan well in advance to foster an environment where workers are empowered by the benefits of AI systems rather than displaced by them. Without targeted policy interventions, the economic impacts of AI could include an acceleration of wealth amongst those who own and control AI systems, which in turn could exacerbate economic inequality.[44]

Example: Impact on precarious workers

To ensure that LLMs provide responses that are aligned with the views of their developers, companies have employed some training based on human feedback which reduces the prevalence of harmful or toxic data in the initial data scraping phase. However, various investigations have shown how this approach can exacerbate exploitative labour practices against the individuals labelling training datasets (often containing very graphic content), and lead to significant yet unseen harms.[45] Moreover, because this work can be easily outsourced, it is often conducted by people in poorer countries with no minimum wage requirements or easily accessible mental welfare provisions. These dynamics place the current glamour, hype and money associated with the AI industry into its wider human context.

Bias and discriminatory impact: Many AI systems have been trained on datasets which reflect the inequitable state of the world and, consequently, can perpetuate and entrench inequities much further into the future. Bias in AI systems can also emerge from sources other than biased training data, for instance if models are deployed on target data whose distribution varies significantly from the training or test data. This is especially concerning in areas of social policy with which most people have some form of interaction over the course of their lives, such as education, health, finance, policing and criminal justice. As well as

[43] "About," Open AI, accessed 28 July 2023, https://openai.com/about.

[44] Aaron Smith and Janna Anderson, "AI, Robotics, and the Future of Jobs," *Pew Research Center Report* (August 2014), https://www.pewresearch.org/internet/2014/08/06/future-of-jobs/.

[45] Billy Perrigo, "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic," *TIME,* 18 January, 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/.

being a source of harm today, bias in AI systems could act as an accelerant that deepens existing inequities and makes them intractable.

Example: Policing and healthcare

Machine learning models trained on police data may replicate and amplify existing biases within training datasets, such as over- or under-policing of certain communities or racial profiling.[46] Individuals from particular socioeconomic or ethnic backgrounds are likely to engage with public services more frequently, meaning that data on these individuals is more readily accessible. This over-representation may then result in higher risk calculations if a model is being used to make a prediction about an individual's future behaviour, such as likelihood to offend.[47]

While the above is an example of overrepresentation of certain communities in sensitive datasets, healthcare offers an example where *underrepresentation* is the primary concern. Given that medical research has historically disproportionately focused on white people, AI systems can be less effective at identifying illness in under-served patient populations.[48] When layered on top of inequitable distribution of healthcare resources, this leads to the underrepresentation of these communities in data about the spread of viruses like COVID-19 and associated mortality rates.

Erosion of epistemic security and freedom of thought: Epistemic security refers to the ability of societies to take informed collective action based on reliable information, in an environment where adversaries seek to undermine informed debate.[49] Information technologies challenge epistemic security in at least four ways: facilitating the spread of misinformation and disinformation; drawing attention away from key issues; allowing the formation and persistence of echo chambers with poor epistemic norms; and allowing bad actors to fake the hallmarks of trustworthy information sources.[50] AI technologies, including recommender systems and generative AI, could exacerbate these challenges, leading to political polarisation and leaving democratic societies unable to sustain informed

[46] Alexander Harris, Eleanor S, Emma Bradford and Ardi Janjeva, "Behavioural Analytics and UK National Security," *CETaS Research Reports* (March 2023).

[47] Alexander Babuta and Marion Oswald, "Data Analytics and Algorithmic Bias in Policing," *RUSI Briefing Paper* (September 2019): 12.

[48] Laleh Seyyed-Kalantari et al., "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature Medicine* 27, no. 2176-2182 (December 2021).

[49] Elizabeth Seger et al., "Tackling threats to informed decision-making in democratic societies," *Alan Turing Institute Paper* (October 2022): 2.

[50] Ibid.

electorates, while giving authoritarian regimes greater tools of control and suppression.[51]

Example: 2024 US presidential campaigns

Generative AI has been used for electoral campaigning during the early phases of the 2024 U.S. presidential campaign. The Republican National Committee have shared AI-generated images of non-existent and dystopian scenarios in their campaign advertising of "An AI-generated look into the country's possible future if Joe Biden is re-elected in 2024," including an attack on Taiwan, as well as boarded-up shops and the imposition of martial law in American cities.[52] Meanwhile the team of Republican candidate Ron DeSantis has circulated media which features real images of rival candidate Donald Trump with Dr. Anthony Fauci interspersed with crude AI-generated images.[53] This highlights a particular risk regarding the merging of real and fake content within a single image, while demonstrating that the threat posed by generative AI to electoral processes is not strictly limited to foreign interference, as evidenced by its adoption by those directly engaged in political campaigns.

Geopolitical instability: The potential for advanced forms of AI to confer strategic advantage on the 'winners' of R&D races creates powerful incentives for international competition.[54] Traditional arms-race logic would dictate that political, commercial, and military actors perceive the stakes of 'winning' and 'losing' to be very high, encouraging the development of more capable systems ahead of rivals and pre-emptive policies which could trigger a downward escalatory spiral. Arms-racing dynamics are directly at odds with AI safety: as systemic competition drives capability development through reprioritisation of investment, actors may take larger risks in hopes of higher payoffs, or in anticipation of what could happen if rivals move first. These international dynamics are inextricably linked to the changing regulatory environment, with the most stringent forms of regulation perceived by some to be an inherent risk when compared to the more permissive environment created by

[51] Josh A. Goldstein et al., "Forecasting potential misuses of language models for disinformation campaigns – and how to reduce risk," *Stanford Internet Observatory Cyber Policy Centre* (January 2023), https://cyber.fsi.stanford.edu/io/news/forecasting-potential-misuses-language-models-disinformation-campaigns-and-how-reduce-risk; Katerina Sedova et al., "AI and the Future of Disinformation Campaigns Part 2: A Threat Model," *CSET Paper* (December 2021).

[52] David Klepper and Ali Swenson, "AI-generated disinformation poses threat of misleading voters in 2024 election," *PBS News Hour,* 14 May, 2023, https://www.pbs.org/newshour/politics/ai-generated-disinformation-poses-threat-of-misleading-voters-in-2024-election.

[53] Nicholas Nehamas "DeSantis Campaign Uses 'Deepfake' Images to Attack Trump on Twitter," 8 June, 2023, https://www.nytimes.com/2023/06/08/us/politics/desantis-deepfakes-trump-fauci.html.

[54] Eric Schmidt, "AI, Great Power Competition & National Security," *Daedalus* 151, no. 2 (Spring 2022): 288-298

perceived rivals and adversaries. For example, although China's recent proposals to regulate generative AI set stringent conditions on private-sector actors developing and deploying generative AI systems, there is no mention of limitations on the Government's use of AI.[55]

Example: Escalating US-China tensions

The concentration of AI capability and investment within two main political blocs – dominated by the US and China – raises a host of issues. For instance, without a meaningful consensus regarding the importance of safe and ethical AI, careless behaviour on the part of one AI leader might be replicated around the world. Crucially, in the military and intelligence context, decision-making informed by AI is likely to operate on much shorter timeframes compared to those which global leaders enjoyed during the Cold War. This emphasises the importance of clear routes for de-escalation to prevent scenarios where AI fosters an inherently more adversarial approach to defence and security strategy.

Concentration of wealth and power: Power and wealth is increasingly concentrating in the hands of the owners and controllers of the most successful AI systems. These actors tend to be profit-driven, function like monopolies and crowd out competitors who may not have the means to scale potentially significant innovations.[56] For countries like the UK operating in a fiscally challenging environment, there is a distinct disadvantage in their ability to finance sector-defining AI projects and attract the expertise that currently resides in the private sector. This contributes to an imbalance of power when implementing regulation and is arguably why governments find it easier to welcome voluntary commitments, of the kind that several US-based companies agreed to in July this year.[57]

[55] Matt O'Shaughnessy, "What a Chinese Regulation Proposal Reveals About AI and Democratic Values," *Carnegie Endowment for International Peace Commentary,* 16 May, 2023, https://carnegieendowment.org/2023/05/16/what-chinese-regulation-proposal-reveals-about-ai-and-democratic-values-pub-89766.

[56] Melissa Heikkilä, "Generative AI risks concentrating Big Tech's power. Here's how to stop it," *MIT Technology Review,* 18 April, 2023, https://www.technologyreview.com/2023/04/18/1071727/generative-ai-risks-concentrating-big-techs-power-heres-how-to-stop-it/.

[57] Kari Paul, Johana Bhuiyan and Dominic Rushe, "Top tech firms commit to AI safeguards amid fears over pace of change," *The Guardian,* 21 July, 2023, https://www.theguardian.com/technology/2023/jul/21/ai-ethics-guidelines-google-meta-amazon.

These leading actors benefit from a snowball effect which reinforces their position as gatekeepers, as they are able to:

1. Attain and deploy large sums of initial investment.
2. Establish state-of-the-art compute infrastructure.
3. Create the conditions to accumulate and store the most data.
4. Build the most state-of-the-art models.
5. Attract the most users and collect further data on them to improve model performance.
6. Reinforce the cycle.

Example: Increasing dominance of industry in AI research

In recent years, the high costs required to advance frontier research in deep learning have tilted dominance in the field away from academia and towards industry. While many research institutes are applying AI systems in innovative ways, massive corporations possess the capital to develop the cutting-edge tools in the first place and therefore set the 'rules of the game'.[58] On current trends there will likely be few alternatives to a small set of leading commercial AI systems. Although there appear to be concerns within companies like Google regarding the prospect of open-source AI challenging their dominant position[59] – as seen with high-performing models like Falcon-40B-Instruct trained in the UAE and subsequently open sourced[60] – they have undoubtedly established a significant head-start, and there are few limits to the resources they can commit to widening the gap again if it were to close.

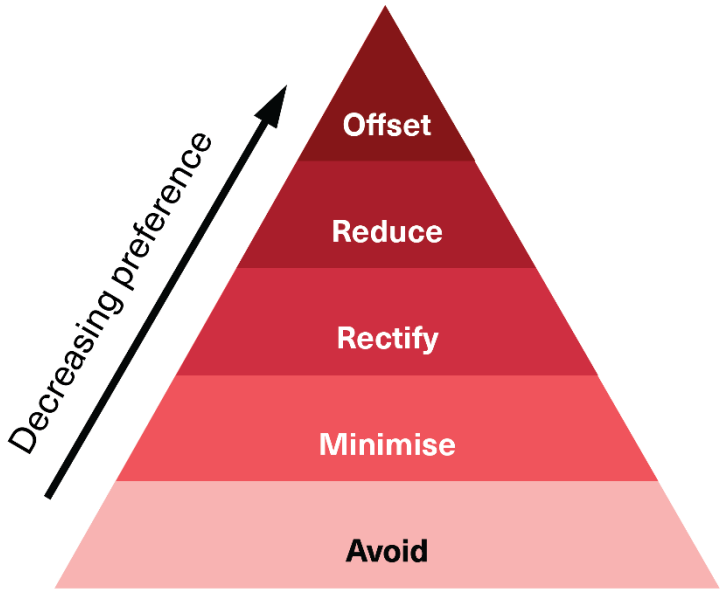
Public initiatives such as the proposed British Exascale compute infrastructure or the proposed American National AI Research Resource could provide a counterweight to these developments, but would require significant funding, protected from fluctuations in the fiscal environment, in order to provide a serious alternative to the market.[61]

[58] Nur Ahmed, Muntasir Wahed and Neil C. Thompson, "The growing influence of industry in AI research," *Science* 379, no. 6635 (March, 2023): 884-886, https://doi.org/10.1126/science.ade2420.
[59] Ibid.
[60] Luis Roque, "Harnessing the Falcon 40B Model, the Most Powerful Open-Source LLM," Towards Data Science, last modified 9 June, 2023, https://towardsdatascience.com/harnessing-the-falcon-40b-model-the-most-powerful-open-source-llm-f70010bc8a10.
[61] "The National Artificial Intelligence Research Resource Task Force (NAIRRTF)," National Artificial Intelligence Initiative, n.d., https://www.ai.gov/nairrtf/; HM Government, *Independent Review of The Future of Compute: Final report and recommendations* (Department for Science, Innovation and Technology: 2023), https://www.gov.uk/government/publications/future-of-compute-review/the-future-of-compute-report-of-the-review-of-independent-panel-of-experts#chap4.

2. Achieving Resilience in the Domestic AI Policy Landscape

The risk pathways outlined above raise important questions for policymakers seeking to make sense of the AI landscape. Considering this range of pathways, the primary goal for policymakers should be to prevent associated harms from materialising as early as possible, while bolstering resilience to minimise damage in the event that AI does inflict harm.

Figure 2. An illustration of how risk mitigation hierarchies can be used to prioritise different strategies to address harmful impacts.[62]



Pyramid shows a range of approaches to acting against a form of risk.

Policy interventions will be most effective if they intervene at the point in the lifecycle where risk first arises. While many risks will be sector and context-specific, we are seeing increasing potential and use of general-purpose systems, where risky features of these systems (emergent capabilities, bias, lack of interpretability, surprising failures) contribute to downstream harms. This demonstrates a need for concurrent policy action to address novel risks from such "frontier" models, alongside more familiar risks from established and use-

[62] Adapted from Saurab Babu, "Mitigation Hierarchy: Levels of mitigation in Environmental Impact Assessment," *Eco-Intelligent,* 11 December, 2016, https://eco-intelligent.com/2016/12/11/levels-of-mitigation-in-environmental-impact-assessment/. Distinct variations of a mitigation hierarchy have also been applied to algorithmic impact assessment. See: David Leslie et al., "Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal," *ArXiv* (February 2022), https://doi.org/10.5281/zenodo.5981676.

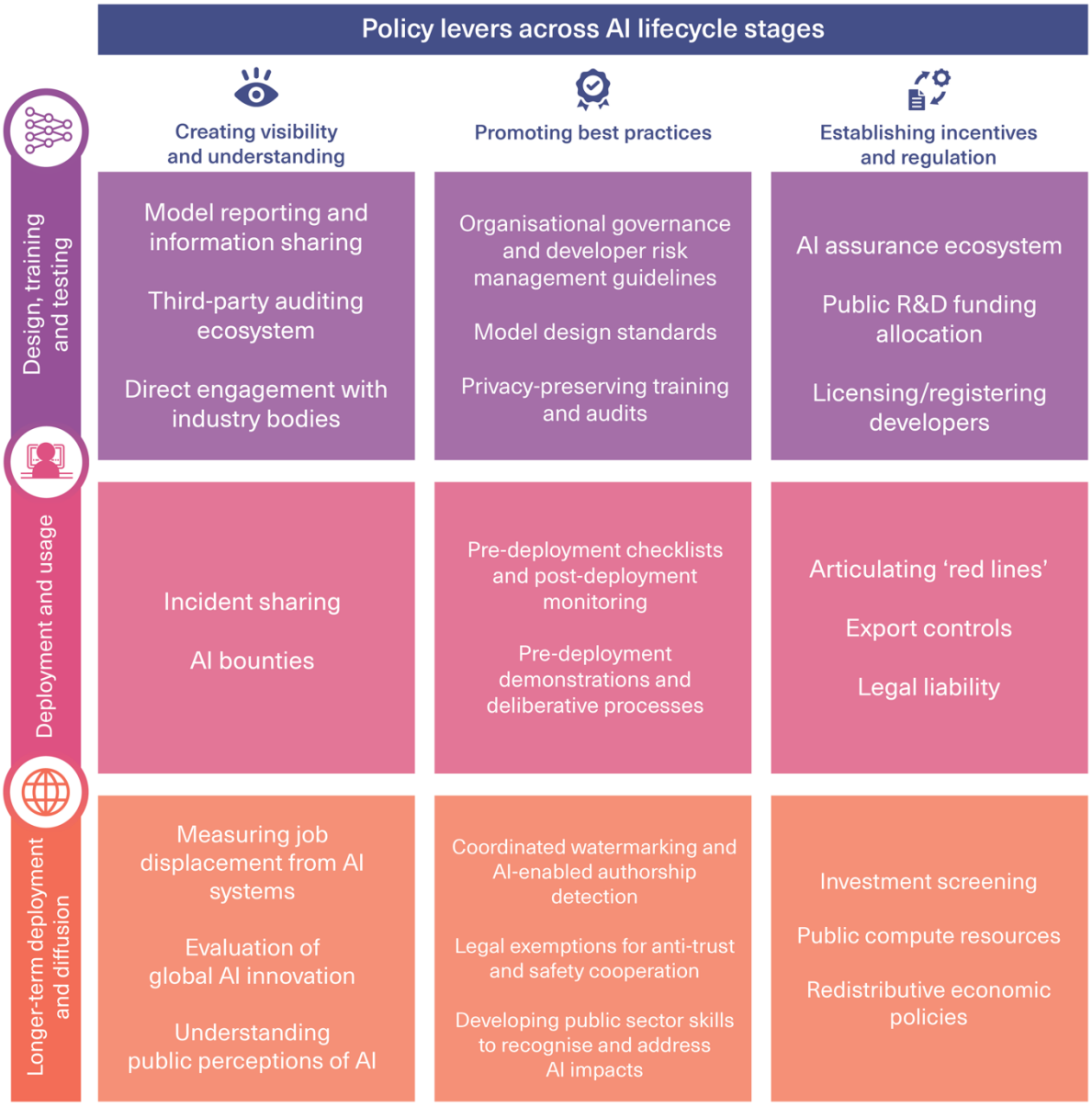
case specific machine learning models. To this end, this section describes three main categories of AI policy interventions, and maps these to the key stages of development and training, deployment and usage, and long-term effects of deployment.

1) **Creating visibility and understanding**: Managing risks effectively requires visibility into technical progress, and where risks arise as part of that progress. Most of this information remains contained within industry, so enhanced transparency and reporting is essential from those at the forefront of AI research and development. Incentives for model developers to provide visibility are often limited, so policy interventions will be crucial in achieving this, albeit in ways that do not unduly restrict innovation.

2) **Promoting best practices**: Policymakers must work in collaboration with AI system designers, academics and impacted groups to build consensus regarding best practices in the development and deployment of AI systems.

3) **Establishing incentives and enforcing regulation**: Incentives and enforcement regimes are required to ensure the adoption of these best practices. Incentives may include allocating public resources and funding for initiatives that increase understanding and adoption of safety measures, while enforcement may involve legislative rules and penalties to mandate adherence to best practices.

Figure 3. Policy levers across AI lifecycle stages.



These three categories are intrinsically complementary: maximising visibility and understanding regarding the capabilities and impacts of AI systems enables policymakers to promote best practices with a greater level of authority, ultimately meaning that they have a clear basis upon which to introduce incentives for developers and a stronger mechanism with which to hold developers accountable.

2.1 Creating visibility and understanding

Policymakers, regulators and oversight bodies require visibility of how AI systems are developed and deployed, where sources of risk and gaps in risk management exist, and (if these risks manifest into harms) how harms are felt by various sections of the public. Attaining a holistic understanding of AI capabilities and potentially dangerous tipping points is preferable to relying on proxy measures such as model parameters or financial investments, particularly if the AI models of the future are better at learning with less data and investment.

The table below describes a series of policy options which would serve the objective of creating better visibility and understanding of AI systems amongst policymakers and regulators, mapping them against the lifecycle phases outlined in the previous section.

Lifecycle phase	Policy lever 1	Policy lever 2	Policy lever 3
Development and training	**Model reporting and information sharing.** An anticipatory approach to AI policy may involve creating an information-sharing regime between AI system developers and government bodies.[63] In the UK, the value of information-sharing is evident in domains such as cybersecurity.[64] Relevant categories of sharing would include models' intended functionality, levels of compute usage during	**Third-party auditing ecosystem.** This is valuable for building trust regarding the robustness of risk management practices. For frontier AI capabilities, auditing based on high-risk capability evaluations may be particularly valuable for assessing accident, misuse, and emergent capability risks	**Direct engagement with industry bodies.** New industry bodies, such as the Frontier Model Forum, are being set up by leading AI companies in an attempt to oversee safe development of the most advanced models.[69] The extent to which government involvement in such bodies will be promoted is as yet unclear, but there should be a consistent approach across UK Government to engagement with these bodies, featuring

[63] Nikhil Mulani and Jess Whittlestone, "Proposing a Foundation Model Information-Sharing Regime for the UK," GovAI Blog, last modified 16 June, 2023, https://www.governance.ai/post/proposing-a-foundation-model-information-sharing-regime-for-the-uk.

[64] "About CISP (Connect Inform Share Project," National Cyber Security Centre, n.d., https://www.ncsc.gov.uk/cisp/home.

[69] Dan Milmo, "Google, Microsoft, OpenAI and startup form body to regulate AI development," *The Guardian*, 26 July, 2023, https://www.theguardian.com/technology/2023/jul/26/google-microsoft-openai-anthropic-ai-frontier-model-forum.

	training, evaluation against performance benchmarks, and details about training datasets. This information could be summarised in model cards released in conjunction with new models,[65] and collated on a centralised model register where key decision-makers are able to access and review model details, thus making informed decisions on their utility and trustworthiness.[66]	ahead of deployment.[67] A multi-layered approach to AI auditing is advisable so that the model itself is scrutinised in addition to the governance and application procedures.[68] Policymakers should encourage future growth in the auditing ecosystem, to include the possible need for accreditation of auditors as trustworthy.	strong representation from the national security community. Engagement with industry bodies should be complemented by engagement with workers' groups representing the interests of those with less power within these organisations, for example unions.
Deployment and usage	**Incident sharing**. A more systematic approach to collecting and analysing risk incidents – whether accidental or malicious – could illuminate crucial patterns that have long-term policy implications. The AI Incident Database is an existing third sector example of such a collection mechanism.[70]	**AI bounties.** Modelled after bug bounties in software security, governments can clarify legality and support industry bounty programmes that incentivise external researchers to identify and responsibly disclose risks of AI systems, including bias, unexpected behaviours, jailbreaks, and adversarial inputs.[71]	

[65] Margaret Mitchell et al., "Model Cards for Model Rerporting," *ArXiv* (October 2018), https://doi.org/10.48550/ArXiv.1810.03993.

[66] "gchq / Bailo," GitHub, n.d., https://github.com/gchq/Bailo.

[67] "An early warning system of novel AI risks," Technical Blog, Google DeepMind, last modified 25 May, 2023, https://www.deepmind.com/blog/an-early-warning-system-for-novel-ai-risks.

[68] Jakob Mökander et al., "Auditing large language models: a three-layered approach," *ArXiv* (February 2023), https://doi.org/10.48550/ArXiv.2302.08500.

[70] "Welcome to the AI Incident Database," AI Incident Database, n.d., https://incidentdatabase.ai.

[71] Miles Brundage et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," *ArXiv* (April 2020), https://doi.org/10.48550/ArXiv.2004.07213; "The Crash Project (formerly the Algorithmic Vulnerability Bounty Project)," AIJ, n.d., https://www.ajl.org/crash-project; Kyra Yee and Irene Font Peradejordi, "Sharing learnings from the first algorithmic bias bounty challenge," Insights, Twitter Blog, last modified 7 September, 2021, https://blog.twitter.com/engineering/en_us/topics/insights/2021/learnings-from-the-first-algorithmic-bias-bounty-challenge.

Longer-term deployment and diffusion	**Measuring job displacement from AI systems.** Having a real-time, economy-wide picture of the areas of the labour market approaching an inflection point for automation will be vital to an anticipatory policy approach.[72] It will be important for policymakers to be aware of both job displacement and possible erosion of quality of work.[73]	**Evaluation of global AI innovation**. Policymakers need a rigorous mechanism for understanding the global pace of change in AI and its ramifications for the geopolitical landscape. Utilising bibliometric analysis and the uplift in open-source intelligence capabilities announced in the Integrated Review Refresh will be integral to this effort.	**Understanding public perceptions of AI**. As well as making developments in AI capability visible to policymakers, public attitudes towards AI should be tracked. A recent Alan Turing Institute-Ada Lovelace Institute collaboration exemplified how this could be done across the public at large.[74] Beyond this, policymakers should pay particular attention to incorporating the views of those most harmed by the rollout of AI. This might include workers in the platform economy whose employment conditions have been worsened by algorithmic decision-making,[75] creative professionals whose work has been used to power generative AI tools such as DALL-E and ChatGPT,[76] and school students who have been impacted by discriminatory algorithmic marking of exams.[77]

[72] PwC, *The Potential Impact of Artificial Intelligence on UK Employment and the Demand for Skills*, Research Report no. 2021/042 (Department for Business, Energy and Industrial Strategy: 2021), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1023590/impact-of-ai-on-jobs.pdf.

[73] "The Good Work Charter," Institute for the Future of Work, last modified 18 October, 2028, https://www.ifow.org/publications/the-ifow-good-work-charter.

[74] The Alan Turing Institute and the Ada Lovelace Institute, *How do people feel about AI? A nationally representative survey of the British public* (2023), https://www.turing.ac.uk/news/publications/how-do-people-feel-about-ai.

[75] Zane Muller, "Algorithmic Harms to Workers in the Platform Economy: The Case of Uber," *Columbia Journal of Law and Social Problems* 53, no. 2 (2020): 167-210.

[76] Alain Strowel, "ChatGPT and Generative AI Tools: Theft of Intellectual Labor?," *IIC* 54 (2023): 491-494, https://doi.org/10.1007/s40319-023-01321-y.

[77] Daan Kolman, "F**k the algorithm"?: What the world can learn from the UK's A-level grading fiasco," LSE Blog, last modified 26 August, 2023, https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/.

2.2 Promoting best practices

As policymakers benefit from better visibility and understanding of the risks and opportunities presented by AI, they will be in a better position to identify and build upon best practice guidelines for the development, use and governance of AI. At present, many of these best practices are developed within industry or by academia. Concerns have been raised over AI risk frameworks which either promote industry interests or are too embedded in academic theory to be practically useful. In this context, government can play an essential role in helping to build consensus across the AI ecosystem nationally and internationally, such that best practices are identified across the full range of sectors and promoted in a consistent manner.

Lifecycle phase	Policy lever 1	Policy lever 2	Policy lever 3
Development and training	**Organisational governance and developer risk management guidelines**. NIST's AI Risk Management Framework Playbook addresses legal compliance, oversight and systems testing[78] – an equivalent of this is required in the UK.	**Model design standards.** Clearer guidelines on the necessary tests and evaluations, and the results expected to verify that models meet a safety threshold, would be particularly useful for the developer community. Ethical and professional standards should be included alongside technical requirements.[79]	**Privacy-preserving training and audits.** To address data privacy concerns, technical approaches such as federated learning could enable more privacy-preserving model training.
Deployment and usage	**Pre-deployment checklists and post-deployment monitoring.** In domains involving a higher risk of accidents or misuse, or where systems are required to have novel characteristics and may behave	**Pre-deployment demonstrations and deliberative processes.** For technologies that present transformative and disruptive potential, input from a broad range of perspectives may help flag societal concerns	

[78] "Govern," NIST Trustworthy & Responsible AI Resource Center, n.d., https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook/Govern#Govern%201.2.

[79] "Different types of standards," Standards at a glance, AI Standards Hub, n.d., https://aistandardshub.org/resource/main-training-page-example/2-different-types-of-standards/.

	unpredictably, best practices for deployment are particularly important. This may involve pre-deployment checklists and post-deployment monitoring requirements for frontier AI systems, and clearer communication regarding the security risks of open-sourcing models of a certain capability level.	before deployment, guide deployment choices, and design anticipatory governance and mitigations. Measures such as public demonstrations and citizen assemblies can draw on broad societal perspectives and lead to better, and more legitimate, governance outcomes.[80]	
Longer-term deployment and diffusion	**Coordinated watermarking and AI-enabled authorship detection.** A society-wide deterioration in the ability to produce, distribute, acquire and access reliable information is a major concern. A coordinated approach to watermarking of AI-generated content and AI-enabled authorship detection are two key challenges which, if solved, could be pivotal to addressing the pollution in public discourse. Although the nature of the information ecosystem makes this a global endeavour, the UK could fund small-scale pilot projects to demonstrate proofs of concept.	**Legal exemptions for anti-trust and safety cooperation**. Concerns over anti-trust regulation may prevent AI developers from sharing knowledge or coordinating best practices that could improve AI safety.[81] Policymakers could explore legal exemptions that allow for safety-motivated collaboration between companies building AI systems, thereby reducing the chances of system flaws going unnoticed and unaddressed.	**Developing public sector skills to recognise and address AI impacts**. Officials in the public sector need to be upskilled, both to use AI tools to efficiently deliver public services, and to improve understanding of the accompanying risks. With regard to LLMs, their outputs are heavily dependent on the formatting of the user's prompt or input, so training a user how to frame questions in the right way is essential. Additional private, academic, and civil society partnerships can address persistent skills gaps.[82]

[80] Elizabeth Seger at al., "Democratising AI: Multiple Meanings, Goals, and Methods," *ArXiv* (March 2023), https://doi.org/10.48550/ArXiv.2303.12642.

[81] Forthcoming GovAI report (Alaga and Schuett, 2023).

[82] Slava Jankin Mikhaylov, Marc Esteve and Averill Campion, "Artificial intelligence for the public sector: opportunities and challenges of cross-sector collaboration," *Phil. Trans. R. Soc. A*., 2128, no. 376 (August 2018) https://doi.org/10.1098/rsta.2017.0357.

2.3 Establishing incentives and enforcing regulation

Due to the vastly different incentive structures of industry, academia and civil society, promoting best practices for the safe development and deployment of AI systems will often be insufficient on its own. To go further, government can encourage adherence to best practices and hold those developing and deploying systems accountable through a combination of soft incentives and enforced regulation. Soft incentives might include actions such as government-supported auditing, or public availability of resources such as funding and compute for socially beneficial research. Legal enforcement may involve many components including liabilities, certifications, or licensing schemes for developers and deployers of high-risk AI systems. What levers are appropriate will depend on the level of risk being addressed and the extent to which existing incentives seem sufficient to lead to safe and responsible behaviour.

Lifecycle phase	Policy lever 1	Policy lever 2	Policy lever 3
Development and training	**AI assurance ecosystem**. AI assurance is about "building confidence or trust" in AI systems.[83] Third party audits, AI standards and risk assessments can help to assess the properties of an AI system against a range of technical and ethical criteria. Much progress has been made towards compiling best practice in AI assurance by the OECD[84] and Centre for	**Public R&D funding allocation**. A step-change in the way that AI is integrated into the UK economy calls for a recalibration in the way that existing research funding vehicles allocate resources. This involves scaling up explicit pathways for research which focus on bias-reducing, privacy-protecting and safety-improving approaches to	**Licensing/registering developers**. Government-administrated licensing regimes could be part of an enforceable legal framework for filtering out the most ethically dubious AI use cases at an earlier stage.[87] Licensing could be defined across several dimensions including compute thresholds, capability evaluations, algorithm design, and intended use-cases. A preliminary step to a licensing regime – which could take many years to develop and implement

[83] Centre for Data Ethics and Innovation, *The roadmap to an effective AI assurance ecosystem* (UK Government: 2021) https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem.

[84] "Catalogue of Tools & Metrics for Trustworthy AI," Tools & Metrics, About the catalogue, OECD.AI, n.d., https://oecd.ai/en/catalogue/faq.

[87] Teralyn Whipple, "Experts debate artificial intelligence licensing legislation," *Broadband Breakfast*, last modified 23 May, 2023, https://broadbandbreakfast.com/2023/05/experts-debate-artificial-intelligence-licensing-legislation/.

	Data Ethics and Innovation.[85] A robust AI assurance ecosystem may require public sector procurement strategies mandating thorough testing of AI systems, independent oversight of industry assurance practices and funding to improve existing tools and metrics for trustworthy AI.	AI.[86] Additional streams of funding could be dedicated to foundational measurement theory for AI systems, which is likely to be necessary for creating accurate technical standards in this domain.	– could be a registration process, where the UK Government gathers basic information about who is training the most sophisticated LLMs and whether there is substantial risk that their use might violate export controls or other laws. Analogously, corporations are subject to registration requirements which give people and businesses confidence in the integrity of financial transactions, as are companies handling nuclear materials for civilian purposes and labs handling dangerous biological or chemical materials.[88]
Deployment and usage	**Articulating 'red lines'.** There may be specific contexts where the integration of AI into decision-making functions will be undesirable *for the foreseeable future*. This is particularly true of 'autonomous agents' – systems which can generate a sequence of tasks that a model works on until the desired 'goal' is reached. Similarly, there may be red lines beyond which systems must not take an irreversible	**Export controls**. These are tools of economic statecraft which could be used to limit who is able to purchase AI software, and inputs for developing AI systems such as advanced chips, developed in the UK. The UK Government has shown an appetite for this already when working with the US Government on blocking China's access to high-	**Legal liability**. Laws regarding remedies for injuries, damages or harms caused by AI systems could have an important role to play in incentivising AI developers (those actors with the most information about AI systems) to weigh potential societal harms against the desire for continued innovation. The EU's AI Liability Directive, for example, aims to create a rebuttable 'presumption of causality, to ease the burden of proof for victims to establish

[85] HM Government, *CDEI portfolio of AI assurance techniques* (Centre for Data Ethics and Innovation and Department for Science, Innovation and Technology: 2023), https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques.

[86] See for example: https://www.nsf.gov/pubs/2023/nsf23562/nsf23562.htm.

[88] Gillian Hadfield, Mariano-Florentino (Tino) Cuéllar and Tim O'Reilly, "It's Time to Create a National Registry for Large AI Models," Carnegie Endowment, Commentary, last modified 12 July, 2023, https://carnegieendowment.org/2023/07/12/it-s-time-to-create-national-registry-for-large-ai-models-pub-90180.

	action without direct human oversight or authorisation. Formalising these in policy documents and/or codes of practice will be important.	performance chip designs produced by Arm.[89]	damage caused by an AI system.[90]
Longer-term deployment and diffusion	**Investment screening**. This is a tool designed to limit foreign influence over the trajectory of AI development in specific, often security-related contexts. Already, foreign-led AI and data infrastructure investments above a certain size are subject to mandatory review by the Government's Investment Security Unit.[91] Blocked deals have so far included Chinese-owned companies attempting to acquire U.K.-based companies working on robotic vision-sensing technology, semiconductor chip design software, and semiconductor manufacturing technology.[92]	**Public compute resources**. Investment in such infrastructure could substantially lower the costs of training, testing and evaluating AI models, thereby reducing the prospect of the most powerful models being heavily concentrated amongst a few companies. To minimise the risk of unsafe development, this would need significant guardrails, such as tiered access and model review requirements.[93]	**Redistributive economic policies**. If advanced forms of AI pave the way for more widespread displacement rather than augmentation of labour, more serious consideration could be given to 'windfall clauses' on companies with the greatest AI market share.[94]

[89] "U.S., UK export controls hit China's access to Arm's chip designs -FT," *Reuters*, last modified 14 December, 2022, https://www.reuters.com/technology/export-controls-hit-chinas-access-arms-chip-designs-ft-2022-12-14/.

[90] European Parliamentary Research Service Briefing, "Artificial intelligence liability directive," 10 February 2023, https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf.

[91] Department for Business, Energy & Industrial Strategy, "National Security and Investment report shows new system is working," Press release, last modified 16 June, 2022, https://www.gov.uk/government/news/national-security-and-investment-report-shows-new-system-is-working.

[92] Debevoise & Plimpton, "UK Government Prohibits Third Deal Under Its New National Security and Investment Act," *Debevoise Update*, last modified 21 November, 2022, https://www.debevoise.com/insights/publications/2022/11/uk-government-prohibits-third-deal.

[93] Lennart Heim and Markus Anderljung, "Comments on the interim report of the National Artificial Intelligence Research Resource Task," Submission to the Request for Information (RFI) on Implementing Initial Findings and Recommendations of the NAIRR Task Force, Centre for the Governance of AI, last modified 30 June, 2022, https://www.governance.ai/research-paper/submission-nairr-task-force.

[94] Cullen O'Keefe et al., "The Windfall Clause: Distributing the Benefits of AI for the Common Good," *ArXiv* (December 2019), https://doi.org/10.48550/ArXiv.1912.11595.

3. Global AI Policy Challenges

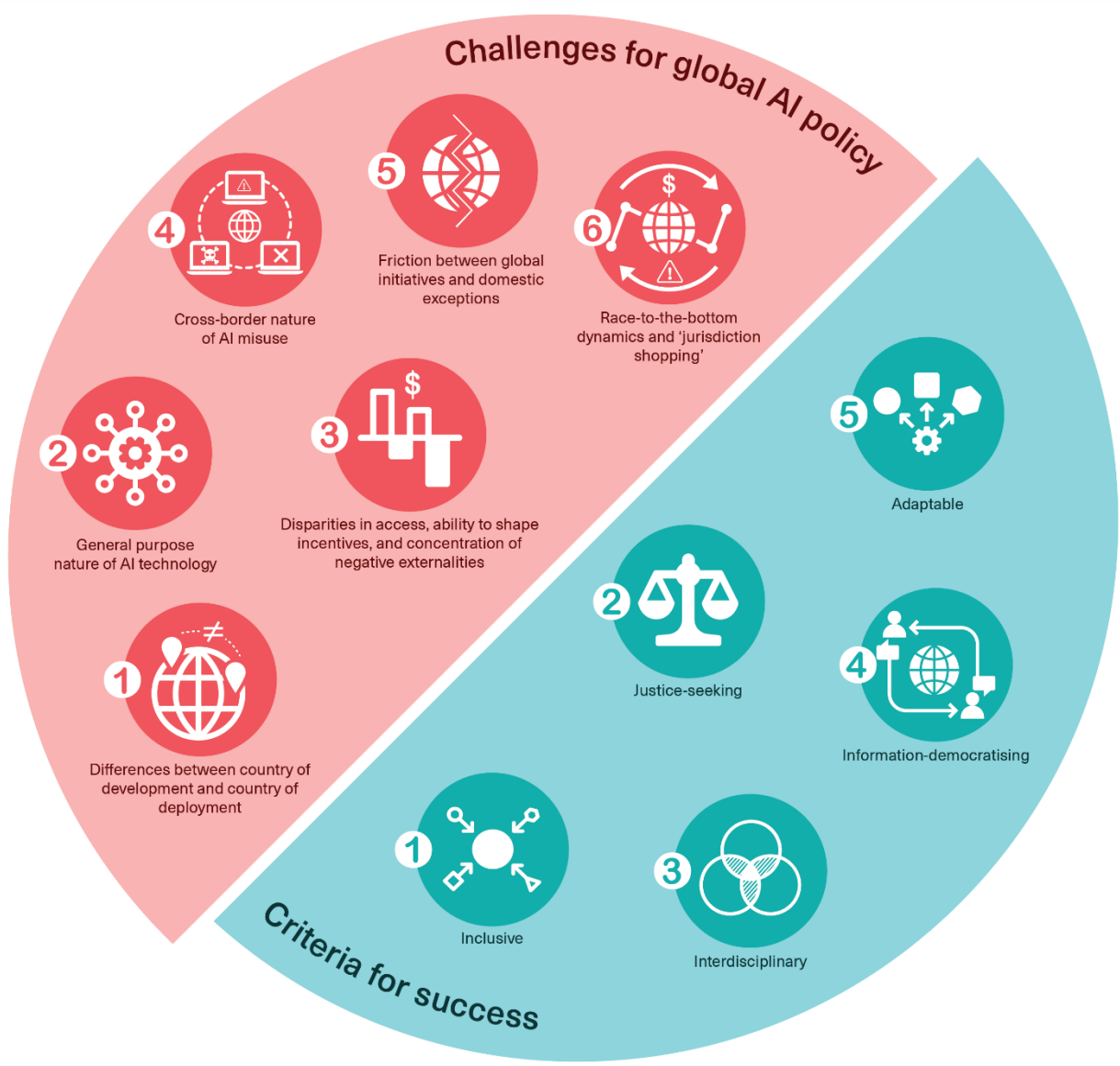
The policy options discussed in the previous section are a starting point for a comprehensive and proportionate domestic response to both the current and potential future risks posed by AI. But enacting change at the domestic level will only take individual countries so far. The AI ecosystem is global in terms of its cross-border supply chains and consumer bases, regulatory approaches can vary widely across jurisdictions and the harms of AI will not respect national boundaries. For the UK to influence approaches to AI standards and development globally, it will need a clear vision of its role in the global AI landscape, and the appetite to expend significant time and resources to achieve ambitious targets in this area.

To that end, it is worth summarising the challenges that global AI policy needs to solve, the characteristics that global AI policy must have to be successful, and the global AI policy levers available to the UK Government.

The 'criteria for success' outlined below introduce unavoidable trade-offs. For example, inclusivity will likely come at some cost to speed and bureaucracy, while interdisciplinarity may come at some cost to time taken to reach consensus. By making these criteria explicit, we hope to start conversations about where and how these trade-offs can and should be resolved.

3.1 Challenges to solve and criteria for success

Figure 4. An illustration of six challenges for global AI policy to solve and five criteria to be met to increase the likelihood of success.



Challenges for global AI policy
1. **The country of development and country of deployment of an AI system are often different**. There are information asymmetries in the global AI landscape, where the countries in which AI systems are often deployed have no insight or input into the model development process. This is compounded by the fact that cutting-edge AI is expensive to build from scratch, meaning that a disproportionate amount of global AI capability is concentrated in relatively small pockets of industry. The difficulty of achieving scale and sustainability as a company outside of these groupings means that even where smaller AI companies show promise, there is a good chance they will be bought out by a larger, more established player, as happened with Google's takeover of DeepMind in 2014, or re-prioritise revenue generation over social impact, as happened with OpenAI's updated charter in 2018.[95]
2. **General purpose nature of AI technology.** The range of possible applications for AI across societies and economies is vast, creating challenges for information gathering, monitoring and analysis. This challenge is significant even at the domestic level, and is further heightened at the international level, where different governments may have different systems for collecting, analysing and reporting information about AI, based on different values, technology stacks, ontologies and bureaucracies.
3. **Disparities in access, ability to shape incentives, and concentration of negative externalities**. There is a limited degree to which the country of deployment can shape incentives that are specific to their context, and they will therefore have limited means of preventing the proliferation of negative externalities without withdrawing from the AI ecosystem entirely.
4. **Misuse of AI by non-state actors easily crosses borders**. Without a globally coordinated approach across countries and sectors, it will prove even more difficult to prevent the harms produced by powerful AI systems from crossing into other jurisdictions.
5. **Friction between global initiatives and domestic exceptions.** The extent to which global agreements and domestic regulation apply to governments' internal AI research and innovation will be a source of inevitable tension, especially in the area of national security.
6. **Race-to-the-bottom dynamics and 'jurisdiction shopping'**. Global AI competition can be a force for good, but there is a risk that if left unbounded by concrete guardrails, it will fuel a race-to-the bottom on two fronts: first on the part of countries that seek to 'go faster and break things' to achieve and solidify a first-mover advantage, and second on the part of companies that gain leverage as a result of this dynamic and can extract concessions from governments eager for their investment.

[95] Chloe Xang, "OpenAI Is Now Everything It Promised Not to Be: Corporate, Closed-Source, and For-Profit," Motherboard, Tech by Vince, last modified 28 February, 2023, https://www.vice.com/en/article/5d3naz/openai-is-now-everything-it-promised-not-to-be-corporate-closed-source-and-for-profit.

Criteria for success
1. **Inclusive**. If AI is the era-defining technology that many world leaders believe it to be, with the potential to destabilise the geopolitical landscape, then high-democracies (predominantly those with leading roles in fora like the OECD, G7 and Global Partnership on AI) must re-evaluate their tolerance for engaging with a broader range of countries. Global AI competition can and should continue, but it is not in the global interest for transnational dialogue on AI safety to be limited to a select few countries.
2. **Justice-seeking**. Global justice issues must be foregrounded in global AI policy discussions – the most important conversations about AI need to involve governments who hold less power on the global stage, yet whose populations may be disproportionately affected by decisions made at this level. These country-to-country inequities are mirrored by inequities experienced by specific communities within countries; this distinction is important to remember given that politicians' priorities at the international level may not reflect the experiences of people at the community level.
3. **Interdisciplinary.** A wide variety of disciplines must guide global AI policy discussions to capture the best available information regarding frontier capabilities and their real-world impacts across different communities and sectors. Governments must exercise caution to avoid giving primacy to commercial interests as the global AI policy agenda unfolds. Global discussions concerning AI safety must provide a platform to researchers in universities, research institutes and civil society organisations who have long been documenting the risks posed by AI.
4. **Information-democratising**. Whether countries, companies, or individual researchers, there are inherent disparities in information held by actors entering into global AI policy discussions. The aim of any approach which seeks to foster coordination and shared understandings of norms, standards and red lines should be to narrow those disparities. This will require governments pushing companies to communicate more transparently about frontier capabilities, while not exposing the details that adversaries would require to implement those capabilities in malicious ways.
5. **Adaptable**. Relatively short periods of activity and innovation in the field of AI can have outsized effects on the momentum behind big policy initiatives, so actors shaping global AI policy discussions must possess a range of tools to shape behaviours in a rapidly changing socio-technical landscape. This will require staffing by technically informed personnel who keep abreast of the latest developments, the ability to revise processes and institutional structures with agility, and flexibility in legislation and regulation (e.g. licensing requirements) that reference regularly-updated expert opinion or more open references to "best practices."

3.2 Options for a comprehensive global strategy

As with domestic policy, the UK's successful international leadership on AI requires using policy levers available across each phase of the AI lifecycle. Policies should prioritise creating more visibility around AI risks, promoting best practices to identify and address these risks, and establishing incentives and enforcement to increase adherence to those practices.

The UK can build its leadership on these issues through piloting levers across each of these phases and evaluating their viability for multilateral implementation. The UK Government's ability to define an AI strategy separately from the EU, US, or China is an asset in this endeavour. It is particularly well-positioned to build bridges and serve as a convener between countries that would otherwise be at odds.

Global considerations raise policy challenges across all three of the policy goals outlined earlier:

- **Creating visibility and understanding**: A country may have little visibility into risks in AI development and deployment occurring outside of its jurisdiction, despite the fact that foreign activities can have cascading effects into other jurisdictions. International measures to increase shared visibility and understanding of AI risks could ameliorate arms race dynamics through effective cooperation built on a foundation of mutual assurance about the future development and use of AI. One potential model for creating greater visibility and understanding on an international level is a multilateral organisation similar to the Intergovernmental Panel on Climate Change to capture the global state of affairs in a rapidly changing landscape, and build consensus internationally on AI progress and its risks.

- **Promoting best practices:** Best practices for deploying AI in public services such as healthcare or education will differ between countries due to variance in methods of providing public services, as well as cultural beliefs regarding governments' role in the provision of such services. However, coordination on establishing global standards in certain high-risk areas of development and deployment, where there are shared threats all countries can agree on, will be crucial for encouraging safer practices and reducing risks of global damage. For example, we may want to prioritise ensuring that there is global agreement on minimum standards for system safety and adherence to human rights in applications, while allowing for some variation in how different value trade-offs are made. Thinking carefully about which

areas warrant global cooperation, and where varying national approaches may be acceptable or even desirable, will be essential here.

- **Establishing incentives and enforcing regulation:** On a global level, the successful establishment of a responsible AI regime requires strong multilateral institutions that have necessary expertise to monitor cross-border AI development and deployment, access to policy levers for establishing incentives, and could be capable of holding a national government accountable through some capability for imposing penalties. There are numerous options for incentive shaping and alignment in the global arena, from negotiated international treaties such as the Nuclear Non-Proliferation Treaty[96], multilateral agreements such as the Wassenaar Arrangements[97] on export controls, provisions relating to trade covered by the World Trade Organisation, and bilateral trade agreements. Design choices will need to be informed by the perceived relative strengths, vulnerabilities and preferences of different parties, including major corporations in addition to states and regional bodies.

[96] "Treaty on the Non-Proliferation of Nuclear Weapons (NPT)," International Atomic Energy Agency, n.d., https://www.iaea.org/publications/documents/treaties/npt.

[97] Wassenaar Arrangement Secretariat, "Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies," *Public Documents,* Vol 1, Founding Documents (February 2017), https://www.wassenaar.org/app/uploads/2015/06/WA-DOC-17-PUB-001-Public-Docs-Vol-I-Founding-Documents.pdf.

Conclusion

In recent months, attention to AI risk has increased drastically, including at the most senior levels of government. This paper aims to help ensure this moment of heightened interest can translate into practical action by providing a structured framework to identify, understand, and respond to these risks. In doing so, we hope to help policymakers effectively build upon the discussions on AI risks and harms that have developed over decades among academics, activists, and technologists, and to draw lessons from prior research.

Effectively addressing the most serious risks will depend upon trust and at times collaboration between government, industry, and civil society on an international scale. The current landscape presents a valuable opportunity for the UK to assume a leading role as an international convener on the topic of AI risk. To realise that ambition, it will be vital to harness the current momentum in AI policy to level up internal government understanding of technical and ethical risks. This process will introduce policymakers to a wider breadth of perspectives and provide them with a more authoritative basis upon which to build a clear and realistic vision of the UK's role in setting the global agenda. This vision must involve demonstrating the UK is a place which sets best practices in AI development and risk reduction, and establishes incentives and regulation which ensure that these are adhered to.

The UK is currently inadequately resilient to the risks posed by AI. Now is the time for the UK Government to act decisively on the big societal challenges facing our population. Any further delay will not only jeopardise the UK Government's ambitions of a global leadership role, but also make it all the more challenging to prevent AI risks from manifesting into real-world widespread harms.

About the Authors

Ardi Janjeva is a Research Associate at CETaS. His research interests are divided into three main areas: artificial intelligence innovation and disruption, intelligence tradecraft and investigatory powers, and emerging technology, political economy and strategy. He has worked closely with national and international partners across government, academia, civil society and the private sector on these topics, producing research which has been cited in various academic journals and mainstream media outlets.

Rosamund Powell is a Research Associate at CETaS. Prior to joining CETaS, Rosamund worked as an Ethics Research Assistant within The Alan Turing Institute, contributing to the public policy programme research agenda. While at the Turing, she has co-authored research reports exploring topics at the intersection of artificial intelligence, ethics, and policy.

Jess Whittlestone is Head of AI Policy at CLTR, where she leads their work on developing and advocating for policy recommendations to reduce extreme risks from AI. She has been working on various aspects of AI risk and policy since 2018, and was previously a Senior Research Fellow and Deputy Director of the AI: Futures and Responsibility Programme at the University of Cambridge. She holds a PhD in Behavioural Science from the University of Warwick.

Nikhil Mulani is a researcher collaborating with CLTR on developing policy proposals for reducing risks from frontier AI. He previously was a Winter Fellow at the Centre for the Governance of AI (GovAI). Nikhil's recent areas of interest in AI policy include information sharing, information security, and board governance. Before policy research, he worked as a product manager building machine learning software and as a consultant advising commercial and government clients on their technology strategies.

Shahar Avin is a senior researcher associate at the Centre for the Study of Existential Risk (CSER), at the University of Cambridge. He works with CSER researchers and others in the global catastrophic risk community to identify and design risk prevention strategies, through organising workshops, building agent-based models, and by frequently asking naive questions. He was co-lead author of two influential reports, The Malicious Use of Artificial Intelligence (2018) and Toward Trustworthy AI Development (2020). He has worked with various startups in Israel and the UK, as a software developer and ethics advisor. His work on this paper was carried out while he was a consultant for CLTR.

Centre for Emerging Technology and Security

BRIEFING PAPER