Department for
Science, Innovation
& Technology

A pro-innovation approach to AI regulation

March 2023

978-1-5286-4009-1

E02886733 03/23

CP 815

A pro-innovation approach to AI regulation

Presented to Parliament
by the Secretary of State for Science, Innovation and Technology
by Command of His Majesty

March 2023

CP 815

This publication is available at www.gov.uk/official-documents.

Any enquiries regarding this publication should be sent to us at: evidence@officeforai.gov.uk

ISBN 978-1-5286-4009-1

E02886733 03/23

Printed on paper containing 40% recycled fibre content minimum

Printed in the UK by HH Associates Ltd. on behalf of the Controller of His Majesty's Stationery Office

CORRECTION SLIP

Title: A pro-innovation approach to AI regulation

Session: 2022−23

CP 815

ISBN: 978-1-5286-4009-1

Correction:

Text currently reads **in Annex C**:

1. Do you agree that requiring organisations to make it clear when they are using AI would adequately ensure transparency?

2. What other transparency measures would be appropriate, if any?

3. Do you agree that current routes to contestability or redress for AI-related harms are adequate?

4. How could routes to contestability or redress for AI-related harms be improved, if at all?

[…]

L3. If you are a business that develops, uses, or sells AI, how do you currently manage AI risk including through the wider supply chain? How could government support effective AI-related risk management?

[…]

S1. Which of the sandbox models described in section 3.3.4 would be most likely to support innovation?

Text should read:

1: Do you agree that requiring organisations to make it clear when they are using AI would improve transparency?

2: Are there other measures we could require of organisations to improve transparency for AI?

3: Do you agree that current routes to contest or get redress for AI -related harms are adequate?

4: How could current routes to contest or seek redress for AI-related harms be improved, if at all?

[…]

L3: If you work for a business that develops, uses, or sells AI, how do you currently manage AI risk including through the wider supply chain? How could government support effective AI-related risk management?

[…]

S1: To what extent would the sandbox models described in section 3.3.4 support innovation?

Date of correction: 04 July 2023

Contents

Ministerial foreword

The Rt Hon Michelle Donelan MP,
Secretary of State for Science, Innovation and Technology

I believe that a common-sense, outcomes-oriented approach is the best way to get right to the heart of delivering on the priorities of people across the UK. Better public services, high quality jobs and opportunities to learn the skills that will power our future – these are the priorities that will drive our goal to become a science and technology superpower by 2030.

Artificial intelligence (AI) will play a central part in delivering and enabling these goals, and this white paper will ensure we are putting the UK on course to be the best place in the world to build, test and use AI technology. But we are not starting from zero. Having invested over £2.5 billion in AI since 2014, this paper builds on our recent announcements of £110 million for our AI Tech Missions Fund, £900 million to establish a new AI Research Resource and to develop an exascale supercomputer capable of running large AI models – backed up by our new £8 million AI Global Talent Network and £117 million of existing funding to create hundreds of new PhDs for AI researchers.

Most of us are only now beginning to understand the transformative potential of AI as the technology rapidly improves. But in many ways, AI is already delivering fantastic social and economic benefits for real people – from improving NHS medical care to making transport safer. Recent advances in things like generative AI give us a glimpse into the enormous opportunities that await us in the near future if we are prepared to lead the world in the AI sector with our values of transparency, accountability and innovation.

My vision for an AI-enabled country is one where our NHS heroes are able to save lives using AI technologies that were unimaginable just a few decades ago. I want our police, transport networks and climate scientists and many more to be empowered by AI technologies that will make the UK the smartest, healthiest, safest and happiest place to live and work. That is why AI is one of this government's five technologies of tomorrow - bringing stronger growth, better jobs, and bold new discoveries. It is a vision that has been shaped by stakeholders and experts in AI, whose expertise and ideas I am determined to see reflected in our department.

The UK has been at the forefront of this progress, placing third in the world for AI research and development. We are home to a third of Europe's total AI companies and twice as many as any other European country. Our world-leading status is down to our thriving research base and the pipeline of

expertise graduating through our universities, the ingenuity of our innovators and the government's long-term commitment to invest in AI.

To ensure we become an AI superpower, though, it is crucial that we do all we can to create the right environment to harness the benefits of AI and remain at the forefront of technological developments. That includes getting regulation right so that innovators can thrive and the risks posed by AI can be addressed.

These risks could include anything from physical harm, an undermining of national security, as well as risks to mental health. The development and deployment of AI can also present ethical challenges which do not always have clear answers. Unless we act, household consumers, public services and businesses will not trust the technology and will be nervous about adopting it. Unless we build public trust, we will miss out on many of the benefits on offer.

Indeed, the pace of change itself can be unsettling. Some fear a future in which AI replaces or displaces jobs, for example. Our white paper and our vision for a future AI-enabled country is one in which our ways of working are complemented by AI rather than disrupted by it. In the modern world, too much of our professional lives are taken up by monotonous tasks – inputting data, filling out paperwork, scanning through documents for one piece of information and so on. AI in the workplace has the potential to free us up from these tasks, allowing us to spend more time doing the things we trained for – teachers with more time to teach, clinicians with more time to spend with patients, police officers with more time on the beat rather than behind a desk – the list goes on.

Indeed, since AI is already in our day-to-day lives, there are numerous examples that can help to illustrate the real, tangible benefits that AI can bring once any risks are mitigated. Streaming services already use advanced AI to recommend TV shows and films to us. Our satnav uses AI to plot the fastest routes for our journeys, or helps us avoid traffic by intelligently predicting where congestion will be on our journey. And of course, almost all of us carry a smartphone in our pockets that uses advanced AI in all sorts of ways. These common devices all carried risks at one time or another, but today they benefit us enormously.

That is why our white paper details how we intend to support innovation while providing a framework to ensure risks are identified and addressed. However, a heavy-handed and rigid approach can stifle innovation and slow AI adoption. That is why we set out a proportionate and pro-innovation regulatory framework. Rather than target specific technologies, it focuses on the context in which AI is deployed. This enables us to take a balanced approach to weighing up the benefits versus the potential risks.

We recognise that particular AI technologies, foundation models for example, can be applied in many different ways and this means the risks can vary hugely. For example, using a chatbot to produce a summary of a long article presents very different risks to using the same technology to provide medical advice. We understand the need to monitor these developments in partnership with innovators while also avoiding placing unnecessary regulatory burdens on those deploying AI.

To ensure our regulatory framework is effective, we will leverage the expertise of our world class regulators. They understand the risks in their sectors and are best placed to take a proportionate approach to regulating AI. This will mean supporting innovation and working closely with business, but also stepping in to address risks when necessary. By underpinning the framework with a set of principles, we will drive consistency across regulators while also providing them with the flexibility needed.

For innovators working at the cutting edge and developing novel technologies, navigating regulatory regimes can be challenging. That's why we are confirming our commitment to taking forward a key recommendation made by Sir Patrick Vallance to establish a regulatory sandbox for AI. This will bring together regulators to support innovators directly and help them get their products to market. The sandbox will also enable us to understand how regulation interacts with new technologies and refine this interaction where necessary.

Having exited the European Union we are free to establish a regulatory approach that enables us to establish the UK as an AI superpower. It is an approach that will actively support innovation while

addressing risks and public concerns. The UK is home to thriving start-ups, which our framework will support to scale-up and compete internationally. Our pro-innovation approach will also act as a strong incentive when it comes to AI businesses based overseas establishing a presence in the UK. The white paper sets out our commitment to engaging internationally to support interoperability across different regulatory regimes. Not only will this ease the burden on business but it will also allow us to embed our values as global approaches to governing AI develop.

Our approach relies on collaboration between government, regulators and business. Initially, we do not intend to introduce new legislation. By rushing to legislate too early, we would risk placing undue burdens on businesses. But alongside empowering regulators to take a lead, we are also setting expectations. Our new monitoring functions will provide a real time assessment of how the regulatory framework is performing so that we can be confident that it is proportionate. The pace of technological development also means that we need to understand new and emerging risks, engaging with experts to ensure we take action where necessary. A critical component of this activity will be engaging with the public to understand their expectations, raising awareness of the potential of AI and demonstrating that we are responding to concerns.

The framework set out in this white paper is deliberately designed to be flexible. As the technology evolves, our regulatory approach may also need to adjust. Our principles based approach, with central functions to monitor and drive collaboration, will enable us to adapt as needed while providing industry with the clarity needed to innovate. We will continue to develop our approach, building on our commitment to making the UK the best place in the world to be a business developing and using AI. Responses to the consultation will inform how we develop the regulatory framework - I encourage all of those with an interest to respond.

RT HON MICHELLE DONELAN MP
Secretary of State for Science, Innovation and Technology
Department for Science, Innovation and Technology

Executive summary

Artificial intelligence – the opportunity and the challenge

1. Artificial intelligence (AI) is already delivering wide societal benefits, from medical advances[1] to mitigating climate change.[2] For example, an AI technology developed by DeepMind, a UK-based business, can now predict the structure of almost every protein known to science.[3] This breakthrough will accelerate scientific research and the development of life-saving medicines – it has already helped scientists to make huge progress in combating malaria, antibiotic resistance, and plastic waste.

2. The UK Science and Technology Framework[4] sets out government's strategic vision and identifies AI as one of five critical technologies. The framework notes the role of regulation in creating the environment for AI to flourish. We know that we have yet to see AI technologies reach their full potential. Under the right conditions, AI will transform all areas of life[5] and stimulate the UK economy by unleashing innovation and driving productivity,[6] creating new jobs and improving the workplace.

3. Across the world, countries and regions are beginning to draft the rules for AI. The UK needs to act quickly to continue to lead the international conversation on AI governance and demonstrate the value of our pragmatic, proportionate regulatory approach. The need to act was highlighted by Sir Patrick Vallance in his recent Regulation for Innovation review. The report identifies the short time frame for government intervention to provide a clear, pro-innovation regulatory environment in order to make the UK one of the top places in the world to build foundational AI companies.[7]

4. While we should capitalise on the benefits of these technologies, we should also not overlook the new risks that may arise from their use, nor the unease that the complexity of AI technologies can produce in the wider public. We already know that some uses of AI could damage our physical[8] and mental health,[9] infringe on the privacy of individuals[10] and undermine human rights.[11]

5. Public trust in AI will be undermined unless these risks, and wider concerns about the potential for bias and discrimination, are addressed. By building trust, we can accelerate the adoption of AI across the UK to maximise the economic and social benefits that the technology can deliver,

---

1 The use of AI in healthcare and medicine is booming, Insider Intelligence, 2023.

2 How to fight climate change using AI, Forbes, 2022; Tackling Climate Change with Machine Learning, Rolnick et al., 2019.

3 DeepMind's protein-folding AI cracks biology's biggest problem, New Scientist, 2022; Improved protein structure prediction using potentials from deep learning, Senior et al., 2020.

4 The UK Science and Technology Framework, Department for Science, Innovation and Technology, 2023.

5 Six of the best future uses of Artificial Intelligence, Technology Magazine, 2023; Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy, Dwivedi et al., 2021.

6 Large dedicated AI companies make a major contribution to the UK economy, with GVA (gross value added) per employee estimated to be £400k, more than double that of comparable estimates of large dedicated firms in other sectors. See AI Sector Study 2022, DSIT, 2023.

7 Pro-innovation Regulation of Technologies Review: Digital Technologies, HM Treasury, 2023.

8 AI Barometer Part 4 – Transport and logistics, Centre for Data Ethics and Innovation, 2021.

9 How TikTok Reads Your Mind, New York Times, 2021.

10 Privacy Considerations in Large Language Models, Google Research, 2020.

11 Artificial Intelligence, Human Rights, Democracy, and the Rule of Law, Alan Turing Institute and Council of Europe, 2021.

while attracting investment and stimulating the creation of high-skilled AI jobs.[12] In order to maintain the UK's position as a global AI leader, we need to ensure that the public continues to see how the benefits of AI can outweigh the risks.[13]

6. Responding to risk and building public trust are important drivers for regulation. But clear and consistent regulation can also support business investment and build confidence in innovation. Throughout our extensive engagement, industry repeatedly emphasised that consumer trust is key to the success of innovation economies. We therefore need a clear, proportionate approach to regulation that enables the responsible application of AI to flourish. Instead of creating cumbersome rules applying to all AI technologies, our framework ensures that regulatory measures are proportionate to context and outcomes, by focusing on the use of AI rather than the technology itself.

7. People and organisations develop and use AI in the UK within the rules set by our existing laws, informed by standards, guidance and other tools. But AI is a general purpose technology and its uses can cut across regulatory remits. As a result, AI technologies are currently regulated through a complex patchwork of legal requirements. We are concerned by feedback from across industry that the absence of cross-cutting AI regulation creates uncertainty and inconsistency which can undermine business and consumer confidence in AI, and stifle innovation. By providing a clear and unified approach to regulation, our framework will build public confidence, making it clear that AI technologies are subject to cross-cutting, principles-based regulation.

Our pro-innovation framework

8. The government will put in place a new framework to bring clarity and coherence to the AI regulatory landscape. This regime is designed to make responsible innovation easier. It will strengthen the UK's position as a global leader in AI, harness AI's ability to drive growth and prosperity,[14] and increase public trust in its use and application.

9. We are taking a deliberately agile and iterative approach, recognising the speed at which these technologies are evolving. Our framework is designed to build the evidence base so that we can learn from experience and continuously adapt to develop the best possible regulatory regime. Industry has praised our pragmatic and proportionate approach.

10. Our framework is underpinned by five principles to guide and inform the responsible development and use of AI in all sectors of the economy:

    - Safety, security and robustness
    - Appropriate transparency and explainability
    - Fairness
    - Accountability and governance
    - Contestability and redress

11. We will not put these principles on a statutory footing initially. New rigid and onerous legislative requirements on businesses could hold back AI innovation and reduce our ability to respond quickly and in a proportionate way to future technological advances. Instead, the principles will be issued on a non-statutory basis and implemented by existing regulators. This approach

[12] Demand for AI skills in jobs, OECD iLibrary, 2021.

[13] Public expectations for AI governance (transparency, fairness and accountability), Centre for Data Ethics and Innovation, 2023.

[14] The AI sector is estimated to contribute £3.7bn in GVA (Gross Value Added) to the UK economy. AI Sector Study 2022, DSIT, 2023.

makes use of regulators' domain-specific expertise to tailor the implementation of the principles to the specific context in which AI is used. During the initial period of implementation, we will continue to collaborate with regulators to identify any barriers to the proportionate application of the principles, and evaluate whether the non-statutory framework is having the desired effect.

12. Following this initial period of implementation, and when parliamentary time allows, we anticipate introducing a statutory duty on regulators requiring them to have due regard to the principles. Some feedback from regulators, industry and academia suggested we should implement further measures to support the enforcement of the framework. A duty requiring regulators to have regard to the principles should allow regulators the flexibility to exercise judgement when applying the principles in particular contexts, while also strengthening their mandate to implement them. In line with our proposal to work collaboratively with regulators and take an adaptable approach, we will not move to introduce such a statutory duty if our monitoring of the framework shows that implementation is effective without the need to legislate.

13. In the 2022 AI regulation policy paper,[15] we proposed a small coordination layer within the regulatory architecture. Industry and civil society were supportive of our intention to ensure coherence across the AI regulatory framework. However, feedback often argued strongly for greater central coordination to support regulators on issues requiring cross-cutting collaboration and ensure that the overall regulatory framework functions as intended.

14. We have identified a number of central support functions required to make sure that the overall framework offers a proportionate but effective response to risk while promoting innovation across the regulatory landscape:

    - Monitoring and evaluation of the overall regulatory framework's effectiveness and the implementation of the principles, including the extent to which implementation supports innovation. This will allow us to remain responsive and adapt the framework if necessary, including where it needs to be adapted to remain effective in the context of developments in AI's capabilities and the state of the art.
    - Assessing and monitoring risks across the economy arising from AI.
    - Conducting horizon scanning and gap analysis, including by convening industry, to inform a coherent response to emerging AI technology trends.
    - Supporting testbeds and sandbox initiatives to help AI innovators get new technologies to market.
    - Providing education and awareness to give clarity to businesses and empower citizens to make their voices heard as part of the ongoing iteration of the framework.
    - Promoting interoperability with international regulatory frameworks.

15. The central support functions will initially be provided from within government but will leverage existing activities and expertise from across the broader economy. The activities described above will neither replace nor duplicate the work undertaken by regulators and will not involve the creation of a new AI regulator.

16. Our proportionate approach recognises that regulation is not always the most effective way to support responsible innovation. The proposed framework is aligned with, and supplemented by,

[15] Establishing a pro-innovation approach to regulating AI, Office for Artificial Intelligence, 2022.

a variety of tools for trustworthy AI, such as assurance techniques, voluntary guidance and technical standards. Government will promote the use of such tools. We are collaborating with partners like the UK AI Standards Hub to ensure that our overall governance framework encourages responsible AI innovation (see part four for details).

17. In keeping with the global nature of these technologies, we will also continue to work with international partners to deliver interoperable measures that incentivise the responsible design, development and application of AI. During our call for views, industry, academia and civil society stressed that international alignment should support UK businesses to capitalise on global markets and protect UK citizens from cross-border harms.

18. The UK is frequently ranked third in the world across a range of measures, including level of investment, innovation and implementation of AI.[16] To make the UK the most attractive place in the world for AI innovation and support UK companies wishing to export and attract international investment, we must ensure international compatibility between approaches. Countries around the world, as well as multilateral forums, are exploring approaches to regulating AI. Thanks to our reputation for pragmatic regulation, the UK is rightly seen by international partners as a leader in this global conversation.

[16] Global AI Index, Tortoise Media, 2022.

Part One: Introduction

1.1 The power and potential of artificial intelligence

19. AI is already delivering major advances and efficiencies in many areas. AI quietly automates aspects of our everyday activities, from systems that monitor traffic to make our commutes smoother,[17] to those that detect fraud in our bank accounts.[18] AI has revolutionised large-scale safety-critical practices in industry, like controlling the process of nuclear fusion.[19] And it has also been used to accelerate scientific advancements, such as the discovery of new medicine[20] or the technologies we need to tackle climate change.[21]

20. But this is just the beginning. AI can be used in a huge variety of settings and has the extraordinary potential to transform our society and economy.[22] It could have as much impact as electricity or the internet, and has been identified as one of five critical technologies in the UK Science and Technology Framework.[23] As AI becomes more powerful, and as innovators explore new ways to use it, we will see more applications of AI emerge. As a result, AI has a huge potential to drive growth[24] and create jobs.[25] It will support people to carry out their existing jobs, by helping to improve workforce efficiency and workplace safety.[26] To remain world leaders in AI, attract global talent and create high-skilled jobs in the UK, we must create a regulatory environment where such innovation can thrive.

21. Technological advances like large language models (LLMs) are an indication of the transformative developments yet to come.[27] LLMs provide substantial opportunities to transform the economy and society. For example, LLMs can automate the process of writing code and fixing programming bugs. The technology can support genetic medicine by identifying links between genetic sequences and medical conditions. It can support people to review and summarise key points from lengthy documents. In the last four years, LLMs have been developed beyond expectations and they are becoming applicable to an increasingly wide range of tasks.[28] We expand on the development of LLM and other foundation models in section 3.3.3 below.

---

[17] Transport apps like Google Maps, and CityMapper, use AI.

[18] Artificial Intelligence in Banking Industry: A Review on Fraud Detection, Credit Management, and Document Processing, ResearchBerg Review of Science and Technology, 2018.

[19] Accelerating fusion science through learned plasma control, Deepmind, 2022; Magnetic control of tokamak plasmas through deep reinforcement learning, Degrave et al., 2022.

[20] Why Artificial Intelligence Could Speed Drug Discovery, Morgan Stanley, 2022.

[21] AI Is Essential for Solving the Climate Crisis, BCG, 2022.

[22] General Purpose Technologies – Handbook of Economic Growth, National Bureau of Economic Research, 2005.

[23] The UK Science and Technology Framework, Department for Science, Innovation and Technology, 2023.

[24] In 2022 annual revenues generated by UK AI companies totalled an estimated £10.6 billion. AI Sector Study 2022, DSIT, 2023.

[25] DSIT analysis estimates over 50,000 full time workers are employed in AI roles in AI companies. AI Sector Study 2022, DSIT, 2023.

[26] For example, AI can potentially improve health and safety in mining while also improving efficiency. See AI on-side: how artificial intelligence is being used to improve health and safety in mining, Axora, 2023. Box 1.1 gives further examples of AI driving efficiency improvements.

[27] Large Language Models Will Define Artificial Intelligence, Forbes, 2023; Scaling Language Models: Methods, Analysis & Insights from Training Gopher, Borgeaud et al., 2022.

[28] See, for example, What are Large Language Models used for? NVIDIA, 2023.

Box 1.1: Examples of AI opportunities

AI helps piece together the first complete image of a black hole

AI can enable scientific discovery. A computer vision model was used to piece together the first ever image of a black hole 55 million light years away, combining images from eight telescopes around the world.[29]

AI solves decades old protein-folding puzzle

An AI company based in the UK trained neural networks to predict the structures of proteins, solving a problem that had long stumped scientists. The predictions are advancing the field of structural biology: scientists have already used them to prevent antibiotic resistance,[30] advance disease research,[31] and accelerate the fight against plastic pollution.[32] As we find more uses for AI, it will rewrite scientific fields and change the way we learn about our world.

Deep learning AI could improve breast cancer screening

AI could transform how diseases are detected, prevented, and treated. Doctors are testing if deep learning can be applied to breast cancer screening. Currently, every mammogram is double-checked by radiologists but this is labour-intensive and causes diagnosis delays. A UK medical technology company is working with the NHS to test AI for the second screening, meaning greater numbers of patients could be screened faster and clinicians could spend more time with patients and provide faster access to treatment.[33]

Farming efficiency increased by AI robots

Applying robotics and AI to field management can make farming more efficient, sustainable and productive. Lightweight, autonomous mapping and monitoring robots operating across the UK can spend hours on the field in all conditions and significantly reduce soil compaction. These systems can digitise the field, providing farmers with data to improve weed and pest management. If these systems become widely used, they could contribute to agricultural and horticultural productivity, reduce the pressure of labour shortages and better preserve the environment.[34]

AI helps accelerate the discovery of new medicines

Significant time and resources are currently needed to develop new and effective medicines. AI can accelerate the discovery of new medicines by quickly identifying potential biologically active compounds from millions of candidates within a short

[29] Black hole pictured for first time – in spectacular detail, Nature, 2019.
[30] Accelerating the race against antibiotic resistance, Deepmind, 2022.
[31] Stopping malaria in its tracks, Deepmind, 2022.
[32] Creating plastic-eating enzymes that could save us from pollution, Deepmind, 2022.
[33] Mia mammography intelligent assessment, NHS England, 2021.
[34] Robotics and Autonomous Systems for Net Zero Agriculture, Pearson et al., 2022.

period.[35] Scientists may also have succeeded in using generative AI to design antibodies that bind to a human protein linked to cancer.[36]

AI is used in the fight against the most serious and harmful crimes

The Child Images Abuse Database[37] uses the powerful data processing capabilities of AI to identify victims and perpetrators of child sexual abuse. The quick and effective identification of victims and perpetrators in digital abuse images allows for real world action to remove victims from harm and ensure their abusers are held to account. The use of AI increases the scale and speed of analysis while protecting staff welfare by reducing their exposure to distressing content.

AI increases cyber security capabilities

Companies providing cyber security services are increasingly using AI to analyse large amounts of data about malware and respond to vulnerabilities in network security at faster-than-human speeds.[38] As the complexity of the cyber threat landscape evolves, the pattern-recognition and recursive learning capabilities of AI are likely to play an increasingly significant role in proactive cyber defence against malicious actors.

1.2 Managing AI risks

22. The concept of AI is not new, but recent advances in data generation and processing have changed the field and the technology it produces. For example, while recent developments in the capabilities of generative AI models have created exciting opportunities, they have also sparked new debates about potential AI risks.[39] As AI research and development continues at pace and scale, we expect to see even greater impact and public awareness of AI risks.[40]

23. We know that not all AI risks arise from the deliberate action of bad actors. Some AI risks can emerge as an unintended consequence or from a lack of appropriate controls to ensure responsible AI use.[41]

24. We have made an initial assessment of AI-specific risks and their potential to cause harm, with reference in our analysis to the values that they threaten if left unaddressed. These values include safety, security, fairness, privacy and agency, human rights, societal well-being and prosperity.

25. Our assessment of cross-cutting AI risk identified a range of high-level risks that our framework will seek to prioritise and mitigate with proportionate interventions. For example, safety risks

[35] Artificial intelligence, big data and machine learning approaches to precision medicine and drug discovery, Current Drug Targets, 2021.
[36] Unlocking *de novo* antibody design with generative artificial intelligence, Shanehsazzadeh et al., 2023.
[37] Pioneering new tools to be rolled out in fight against child abusers, Home Office, 2019.
[38] Intelligent security tools, National Cyber Security Centre, 2019.
[39] What is generative AI, and why is it suddenly everywhere?, Vox, 2023.
[40] See, for example, The Benefits and Harms of Algorithms, The Digital Regulation Cooperation Forum, 2022; Harms of AI, Acemoglu, 2021.
[41] AI Accidents: An Emerging Threat, Center for Security and Emerging Technology, 2021.

include physical damage to humans and property, as well as damage to mental health.[42] AI creates a range of new security risks to individuals, organisations, and critical infrastructure.[43] Without government action, AI could cause and amplify discrimination that results in, for example, unfairness in the justice system.[44] Similarly, without regulatory oversight, AI technologies could pose risks to our privacy and human dignity, potentially harming our fundamental liberties.[45] Our regulatory intervention will ensure that AI does not cause harm at a societal level, threatening democracy[46] or UK values.

Box 1.2: Illustrative AI risks

The patchwork of legal frameworks that currently regulate some uses of AI may not sufficiently address the risks that AI can pose. The following examples are **hypothetical scenarios** designed to illustrate AI's potential to create harm.

Risks to human rights

Generative AI is used to generate deepfake pornographic video content, potentially damaging the reputation, relationships and dignity of the subject.

Risks to safety

An AI assistant based on LLM technology recommends a dangerous activity that it has found on the internet, without understanding or communicating the context of the website where the activity was described. The user undertakes this activity causing physical harm.

Risks to fairness[47]

An AI tool assessing credit-worthiness of loan applicants is trained on incomplete or biased data, leading the company to offer loans to individuals on different terms based on characteristics like race or gender.

Risks to privacy and agency

Connected devices in the home may constantly gather data, including conversations, potentially creating a near-complete portrait of an individual's home life. Privacy risks are compounded the more parties can access this data.

Risks to societal wellbeing

Disinformation generated and propagated by AI could undermine access to reliable information and trust in democratic institutions and processes.

Risks to security

[42] AI for radiographic COVID-19 detection selects shortcuts over signal, DeGrave, Janizek and Lee, 2021; Pathways: How digital design puts children at risk, 5Rights Foundation, 2021.

[43] The Malicious Use of Artificial Intelligence, Malicious AI Report, 2018.

[44] Constitutional Challenges in the Algorithmic Society, Micklitz et al., 2022.

[45] Smart Speakers and Voice Assistants, CDEI, 2019; Deepfakes and Audiovisual disinformation, CDEI, 2019.

[46] Artificial Intelligence, Human Rights, Democracy and the Rule of Law, Leslie et al., 2021.

[47] Government has already committed to addressing some of these issues more broadly. See, for example, the Inclusive Britain report, Race Disparity Unit, 2022.

AI tools can be used to automate, accelerate and magnify the impact of highly targeted cyber attacks, increasing the severity of the threat from malicious actors. The emergence of LLMs enables hackers[48] with little technical knowledge or skill to generate phishing campaigns with malware delivery capabilities.[49]

1.3 A note on terminology

Terminology used in this paper:[50]

AI or **AI system** or **AI technologies**: products and services that are 'adaptable' and 'autonomous' in the sense outlined in our definition in section 3.2.1.

AI supplier: any organisation or individual who plays a role in the research, development, training, implementation, deployment, maintenance, provision or sale of AI systems.

AI user: any individual or organisation that uses an AI product.

AI life cycle: all events and processes that relate to an AI system's lifespan, from inception to decommissioning, including its design, research, training, development, deployment, integration, operation, maintenance, sale, use and governance.

AI ecosystem: the complex network of actors and processes that enable the use and supply of AI throughout the AI life cycle (including supply chains, markets, and governance mechanisms).

Foundation model: a type of AI model that is trained on a vast quantity of data and is adaptable for use on a wide range of tasks. Foundation models can be used as a base for building more specific AI models. Foundation models are discussed in more detail in section 3.3.3 below.[51]

Impacted third party: an individual or company that is impacted by the outcomes of the AI systems that they do not use or supply themselves.

[48] 'Is ChatGPT a cybersecurity threat?' TechCrunch, 2023.

[49] OPWNAI: Cybercriminals starting to use ChatGPT, Check Point Research, 2023.

[50] These are not intended to be legal definitions for the purposes of the framework.

[51] The value chain of general-purpose AI, Ada Lovelace Institute, 2023.

Part Two: The current regulatory environment

2.1 Navigating the current landscape

26. The UK's AI success is, in part, due to our reputation for high-quality regulators and our strong approach to the rule of law, supported by our technology-neutral legislation and regulations. UK laws, regulators and courts already address some of the emerging risks posed by AI technologies (see box 2.1 for examples). This strong legal foundation encourages investment in new technologies, enabling AI innovation to thrive,[52] and high-quality jobs to flourish.[53]

Box 2.1: Example of legal coverage of AI in the UK and potential gaps

Discriminatory outcomes that result from the use of AI may contravene the protections set out in the Equality Act 2010.[54] AI systems are also required by data protection law to process personal data fairly.[55] However, AI can increase the risk of unfair bias or discrimination across a range of indicators or characteristics. This could undermine public trust in AI.

Product safety laws ensure that goods manufactured and placed on the market in the UK are safe. Product-specific legislation (such as for electrical and electronic equipment,[56] medical devices,[57] and toys[58]) may apply to some products that include integrated AI. However, safety risks specific to AI technologies should be monitored closely. As the capability and adoption of AI increases, it may pose new and substantial risks that are unaddressed by existing rules.

Consumer rights law[59] may protect consumers where they have entered into a sales contract for AI-based products and services. Certain contract terms (for example, that goods are of satisfactory quality, fit for a particular purpose, and as described) are relevant to consumer contracts. Similarly, businesses are prohibited from including certain terms in consumer contracts. Tort law provides a complementary regime that may provide redress where a civil wrong has caused harm. It is not yet clear whether consumer rights law will provide the right level of protection in the context of products that include integrated AI or services based on AI, or how tort law may apply to fill any gap in consumer rights law protection.

52 Global Innovation Index 2022, GII 2022; Global Indicators of Regulatory Governance, World Bank, 2023.

53 Demand for AI skills in jobs, OECD Science, Technology and Industry Working Papers, 2021.

54 The protected characteristics are age, disability, gender reassignment, marriage and civil partnership, race, religion or belief, sex, and sexual orientation.

55 Article 5(1)(a) Principles relating to processing of personal data, HM Government, 2016.

56 Electrical Equipment (Safety) Regulations, HM Government, 2016.

57 Medical Devices Regulation, HM Government, 2002.

58 Toys (Safety) Regulations, HM Government, 2011.

59 Consumer Rights Act 2015; Consumer Protection from Unfair Trading Regulations, HM Government, 2008.

27. While AI is currently regulated through existing legal frameworks like financial services regulation,[60] some AI risks arise across, or in the gaps between, existing regulatory remits. Industry told us that conflicting or uncoordinated requirements from regulators create unnecessary burdens and that regulatory gaps may leave risks unmitigated, harming public trust and slowing AI adoption.

28. Industry has warned us that regulatory incoherence could stifle innovation and competition by causing a disproportionate amount of smaller businesses to leave the market. If regulators are not proportionate and aligned in their regulation of AI, businesses may have to spend excessive time and money complying with complex rules instead of creating new technologies. Small businesses and start-ups often do not have the resources to do both.[61] With the vast majority of digital technology businesses employing under 50 people,[62] it is important to ensure that regulatory burdens do not fall disproportionately on smaller companies, which play an essential role in the AI innovation ecosystem and act as engines for economic growth and job creation.[63]

29. Regulatory coordination will support businesses to invest confidently in AI innovation and build public trust by ensuring real risks are effectively addressed. While some regulators already work together to ensure regulatory coherence for AI through formal networks like the AI and digital regulations service in the health sector[64] and the Digital Regulation Cooperation Forum (DRCF), other regulators have limited capacity and access to AI expertise. This creates the risk of inconsistent enforcement across regulators. There is also a risk that some regulators could begin to dominate and interpret the scope of their remit or role more broadly than may have been intended in order to fill perceived gaps in a way that increases incoherence and uncertainty. Industry asked us to support further system-wide coordination to clarify who is responsible for addressing cross-cutting AI risks and avoid duplicate requirements across multiple regulators.

[60] Such as the Financial Services and Markets Act, HM Government, 2000.

[61] Evidence to support the analysis of impacts for AI governance, Frontier Economics, 2023.

[62] In 2019, 98.8% of businesses in the digital sector had less than 50 employees. DCMS Sectors Economic Estimates 2019: Business Demographics, ONS, 2022.

[63] The AI Sector Study found that almost 90% of businesses in the AI sector are small or micro in size. AI Sector Study 2022, DSIT, 2023.

[64] AI and Digital Regulations Service, Care Quality Commission, Health Research Authority, Medicines and Healthcare Products Regulatory Agency, National Institute for Health and Care Excellence, 2023.

Case study 2.1: Addressing AI fairness under the existing legal and regulatory framework

A fictional company, *"AI Fairness Insurance Limited",* is designing a new AI-driven algorithm to set prices for insurance premiums that accurately reflect a client's risk. Setting fair prices and building consumer trust is a key component of AI Fairness Insurance Limited's brand so ensuring it complies with the relevant legislation and guidance is a priority.

Fairness in AI systems is covered by a variety of regulatory requirements and best practice. AI Fairness Insurance Limited's use of AI to set prices for insurance premiums could be subject to a range of legal frameworks, including data protection, equality, and general consumer protection laws. It could also be subject to sectoral rules like the Financial Services and Markets Act 2000.[65]

It can be challenging for a company like AI Fairness Insurance Limited to identify which rules are relevant and confidently apply them to AI use cases. There is currently a lack of support for businesses like AI Fairness Insurance Limited to navigate the regulatory landscape, with no cross-cutting principles and limited system-wide coordination.

30. Government intervention is needed to improve the regulatory landscape. We intend to leverage and build on existing regimes, maximising the benefits of what we already have, while intervening in a proportionate way to address regulatory uncertainty and gaps. This will deliver a pro-innovation regulatory framework that is designed to be adaptable and future-proof, supported by tools for trustworthy AI including assurance techniques and technical standards. This approach will provide more clarity and encourage collaboration between government, regulators and industry to unlock innovation.

65 Financial Services and Markets Act, HM Government, 2000.

Case study 2.2: Adapting regulatory approaches to AI – AI as a medical device

Some UK regulators have led the way and proactively adapted their approaches to AI-enabled technologies.

In 2022, the MHRA (Medicines and Healthcare products Regulatory Agency) published a roadmap clarifying in guidance the requirements for AI and software used in medical devices.[66] The regulator is also updating the regulatory framework for medical devices to protect patients and secure the UK's global reputation for responsible innovation in medical device software.

As part of this work, the MHRA will develop guidance on the transparency and interpretability of AI as a medical device. [67] The MHRA will consider the specific challenges posed by AI in this context, drawing on the applicable AI regulation cross-sectoral principles and ethical principles for AI in health and social care to issue practical guidance on how to meet legal product safety requirements. The MHRA will work with other regulators such as the Information Commissioner's Office (ICO) and the National Data Guardian to consider patients' data protection and trust in medical devices.

This work will provide manufacturers with clear requirements and guidance to attract responsible innovation to the UK.

[66] Software and AI as a Medical Device Change Programme – Roadmap, MHRA, 2022.

[67] The exact relation between the concepts 'interpretability' and 'explainability' is the subject of ongoing academic debate. See Interpretable and explainable machine learning: A methods-centric overview with concrete examples, Marcinkevics and Vogt, 2023. We use 'explainability' as the key term in our AI principle in alignment with the OECD.

Part Three: An innovative and iterative approach

3.1 Aims of the regulatory framework

31. Regulation can increase innovation by giving businesses the incentive to solve important problems while addressing the risk of harm to citizens. For example, product safety legislation has increased innovation towards safer products and services.[68] In the case of AI, a context-based, proportionate approach to regulation will help strengthen public trust and increase AI adoption.[69]

32. The National AI Strategy set out our aim to regulate AI effectively and support innovation.[70] In line with the principles set out in the Plan for Digital Regulation,[71] our approach to AI regulation will be proportionate; balancing real risks against the opportunities and benefits that AI can generate. We will maintain an effective balance as we implement the framework by focusing on the context and outcomes of AI.

33. Our policy paper proposed a pro-innovation framework designed to give consumers the confidence to use AI products and services, and provide businesses the clarity they need to invest in AI and innovate responsibly.[72] This approach was broadly welcomed – particularly by industry. Based on feedback, we have distilled our aims into three objectives that our framework is designed to achieve:

    - **Drive growth and prosperity** by making responsible innovation easier and reducing regulatory uncertainty. This will encourage investment in AI and support its adoption throughout the economy, creating jobs and helping us to do them more efficiently.

      To achieve this objective we must act quickly to remove existing barriers to innovation and prevent the emergence of new ones. This will allow AI companies to capitalise on early development successes and achieve long term market advantage.[73] By acting now, we can give UK innovators a headstart in the global race to convert the potential of AI into long term advantages for the UK, maximising the economic and social value of these technologies and strengthening our current position as a world leader in AI.[74]

    - **Increase public trust in AI** by addressing risks and protecting our fundamental values.

      Trust is a critical driver for AI adoption.[75] If people do not trust AI, they will be reluctant to use it. Such reluctance can reduce demand for AI products and hinder innovation.

---

68 The impact of regulation on innovation, Nesta, 2012.

69 Public expectations for AI governance (transparency, fairness and accountability), Centre for Data Ethics and Innovation, 2023.

70 National AI Strategy, Office for Artificial Intelligence, 2021.

71 Plan for Digital Regulation, DSIT (formerly DCMS), 2022.

72 Establishing a pro-innovation approach to regulating AI, Office for Artificial Intelligence, 2022.

73 Economic impacts of artificial intelligence, European Parliament, 2019.

74 The UK is ranked near the top of the Global AI Index, third only to the US and China. Global AI Index, Tortoise Media, 2022.

75 Trust in Artificial Intelligence: a five country study, KPMG and the University of Queensland, 2021; Evidence to support the analysis of impacts for AI governance, Frontier Economics, 2023.

Therefore we must demonstrate that our regulatory framework (described in section 3.2) effectively addresses AI risks.

- **Strengthen the UK's position as a global leader in AI.** The development of AI technologies can address some of the most pressing global challenges, from climate change to future pandemics. There is also growing international recognition that AI requires new regulatory responses to guide responsible innovation.

  The UK can play a central role in the global conversation by shaping international governance and regulation to maximise opportunities and build trust in the technology, while mitigating potential cross-border risks and protecting our democratic values. There is also an important leadership role for the UK in the development of the global AI assurance industry,[76] including auditing and safety.

  We will ensure that the UK remains attractive to innovators and investors by promoting interoperability with other regulatory approaches and minimising cross-border frictions. We will work closely with global partners through multilateral and bilateral engagements to learn from, influence and adapt as international and domestic approaches to AI regulation continue to emerge (see part 6).

34. The proposed regulatory framework does not seek to address all of the wider societal and global challenges that may relate to the development or use of AI. This includes issues relating to access to data, compute capability, and sustainability, as well as the balancing of the rights of content producers and AI developers. These are important issues to consider – especially in the context of the UK's ability to maintain its place as a global leader in AI – but they are outside of the scope of our proposals for a new overarching framework for AI regulation.

35. Government is taking wider action to ensure the UK retains its status as a global leader in AI, for example by taking forward Sir Patrick Vallance's recommendation relating to intellectual property law and generative AI.[77] This will ensure we keep the right balance between protecting rights holders and our thriving creative industries, while supporting AI developers to access the data they need.

3.2 The proposed regulatory framework

36. Our innovative approach to AI regulation uses a principles-based framework for regulators to interpret and apply to AI within their remits. This collaborative and iterative approach can keep pace with a fast moving technology that requires proportionate action to balance risk and opportunity and to strengthen the UK's position as a global leader in AI. Our agile approach aligns with Sir Patrick Vallance's Regulation for Innovation report,[78] which highlights that flexible regulatory approaches can better strike the balance between providing clarity, building trust and enabling experimentation. Our framework will provide more clarity to innovators by encouraging collaboration between government, regulators, industry and civil society.

---

[76] "Building on the UK's strengths in the professional services and technology sectors, AI assurance will also become a significant economic activity in its own right, with the potential for the UK to be a global leader in a new multi-billion pound industry." See The roadmap to an effective AI assurance ecosystem, Centre for Data Ethics and Innovation, 2021.

[77] Pro-innovation Regulation of Technologies Review: Digital Technologies, HM Treasury, 2023.

[78] Pro-innovation Regulation of Technologies Review: Digital Technologies, HM Treasury, 2023.

37. We have identified the essential characteristics of our regulatory regime. Our framework will be **pro-innovation**, **proportionate**, **trustworthy**, **adaptable**, **clear** and **collaborative**.[79]

    - **Pro-innovation**: enabling rather than stifling responsible innovation.
    - **Proportionate**: avoiding unnecessary or disproportionate burdens for businesses and regulators.
    - **Trustworthy**: addressing real risks and fostering public trust in AI in order to promote and encourage its uptake.
    - **Adaptable**: enabling us to adapt quickly and effectively to keep pace with emergent opportunities and risks as AI technologies evolve.
    - **Clear**: making it easy for actors in the AI life cycle, including businesses using AI, to know what the rules are, who they apply to, who enforces them, and how to comply with them.
    - **Collaborative**: encouraging government, regulators, and industry to work together to facilitate AI innovation, build trust and ensure that the voice of the public is heard and considered.

38. The framework, built around the four key elements below, is designed to empower our existing regulators and promote coherence across the regulatory landscape. The four key elements are:

    - Defining AI based on its unique characteristics to support regulator coordination (section 3.2.1).
    - Adopting a context-specific approach (section 3.2.2).
    - Providing a set of cross-sectoral principles to guide regulator responses to AI risks and opportunities (section 3.2.3).
        i. The principles clarify government's expectations for responsible AI and describe good governance at all stages of the AI life cycle.
        ii. The application of the principles will initially be at the discretion of the regulators, allowing prioritisation according to the needs of their sectors.
        iii. Following this initial non-statutory period of implementation, and when parliamentary time allows, we anticipate introducing a statutory duty requiring regulators to have due regard to the principles.
    - Delivering new central functions to support regulators to deliver the AI regulatory framework, maximising the benefits of an iterative approach and ensuring that the framework is coherent (section 3.2.4).

[79] These characteristics are aligned with existing principles set out in the Plan for Digital Regulation, the report of the independent Taskforce on Innovation, Growth and Regulatory Reform and with the findings of the Pro-innovation Regulation of Technologies Review: Digital Technologies, published in March 2023, which called for a proportionate and agile regulatory approach and acknowledged the importance of achieving a "balance between providing clarity and building public trust, while also enabling development, experimentation, and deployment."

3.2.1 Defining Artificial Intelligence

39. To regulate AI effectively, and to support the clarity of our proposed framework, we need a common understanding of what is meant by 'artificial intelligence'. There is no general definition of AI that enjoys widespread consensus.[80] That is why we have defined AI by reference to the two characteristics that generate the need for a bespoke regulatory response.

    - The 'adaptivity' of AI can make it difficult to explain the intent or logic of the system's outcomes:

        i. AI systems are 'trained' – once or continually – and operate by inferring patterns and connections in data which are often not easily discernible to humans.

        ii. Through such training, AI systems often develop the ability to perform new forms of inference not directly envisioned by their human programmers.

    - The 'autonomy' of AI can make it difficult to assign responsibility for outcomes:

        i. Some AI systems can make decisions without the express intent or ongoing control of a human.

40. The combination of adaptivity and autonomy can make it difficult to explain, predict, or control the outputs of an AI system, or the underlying logic by which they are generated. It can also be challenging to allocate responsibility for the system's operation and outputs. For regulatory purposes, this has potentially serious implications, particularly when decisions are made relating to significant matters, like an individual's health, or where there is an expectation that a decision should be justifiable in easily understood terms, like a legal ruling.

41. By defining AI with reference to these functional capabilities and designing our approach to address the challenges created by these characteristics, we future-proof our framework against unanticipated new technologies that are autonomous and adaptive. Because we are not creating blanket new rules for specific technologies or applications of AI, like facial recognition or LLMs, we do not need to use rigid legal definitions. Our use of these defining characteristics was widely supported in responses to our policy paper,[81] as rigid definitions can quickly become outdated and restrictive with the rapid evolution of AI.[82] We will, however, retain the ability to adapt our approach to defining AI if necessary, alongside the ongoing monitoring and iteration of the wider regulatory framework.

42. Below, we provide some illustrative examples of AI systems to demonstrate their autonomous and adaptive characteristics. While many aspects of the technologies described in these case studies will be covered by existing law, they illustrate how AI-specific characteristics introduce novel risks and regulatory implications.

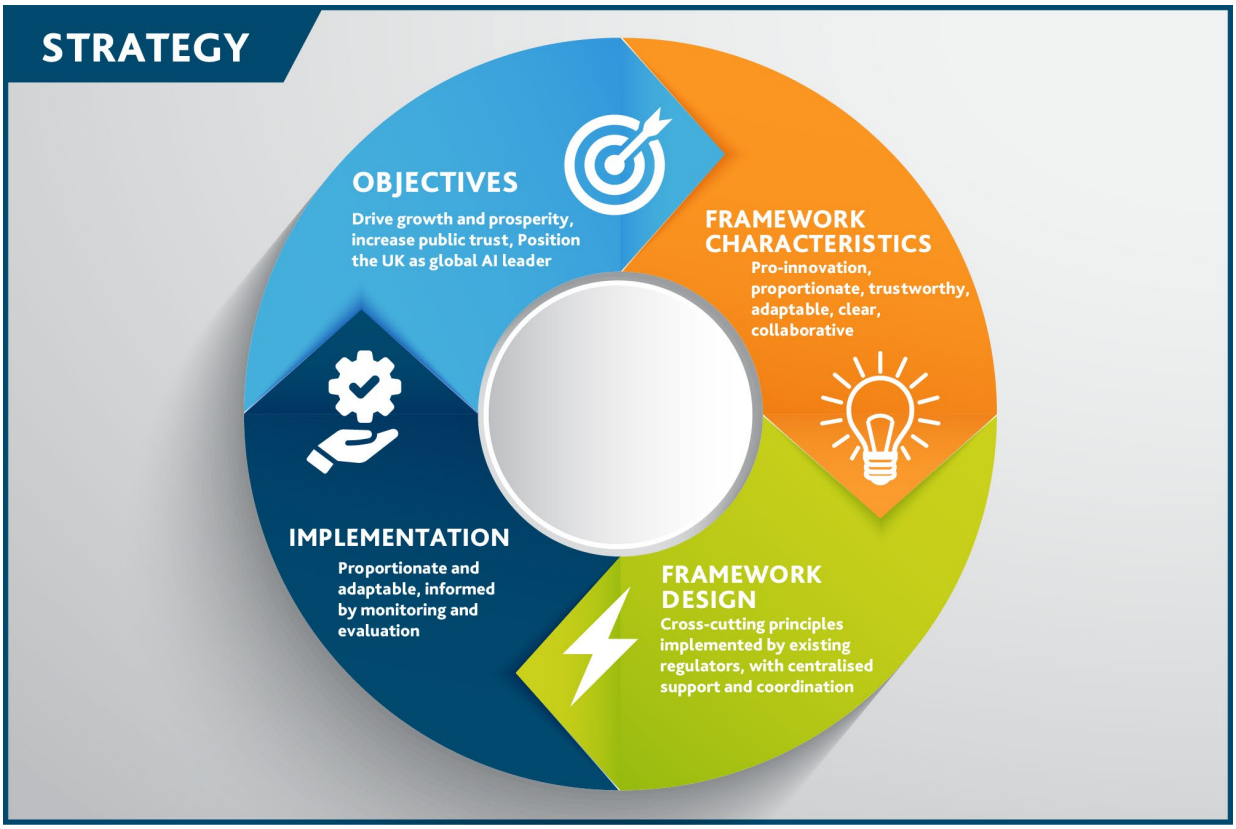
---

[80] One of the biggest problems in regulating AI is agreeing on a definition, Carnegie Endowment for International Peace, 2022.

[81] Establishing a pro-innovation approach to regulating AI, Office for Artificial Intelligence, 2022.

[82] As stated in government guidance on using AI in the public sector, we consider machine learning to be a subset of AI. While machine learning is the most widely-used form of AI and will be captured within our framework, our adaptive and autonomous characteristics ensure any current or future AI system that meets this criteria will be within scope. See A guide to using artificial intelligence in the public sector, Government Digital Service and Office for Artificial Intelligence, 2019.

Figure 1: Illustration of our strategy for regulating AI



Case study 3.1: Natural language processing in customer service chatbots

Adaptivity: Provides responses to real-time customer messages, having been trained on huge datasets to identify statistical patterns in ordinary human speech, potentially increasing personalisation over time as the system learns from each new experience.

Autonomy: Generates a human-like output based on the customer's text input, to answer queries, help customers find products and services, or send targeted updates. Operates with little need for human oversight or intervention.

Illustrative AI-related regulatory implication: Unintentional inclusion of inaccurate or misleading information in training data, producing harmful instructions or convincingly spreading misinformation.

Case study 3.2: Automated healthcare triage systems

Adaptivity: Predicts patient conditions based on the pathology, treatment and risk factors associated with health conditions from the analysis of medical datasets, patient records and real-time health data.

Autonomy: Generates information about the likely causes of a patient's symptoms and recommends potential interventions and treatments, either to a medical professional or straight to a patient.

Illustrative AI-related regulatory implication: Unclear liability for an AI triage system that provides incorrect medical advice, leading to negative health outcomes for a patient and affecting the patient's ability to obtain redress.

Case study 3.3: Text-to-image generators

Adaptivity: Uses large amounts of online content to learn how to create rich, highly specific images on the basis of a short text prompt.

Autonomy: Based on text input, these systems generate images that mimic the qualities of human-created art, with no ongoing oversight from the user.

Illustrative AI-related regulatory implication: Reproduction of biases or stereotyping in training data, leading to offensive language or content.

43. Industry, regulators, and civil society responded positively to our proposed definition, recognising that it supports our context-based and flexible approach to AI regulation. We will monitor how regulators interpret and apply adaptivity and autonomy when formulating domain-specific definitions of AI. Government will support coordination between regulators when we see potential for better alignment between their interpretations and use of our defining characteristics.

44. Active and collaborative horizon scanning will ensure that we can identify developments and emerging trends, and adapt our framework accordingly. We will convene industry, academia

and other key stakeholders to inform economy-wide horizon scanning activity. This work will build on the activity of individual regulators.

3.2.2 Regulating the use – not the technology

45. Our framework is context-specific.[83] We will not assign rules or risk levels to entire sectors or technologies. Instead, we will regulate based on the outcomes AI is likely to generate in particular applications. For example, it would not be proportionate or effective to classify all applications of AI in critical infrastructure as high risk. Some uses of AI in critical infrastructure, like the identification of superficial scratches on machinery, can be relatively low risk. Similarly, an AI-powered chatbot used to triage customer service requests for an online clothing retailer should not be regulated in the same way as a similar application used as part of a medical diagnostic process.

46. A context-specific approach allows regulators to weigh the risks of using AI against the costs of missing opportunities to do so.[84] Regulators told us that AI risk assessments should include the failure to exploit AI capabilities. For example, there can be a significant opportunity cost related to not having access to AI in safety-critical operations, from heavy industry,[85] to personal healthcare (see box 1.1). Sensitivity to context will allow the framework to respond to the level of risk in a proportionate manner and avoid stifling innovation or missing opportunities to capitalise on the social benefits made available by AI.

47. To best achieve this context-specificity we will empower existing UK regulators to apply the cross-cutting principles. Regulators are best placed to conduct detailed risk analysis and enforcement activities within their areas of expertise. Creating a new AI-specific, cross-sector regulator would introduce complexity and confusion, undermining and likely conflicting with the work of our existing expert regulators.

3.2.3 A principles-based approach

48. Existing regulators will be expected to implement the framework underpinned by five values-focused cross-sectoral principles:

    - Safety, security and robustness
    - Appropriate transparency and explainability
    - Fairness
    - Accountability and governance
    - Contestability and redress

    These build on, and reflect our commitment to, the Organisation for Economic Co-operation and Development (OECD) values-based AI principles, which promote the ethical use of AI.

49. The principles set out the key elements of responsible AI design, development and use, and will help guide businesses. Regulators will lead the implementation of the framework, for example by issuing guidance on best practice for adherence to these principles.

---

[83] See Establishing a pro-innovation approach to regulating AI, Office for Artificial Intelligence, 2022. The context-based approach received wide support in feedback received following publication of this policy paper.

[84] FIDO Direct launched as end-to-end solution to solve water loss, Smart Water Magazine, 2023.

[85] AI on-side: how artificial intelligence is being used to improve health and safety in mining, Axora, 2023.

50. Regulators will be expected to apply the principles proportionately to address the risks posed by AI within their remits, in accordance with existing laws and regulations. In this way, the principles will complement existing regulation, increase clarity, and reduce friction for businesses operating across regulatory remits.

51. A principles-based approach allows the framework to be agile and proportionate. It is in line with the Plan for Digital Regulation,[86] the findings from the independent Taskforce on Innovation, Growth and Regulatory Reform,[87] the Regulatory Horizons Council's Closing the Gap report on implementing innovation-friendly regulation,[88] and Sir Patrick Vallance's Regulation for Innovation report.[89]

52. Since publishing the AI regulation policy paper,[90] we have updated and strengthened the principles. We have:

    - Reflected stakeholder feedback by expanding on concepts such as robustness and governance. We have also considered the results of public engagement research that highlighted an expectation for principles such as transparency, fairness and accountability to be included within an AI governance framework.[91]
    - Merged the safety principle with security and robustness, given the significant overlap between these concepts.
    - Better reflected concepts of accountability and responsibility.
    - Refined each principle's definition and rationale.

Principle	**Safety, Security and Robustness**
Definition and explanation	AI systems should function in a robust, secure and safe way throughout the AI life cycle, and risks should be continually identified, assessed and managed. Regulators may need to introduce measures for regulated entities to ensure that AI systems are technically secure and function reliably as intended throughout their entire life cycle.
Rationale for	The breadth of possible uses for AI and its capacity to

[86] Plan for Digital Regulation, DSIT (formerly DCMS), 2021.

[87] The Taskforce on Innovation, Growth and Regulatory Reform independent report, 10 Downing Street, 2021. The report argues for UK regulation that is: proportionate, forward-looking, outcome-focussed, collaborative, experimental, and responsive.

[88] Closing the gap: getting from principles to practices for innovation friendly regulation, Regulatory Horizons Council, 2022.

[89] Pro-innovation Regulation of Technologies Review: Digital Technologies, HM Treasury, 2023.

[90] Establishing a pro-innovation approach to regulating AI, Office for Artificial Intelligence, 2022.

[91] The Centre for Data Ethics and Innovation (CDEI) has engaged with the public to understand their expectations for AI governance. This engagement has informed our policy development. Participants also referred to a privacy principle, which is embedded in the broader regulatory considerations as regulators and AI life cycle actors are expected to comply with the UK's data protection framework. Public expectations for AI governance (transparency, fairness and accountability), Centre for Data Ethics and Innovation, 2023.

the principle	autonomously develop new capabilities and functions mean that AI can have a significant impact on safety and security. Safety-related risks are more apparent in certain domains, such as health or critical infrastructure, but they can materialise in many areas. Safety will be a core consideration for some regulators and more marginal for others. However, it will be important for all regulators to assess the likelihood that AI could pose a risk to safety in their sector or domain, and take a proportionate approach to managing it. Additionally, AI systems should be technically secure and should reliably function as intended and described. System developers should be aware of the specific security threats that could apply at different stages of the AI life cycle and embed resilience to these threats into their systems. Other actors should remain vigilant of security issues when they interact with an AI system. We anticipate that regulators may wish to consider the National Cyber Security Centre (NCSC) principles for securing machine learning models when assessing whether AI actors are adequately prioritising security.[92] When applying this principle, regulators will need to consider providing guidance in a way that is coordinated and coherent with the activities of other regulators. Regulators' implementation of this principle may require the corresponding AI life cycle actors to regularly test or carry out due diligence on the functioning, resilience and security of a system.[93] Regulators may also need to consider technical standards addressing safety, robustness and security to benchmark the safe and robust performance of AI systems and to provide AI life cycle actors with guidance for implementing this principle in their remit.
Principle	**Appropriate transparency and explainability**
Definition and explanation	AI systems should be appropriately transparent and explainable. Transparency refers to the communication of appropriate information about an AI system to relevant people (for example, information on how, when, and for which purposes an AI system is being used). Explainability refers to the extent to which it is possible for relevant parties to access, interpret and understand the decision-making processes of an AI system.[94] An appropriate level of transparency and explainability will mean

[92] Principles for the security of machine learning, National Cyber Security Centre, 2022.

[93] For example, digital security can affect the safety of connected products such as automobiles and home appliances if risks are not appropriately managed. See Principle 1.4: Robustness, security and safety, OECD AI, 2019.

[94] Adapted from IEEE 7001-2021, Standard for Transparency of Autonomous Systems.

	that regulators have sufficient information about AI systems and their associated inputs and outputs to give meaningful effect to the other principles (e.g. to identify accountability). An appropriate degree of transparency and explainability should be proportionate to the risk(s) presented by an AI system. Regulators may need to look for ways to support and encourage relevant life cycle actors to implement appropriate transparency measures, for example through regulatory guidance. Parties directly affected by the use of an AI system should also be able to access sufficient information about AI systems to be able to enforce their rights. In applying the principle to their business processes, relevant life cycle actors may be asked to provide this information in the form and manner required by regulators, including through product labelling. Technical standards could also provide useful guidance on available methods to assess, design, and improve transparency and explainability within AI systems – recognising that consumers, users and regulators will require different information.[95]
Rationale for the principle	Transparency can increase public trust,[96] which has been shown to be a significant driver of AI adoption.[97] When AI systems are not sufficiently explainable, AI suppliers and users risk inadvertently breaking laws, infringing rights, causing harm and compromising the security of AI systems. At a technical level, the explainability of AI systems remains an important research and development challenge. The logic and decision-making in AI systems cannot always be meaningfully explained in a way that is intelligible to humans, although in many settings this poses no substantial risk. It is also true that in some cases, a decision made by AI may perform no worse on explainability than a comparable decision made by a human.[98] Future developments of the technology may pose additional challenges to achieving explainability. AI systems should display levels of explainability that are appropriate to their context, including the level of risk and consideration of what is achievable given the state of the art.

[95] For example IEEE 7001-2021 (Active Standard) describes measurable, testable levels of transparency so that autonomous systems can be objectively assessed, and levels of compliance determined; ISO/IEC TS6254 (Under development) will describe approaches and methods that can be used to achieve explainability objectives of stakeholders with regards to ML models and AI system's behaviours, outputs, and results.

[96] BritainThinks: Complete transparency, complete simplicity, CDEI and CDDO, 2021.

[97] Trust in Artificial Intelligence: a five country study, KPMG and the University of Queensland, 2021; Evidence to support the analysis of impacts for AI governance, Frontier Economics, 2023.

[98] Should AI models be explainable? That depends, Stanford Institute for Human-Centered Artificial Intelligence, 2021.

Principle	Fairness
Definition and explanation	AI systems should not undermine the legal rights of individuals or organisations, discriminate unfairly against individuals or create unfair market outcomes. Actors involved in all stages of the AI life cycle should consider definitions of fairness that are appropriate to a system's use, outcomes and the application of relevant law. Fairness is a concept embedded across many areas of law and regulation, including equality and human rights, data protection, consumer and competition law, public and common law, and rules protecting vulnerable people. Regulators may need to develop and publish descriptions and illustrations of fairness that apply to AI systems within their regulatory domain, and develop guidance that takes into account relevant law, regulation, technical standards,[99] and assurance techniques. Regulators will need to ensure that AI systems in their domain are designed, deployed and used considering such descriptions of fairness. Where concepts of fairness are relevant in a broad range of intersecting regulatory domains, we anticipate that developing joint guidance will be a priority for regulators.
Rationale for the principle	In certain circumstances, AI can have a significant impact on people's lives, including insurance offers, credit scores, and recruitment outcomes. AI-enabled decisions with high impact outcomes should not be arbitrary and should be justifiable. In order to ensure a proportionate and context-specific approach regulators should be able to describe and illustrate what fairness means within their sectors and domains, and consult with other regulators where multiple remits are engaged by a specific use case. We expect that regulators' interpretations of fairness will include consideration of compliance with relevant law and regulation, including: 1) AI systems should not produce discriminatory outcomes, such as those which contravene the Equality Act 2010 or the Human Rights Act 1998. Use of AI by public authorities should comply with the additional duties placed on them by legislation (such as the Public Sector Equality Duty). 2) Processing of personal data involved in the design, training, and use of AI systems should be compliant with requirements under the UK General Data Protection Regulation (GDPR), the

[99] For example, ISO/IEC TR 24027:2021 describes measurement techniques and methods for assessing bias in AI systems across their life cycle, especially in AI-aided decision-making.

	Data Protection Act 2018,[100] particularly around fair processing and solely automated decision-making. 3) Consumer and competition law, including rules protecting vulnerable consumers and individuals.[101] 4) Relevant sector-specific fairness requirements, such as the Financial Conduct Authority (FCA) Handbook.
Principle	**Accountability and governance**
Definition and explanation	Governance measures should be in place to ensure effective oversight of the supply and use of AI systems, with clear lines of accountability established across the AI life cycle. AI life cycle actors should take steps to consider, incorporate and adhere to the principles and introduce measures necessary for the effective implementation of the principles at all stages of the AI life cycle. Regulators will need to look for ways to ensure that clear expectations for regulatory compliance and good practice are placed on appropriate actors in the AI supply chain, and may need to encourage the use of governance procedures that reliably ensure these expectations are met. Regulator guidance on this principle should reflect that "accountability" refers to the expectation that organisations or individuals will adopt appropriate measures to ensure the proper functioning, throughout their life cycle, of the AI systems that they research, design, develop, train, operate, deploy, or otherwise use.
Rationale for the principle	AI systems can operate with a high level of autonomy, making decisions about how to achieve a certain goal or outcome in a way that has not been explicitly programmed or foreseen.[102] Establishing clear, appropriate lines of ownership and accountability is essential for creating business certainty while ensuring regulatory compliance. Doing so for actors in the AI life cycle is difficult, given the complexity of AI supply chains, as well as the adaptivity,

[100] The Data Protection and Digital Information (No. 2) Bill reforms the UK's data protection regime (Data Protection Act 2018 and the UK GDPR).

[101] Guidance on vulnerability includes: FCA guidance on vulnerable consumers, FCA, 2019; Consumer vulnerability protections, Ofgem, 2020; Vulnerable consumers, CMA, 2018.

[102] AI has the potential to learn to solve problems without human intervention instructing it to do so, or cope with situations the systems have not encountered before, producing potentially different associated risks that require clear lines of accountability and governance mechanisms to be in place. For example, see AI is learning how to create itself, MIT Technology Review, 2021.

	autonomy and opacity of AI systems. In some cases, technical standards can provide useful guidance on good practices for AI governance.[103] Assurance techniques like impact assessments can help to identify potential risks early in the development life cycle, enabling their mitigation through appropriate safeguards and governance mechanisms. Regulatory guidance should also reflect the responsibilities such life cycle actors have for *demonstrating* proper accountability and governance (for example, by providing documentation on key decisions throughout the AI system life cycle, conducting impact assessments or allowing audits where appropriate).
Principle	**Contestability and redress**
Definition and explanation	Where appropriate, users, impacted third parties and actors in the AI life cycle should be able to contest an AI decision or outcome that is harmful or creates material risk of harm. Regulators will be expected to clarify existing routes to contestability and redress, and implement proportionate measures to ensure that the outcomes of AI use are contestable where appropriate. We would also expect regulators to encourage and guide regulated entities to make clear routes (including informal channels) easily available and accessible, so affected parties can contest harmful AI outcomes or decisions as needed.
Rationale for the principle	The use of AI technologies can result in different types of harm and can have a material impact on people's lives. AI systems' outcomes may introduce risks such as the reproduction of biases or safety concerns. People and organisations should be able to contest outcomes where existing rights have been violated or they have been harmed. It will be important for regulators to provide clear guidance on this principle so that AI life cycle actors can implement it in practice. This should include clarifying that appropriate transparency and explainability are relevant to good implementation of this contestability and redress principle. The UK's initial non-statutory approach will not create new rights

[103] For example, ISO/IEC 42001 (Under development) will provide guidance for establishing, implementing and maintaining an AI management system within an organisation to develop or use AI systems responsibly. ISO/IEC 23894 (Under development) will provide guidance for establishing AI risk management principles and processes within an organisation.

	or new routes to redress at this stage.

53. We anticipate that regulators will need to issue guidance on the principles or update existing guidance to provide clarity to business. Regulators may also publish joint guidance on one or more of the principles, focused on AI use cases that cross multiple regulatory remits. We are keen to work with regulators and industry to understand the best approach to providing guidance. We expect that practical guidance will support actors in the AI life cycle to adhere to the principles and embed them into their technical and operational business processes. Regulators may also use alternative measures and introduce other tools or resources, in addition to issuing guidance, within their existing remits and powers to implement the principles.

54. Government will monitor the overall effectiveness of the principles and the wider impact of the framework.[104] This will include working with regulators to understand how the principles are being applied and whether the framework is adequately supporting innovation.

Consultation questions:

1. Do you agree that requiring organisations to make it clear when they are using AI would improve transparency?

2. Are there other measures we could require of organisations to improve transparency for AI?

3. Do you agree that current routes to contest or get redress for AI-related harms are adequate?

4. How could current routes to contest or seek redress for AI-related harms be improved, if at all?

5. Do you agree that, when implemented effectively, the revised cross-sectoral principles will cover the risks posed by AI technologies?

6. What, if anything, is missing from the revised principles?

[104] While this activity is likely to be led centrally (see part 3.3.1), this will involve continuation of the existing collaboration across government to ensure alignment with (and appropriate leveraging of) existing work being undertaken in relation to the National Cyber Strategy, UKRI work on Safe and Trusted AI, the work of the Centre for Connected and Autonomous Vehicles, the NHS AI Lab and other examples.

Case Study 3.4: Explainable AI in practice

The level of explainability needed from an AI system is highly specific to its context, including the extent to which an application is safety-critical. The level and type of explainability required will likely vary depending on whether the intended audience of the explanation is a regulator, technical expert, or lay person.

For example, a technical expert designing self-driving vehicles would need to understand the system's decision-making capabilities to test, assess and refine them. In the same context, a lay person may need to understand the decision-making process *only* in order to use the vehicle safely. If the vehicle malfunctioned and caused a harmful outcome,[105] a regulator may need information about how the system operates in order to allocate responsibility – similar to the level of explainability currently needed to hold human drivers accountable.

While AI explainability remains a technical challenge and an area of active research, regulators are already conducting work to address it. In 2021, the ICO and the Alan Turing Institute issued co-developed guidance on explaining decisions made with AI,[106] giving organisations practical advice to help explain the processes, services and decisions delivered or assisted by AI to the individuals affected by them.

The audience for an explanation of AI's outcomes will often be a regulator, who may require a higher standard of explainability depending on the risks represented by an application. The MHRA's Project Glass Box work is addressing the challenge of setting medical device requirements that take into account adequate consideration of human interpretability and its consequences for the safety and effectiveness for AI used in medical devices.[107]

105 Responsible Innovation in Self-Driving Vehicles, CDEI, 2022.

106 Explaining decisions made with AI, ICO and the Alan Turing Institute, 2021.

107 Software and AI as a Medical Device Change Programme – Roadmap, MHRA, 2022.

Case Study 3.5: What the principles mean for businesses in practice

A fictional company, *"Good AI Recruitment Limited"*, provides recruitment services that use a range of AI systems to accelerate the recruitment process, including a service that automatically shortlists candidates based on application forms. While potentially useful, such systems may discriminate against certain groups that have historically not been selected for certain positions.

After the implementation of the UK's new AI regulatory framework, the Equality and Human Rights Commission (EHRC) and the Information Commissioner Office (ICO) will be supported and encouraged to work with the Employment Agency Standards Inspectorate (EASI) and other regulators and organisations in the employment sector to issue joint guidance. The joint guidance could address the cross-cutting principles relating to fairness, appropriate transparency and explainability, and contestability and redress in the context of the use of AI systems in recruitment or employment. Such joint guidance could, for example, make things clearer and easier for Good AI Recruitment Limited by:

1. Clarifying the type of information businesses should provide when implementing such systems
2. Identifying appropriate supply chain management processes such as due diligence or AI impact assessments
3. Suggesting proportionate measures for bias detection, mitigation and monitoring
4. Providing suggestions for the provision of contestability and redress routes.

Good AI Recruitment Limited would also be able to apply a variety of tools for trustworthy AI, such as technical standards, that would supplement regulatory guidance and other measures promoted by regulators. In their published guidance regulators could, where appropriate, refer businesses to existing technical standards on transparency (e.g. IEEE 7001-2021), as well as standards on bias mitigation (e.g. ISO/IEC TR 24027:2021).

By following this guidance Good AI Recruitment Limited would be able to develop and deploy their services responsibly.

3.2.4 Our preferred model for applying the principles

55. Initially, the principles will be issued by government on a non-statutory basis and applied by regulators within their remits. We will support regulators to apply the principles using the powers and resources available to them. This initial period of implementation will provide a valuable opportunity to ensure that the principles are effective and that the wider framework is supporting innovation while addressing risks appropriately.

56. While industry has strongly supported non-statutory measures in the first instance, favouring flexibility and fewer burdens, some businesses and regulators have suggested that government

should go beyond a non-statutory approach to ensure the principles have the desired impact.[108] Some regulators have also expressed concerns that they lack the statutory basis to consider the application of the principles. We are committed to an approach that leverages collaboration with our expert regulators but we agree that we may need to intervene further to ensure that our framework is effective.

57. Following a period of non-statutory implementation, and when parliamentary time allows, we anticipate that we will want to strengthen and clarify regulators' mandates by introducing a new duty requiring them to have due regard to the principles. Such a duty would give a clear signal that we expect regulators to act and support coherence across the regulatory landscape, ensuring that the framework displays the characteristics that we have identified.[109] One of the strengths of this approach is that regulators would still be able to exercise discretion and expert judgement regarding the relevance of each principle to their individual domains.

58. A duty would ensure that regulators retain the ability to exercise judgement when applying the principles in particular contexts – benefiting from some of the flexibility expected through non-statutory implementation. For example, while the duty to have due regard would require regulators to demonstrate that they had taken account of the principles, it may be the case that not every regulator will need to introduce measures to implement every principle. In having due regard to a particular principle, a regulator may exercise their expert judgement and determine that their sector or domain does not require action to be taken. The introduction of the duty will, however, give regulators a clear mandate and incentive to apply the principles where relevant to their sectors or domains.

59. If our monitoring of the effectiveness of the initial, non-statutory framework suggests that a statutory duty is unnecessary, we would not introduce it. Similarly, we will monitor whether particular principles cannot be, or are not being, applied in certain circumstances or by specific regulators because of the interpretation of existing legal requirements or because of technical constraints. Such circumstances may require broader legislative changes. Should we decide there is a need for statutory measures, we will work with regulators to review the interaction of our principles with their existing duties and powers.

Consultation questions:

7. Do you agree that introducing a statutory duty on regulators to have due regard to the principles would clarify and strengthen regulators' mandates to implement our principles while retaining a flexible approach to implementation?

8. Is there an alternative statutory intervention that would be more effective?

[108] Following publication of our policy paper in July 2022.

[109] Pro-innovation, proportionate, adaptable, trustworthy, clear and collaborative – see paragraph 37 above.

3.2.5 The role of individual regulators in applying the principles

60. In some sectors, principles for AI governance will already exist and may even go further than the cross-cutting principles we propose. Our framework gives sectors the ability to develop and apply more specific principles to suit their own domains, where government or regulators identify these are needed.

61. The Ministry of Defence published its own AI ethical principles and policy in June 2022, which determines HM Government's approach regarding AI-enabled military capabilities. We will ensure appropriate coherence and alignment in the application of this policy through a context specific approach and thereby promote UK leadership in the employment of AI for defence purposes. Ahead of introducing any statutory duty to have due regard to our principles, and in advance of introducing other material iterations of the framework, we will consider whether exemptions are needed to allow existing regulators (such as those working in areas like national security) to continue their domain-level approach.

62. Not all principles will be equally relevant in all contexts and sometimes two or more principles may come into conflict. For example, it may be difficult to assess the fairness of an algorithm's outputs without access to sensitive personal data about the subjects of the processing. Regulators will need to use their expertise and judgement to prioritise and apply the principles in such cases, sharing information where possible with government and other regulators about how they are assessing the relevance of each principle. This collaboration between regulators and government will allow the framework to be adapted to ensure it is practical, coherent and supporting innovation.

63. In implementing the new regulatory framework we expect that regulators will:

    - Assess the cross-cutting principles and apply them to AI use cases that fall within their remit.
    - Issue relevant guidance on how the principles interact with existing legislation to support industry to apply the principles. Such guidance should also explain and illustrate what compliance looks like.
    - Support businesses operating within the remits of multiple regulators by collaborating and producing clear and consistent guidance, including joint guidance where appropriate.

64. Regulators will need to monitor and evaluate their own implementation of the framework and their own effectiveness at regulating AI within their remits. We understand that there may be AI-related risks that do not clearly fall within the remits of the UK's existing regulators.[110] Not every new AI-related risk will require a regulatory response and there is a growing ecosystem of tools for trustworthy AI that can support the application of the cross-cutting principles. These are described further in part four.

65. Where prioritised risks fall within a gap in the legal landscape, regulators will need to collaborate with government to identify potential actions. This may include identifying iterations to the

[110] For example, there are only six specific legal services activities that are overseen by regulators in the legal services sector. These "reserved legal activities" are set out in the Legal Services Act, HM Government, 2007 and can only be carried out by those who are authorised (or exempt). AI-driven systems could offer other services like writing wills or contracts (which many might consider to be legal services) without being subject to oversight from legal services regulators.

framework such as changes to regulators' remits, updates to the Regulators' Code,[111] or additional legislative interventions. Our approach benefits from our strong sovereign parliamentary system, which reliably allows for the introduction of targeted and proportionate measures in response to emerging issues, including by adapting existing legislation if necessary.[112]

66. Sir Patrick Vallance's review has highlighted that rushed attempts to regulate AI too early would risk stifling innovation.[113] Our approach aligns with this perspective. We recognise the need to build a stronger evidence base before making decisions on statutory interventions. In doing so, we will ensure that we strike the right balance between retaining flexibility in our iterative approach and providing clarity to businesses. As detailed in section 3.3.1, we will deliver a range of central functions, including horizon scanning and risk monitoring, to identify and respond to situations where prioritised risks are not adequately covered by the framework, or where gaps between regulators' remits are negatively impacting innovation.

Case study 3.6: Responding to regulatory policy challenges – self-driving vehicles

Some aspects of a new AI use case may sit outside regulators' existing remits, meaning they do not have a mandate to address specific harms or support a new product to enter the market.

The advent of self-driving vehicles highlighted such a regulatory and policy challenge. Where sophisticated AI-enabled software is capable of performing the designated driving task, existing regulatory structures – where responsibility for road safety is achieved by licensing human drivers – are not fit for purpose. This creates uncertainty regarding the development and deployment of self-driving vehicles that cannot be addressed by regulators alone.

To achieve the government's ambition to 'make the UK one of the best places in the world to develop and deploy self-driving vehicles technology',[114] manufacturers need clarity about the regulatory landscape they are operating in and the general public needs to have confidence in the safety, fairness and trustworthiness of these vehicles.

The government published its Connected & Automated Mobility 2025 report[115] to address this challenge, describing how the ecosystem could be adapted to spur innovation and secure the economic and social benefits of this technology.

The work of the UK's Centre for Connected and Autonomous Vehicles is an example of government acting to identify regulatory gaps, develop policy and build UK capabilities. A central monitoring and evaluation function, described below, will identify and assess gaps in the regulatory ecosystem that could stifle AI innovation so that government can take action to address them.

111 Regulators' Code, Office for Product Safety and Standards, 2014.
112 What is the UK Constitution?, The Constitution Unit, University College London, 2023.
113 Pro-innovation regulation of technologies review: digital technologies, HM Treasury, 2023.
114 UK on the cusp of a transport revolution, Department for Transport, 2021.
115 Connected & Automated Mobility 2025, Department for Transport, 2022.

3.2.6 Guidance to regulators on applying the principles

67. The proposed regulatory framework is dependent upon the implementation of the principles by our expert regulators. This regulator-led approach has received broad support from across industry, with stakeholders acknowledging the importance of the sector-specific expertise held by individual regulators. We expect regulators to collaborate proactively to achieve the best outcomes for the economy and society. We will work with regulators to monitor the wider framework and ensure that this collaborative approach to implementation is effective. If improvements are needed, including interventions to drive stronger collaboration across regulators, we will take further action.

68. Our engagement with regulators and industry highlighted the need for central government to support regulators. We will work with regulators to develop guidance that helps them implement the principles in a way that aligns with our expectations for how the framework should operate. Existing legal frameworks already mandate and guide regulators' actions. For example, nearly all regulators are bound by the Regulators' Code[116] and all regulators – as public bodies – are required to comply with the Human Rights Act.[117] Our proposed guidance to regulators will seek to ensure that when applying the principles, regulators are supported and encouraged to:

    - Adopt a proportionate approach that promotes growth and innovation by focusing on the risks that AI poses in a particular context.
    - Consider proportionate measures to address prioritised risks, taking into account cross-cutting risk assessments undertaken by, or on behalf of, government.
    - Design, implement and enforce appropriate regulatory requirements and, where possible, integrate delivery of the principles into existing monitoring, investigation and enforcement processes.
    - Develop joint guidance, where appropriate, to support industry compliance with the principles and relevant regulatory requirements.
    - Consider how tools for trustworthy AI like assurance techniques and technical standards can support regulatory compliance.
    - Engage proactively and collaboratively with government's monitoring and evaluation of the framework.

[116] Regulators' Code, Office for Product Safety and Standards, 2014.

[117] Human Rights Act, HM Government, 1998

Case Study 3.7: What this means for businesses

A fictional company, "*AI Fairness Insurance Limited*", has delayed the deployment of a new AI application as – under the current patchwork of relevant regulatory requirements – it has been challenging to identify appropriate compliance actions for AI-driven insurance products.

Following implementation of the UK's new pro-innovation framework to regulate AI, we could expect to see joint guidance produced collaboratively by the Information Commissioner's Office (ICO), Equality and Human Rights Commission (EHRC), Financial Conduct Authority (FCA) and other relevant regulatory authorities. This would provide greater clarity on the regulatory requirements relevant to AI as well as guidance on how to satisfy those requirements in the context of insurance and consumer-facing financial services.

Under the proposed regulatory framework, AI Fairness Insurance Limited could be supported by new or updated guidance issued by regulators to address the AI regulatory principles. The company may also be able to follow joint regulatory guidance, issued as a result of collaboration between regulators, and use a set of tools provided by regulators, such as template risk assessments and transparency measures, and relevant technical standards (e.g. international standards for transparency and bias mitigation). The collaboration between regulators and focus on practical implementation measures will guide the responsible deployment of AI Fairness Insurance Limited's AI product by making it easier for the company to navigate the regulatory landscape and address specific risks such as discrimination.

69. Further details about the implementation of the regulatory framework will be provided through an AI regulation roadmap which will be published alongside the government response to the consultation on this white paper.

3.3.1 New central functions to support the framework

70. Government has a responsibility to make sure the regulatory framework operates proportionately and supports innovation. Feedback to our proposals from businesses has been clear that the current patchwork of regulation, with relatively little in the way of central coordination or oversight, will create a growing barrier to innovation if left unaddressed. Responses from over 130 organisations and individuals to our 2022 policy paper highlighted the need for a greater level of monitoring and coordination to achieve the coherence and improved clarity we need to support innovation. Businesses, particularly small to medium sized enterprises, noted that regulatory coordination could improve business certainty and investment, resulting in more and better jobs in the sector.

71. Government therefore intends to put mechanisms in place to coordinate, monitor and adapt the framework as a whole. Further detail on these functions is set out below. Enhanced monitoring activity will allow us to take a structured approach to gathering feedback from industry on the impact of our regime as it is introduced. These mechanisms will supplement and support the work of regulators, without undermining their independence. Equally, such mechanisms are not intended to duplicate existing activities.

72. Delivering some functions centrally provides government with an overarching view of how the framework is working, where it is effective and where it may need improving. A central suite of functions will also facilitate collaboration by bringing together a wide range of interested parties, including regulators, international partners, industry, civil society organisations such as trade unions and advocacy groups, academia and the general public. Our engagement with these groups has highlighted the need for our proposed central functions. We will continue to convene a wide range of stakeholders to ensure that we hear the full spectrum of viewpoints. This breadth of engagement and collaboration will be integral to government's ability to monitor and improve the framework. The functions will identify and support opportunities for further coordination between regulators, resulting in greater clarity for businesses and stronger consumer trust.

73. We have identified a set of functions that will drive regulatory coherence and support regulators to implement the pro-innovation approach that we have outlined. These functions have been informed by our discussions with industry, research organisations, and regulators following the publication of the AI policy paper.

Box 3.1: Functions required to support implementation of the framework

Monitoring, assessment and feedback[118]

Activities

- Develop and maintain a central monitoring and evaluation (M&E) framework to assess cross-economy and sector-specific impacts of the new regime.
- Ensure appropriate data is gathered from relevant sources – for example, from industry, regulators, government and civil society – and considered as part of the overall assessment of the effectiveness of the framework.
- Support and equip regulators to undertake internal M&E and find ways to support regulators' contributions to the central M&E function.
- Monitor the regime's overall effectiveness including the extent to which it is proportionate and supporting innovation.
- Provide advice to ministers on issues that may need to be addressed to improve the regime, including where additional intervention may be required to ensure that the framework remains effective as the capability of AI and the state of the art develops.

Rationale

This function is at the heart of our iterative approach. We need to know whether the framework is working – for example, whether it is able to respond to and mitigate prioritised risks and whether the framework is actively supporting innovation – and we need the ability to spot issues quickly so we can adapt the framework in response.

M&E needs to be undertaken centrally to determine whether the regime as a

[118] See Box 3.3.

whole is delivering against our objectives. M&E will assess whether our regime is operating in a way that is pro-innovation, clear, proportionate, adaptable, trustworthy and collaborative.

Our engagement with industry, regulators, and civil society has shown us the importance of establishing a feedback loop to measure the effectiveness of the framework. We will ensure mechanisms are in place to gather evidence and insights to inform policy design.

Support coherent implementation of the principles

Activities

- Develop and maintain central regulatory guidance to support regulators in the implementation of the principles.
- Identify barriers that prevent regulators from effectively implementing the principles, such as:
    - Scope of regulatory remit.
    - Insufficient regulatory powers.
    - Insufficient regulatory capabilities.
- Identify conflicts or inconsistencies in the way the principles are interpreted across regulatory remits, and assess the impact this is having on innovation. Some variation across regulators' approaches to implementation is to be expected and encouraged, given the context-based approach that we are taking.
- Work with regulators to resolve discrepancies that are having a significant impact on innovation, and share learning and best practice.
- Monitor and assess the ongoing relevance of the principles themselves.

Rationale

Businesses have noted that, within a context-based regulatory framework, an appropriate degree of central leadership is needed to ensure coherence. To be effective, this function must be performed centrally, as the whole regulatory landscape needs to be considered to:

- Ensure that, as far as necessary to support innovation, regulators interpret and implement the principles in a coherent way.
- Effectively monitor how well the principles are being implemented, as well as their ongoing relevance.

This function will play a central part in delivering a regulatory regime that is:

- Clear: by making it easier for businesses working across regulatory remits (for example, by supporting the development of joint regulatory guidance

where appropriate).

- Proportionate and pro-innovation: as it allows government to find and prevent any application of the principles that has a disproportionate or harmful impact on innovation.
- Adaptable and trustworthy: as it forms part of the feedback loop established by the M&E function to understand how well the regime operates and whether it should be changed.
- Collaborative: by encouraging cross-government cooperation aligned to the principles.

Cross-sectoral risk assessment[119]

Activities

- Develop and maintain a cross-economy, society-wide AI risk register to support regulators' internal risks assessments.
- Monitor, review and re-prioritise known risks.
- Identify and prioritise new and emerging risks (working with the horizon scanning function).
- Work with regulators to clarify responsibilities in relation to new risks or areas of contested responsibility.
- Support join-up between regulators on AI-related risks that cut across remits and facilitate issuing of joint guidance where appropriate.
- Identify where risks are not adequately covered.
- Share risk enforcement best practices.

Rationale

Stakeholders have emphasised that a cross-sectoral assessment of risk is required to ensure that any new risks can be addressed and do not fall in any gaps between regulator remits. To be effective, this function must be performed centrally, as risks need to be considered across the whole economy to:

- Encourage regulators to take a coherent approach to assessing AI-related risks.
- Ensure risks do not fall between regulatory gaps and that appropriate action is taken where cross-sector risks do not have an obvious 'home' within a single regulatory remit.

A centrally delivered risk function will ensure that the framework's approach to risk

[119] See Box 3.2.

is informed by a cross-sector, holistic viewpoint. A cross-cutting approach to risk allows a proportionate but effective response.

This function will play a central part in delivering a regulatory regime that is:

- Clear: by making it easier for businesses working across sectors.
- Proportionate: by ensuring an appropriate response to cross-sector risks.
- Trustworthy: by making sure priority AI-related risks are being addressed.
- Collaborative: by allowing important actors – such as frontier researchers, civil society, international partners and regulators – to be convened to engage in focused, prioritised discussions on AI-related risks.

Support for innovators (including testbeds and sandboxes as detailed in section 3.3.4)

Activities

- Remove barriers to innovation by assisting AI innovators to navigate regulatory complexity and get their product to market while minimising legal and compliance risk (drawing on the expertise of all relevant regulators).
- Identify cross-cutting regulatory issues that are having real-world impacts and stifling innovation, and identify opportunities for improvement to our regulatory framework.

Rationale

We want to make it easy for innovators to navigate the regulatory landscape. Businesses have noted that tools such as regulatory sandboxes can help innovators to navigate the regulatory landscape. Central commissioning or delivery of the sandbox or testbed will also enable information and insights generated from this work to directly inform our implementation of the overall regulatory framework.

This function will play a central part in delivering a regulatory regime that is:

- Clear: by making it easier for businesses working across sectors.
- Adaptable and trustworthy: as it forms an important part of the feedback loop to understand how well the regime is functioning and how it should iterate.

To support innovators, we will take forward Sir Patrick Vallance's recommendation for a multi-regulator AI sandbox to be established[120] (see section 3.3.4 for more detail).

[120] Pro-innovation Regulation of Technologies Review: Digital Technologies, HM Treasury, 2023.

Education and awareness

Activities

- Provide guidance to businesses seeking to navigate the AI regulatory landscape.
- Raise awareness and provide guidance to consumers and the general public to ensure that these groups are empowered and encouraged to engage with the ongoing monitoring and iteration of the framework.
- Encourage regulators to promote awareness campaigns to educate consumers and users on AI regulation and risks.[121]

Rationale

To be effective, this function must be performed centrally, as the whole regulatory landscape needs to be considered to provide useful guidance to businesses and consumers on navigating it. This will ensure that businesses and consumers are able to contribute to the monitoring and evaluation of the framework and its ongoing iteration.

This function will help deliver a regulatory regime that is:

- Clear: by helping businesses working across sectors to navigate the regulatory landscape.
- Trustworthy: by increasing awareness of the framework and its requirements among consumers and businesses.
- Collaborative: by educating and raising awareness to empower businesses and consumers to participate in the ongoing evaluation and iteration of the framework.
- Pro-innovation: by enhancing trust, which is shown to increase AI adoption.

Horizon scanning

Activities

- Monitor emerging trends and opportunities in AI development to ensure that the framework can respond to them effectively.
- Proactively convene industry, frontier researchers, academia and other key stakeholders to establish how the AI regulatory framework could support the UK's AI ecosystem to maximise the benefits of emerging opportunities

[121] The Centre for Data Ethics and Innovation (CDEI) Public attitudes report states that the public continue to have limited awareness of AI, with knowledge mainly of low-risk use cases that are already in use but showing low familiarity with more complex AI applications. Public expectations for AI governance (transparency, fairness and accountability), Centre for Data Ethics and Innovation, 2023.

whilst continuing to take a proportionate approach to AI risk.

- Support the risk assessment function to identify and prioritise new and emerging AI risks, working collaboratively with industry, academia, global partners, and regulators.

Rationale

This function will support horizon-scanning activities in individual regulators but a central function is also necessary. As stakeholders have highlighted, an economy-wide view is required to anticipate opportunities that emerge across the landscape, particularly those that cut across regulatory remits or fall in the gaps between them.

This function will help deliver a regulatory regime that is:

- Adaptable: by identifying emerging trends to enable intelligent, coordinated adaptation of the regulatory framework.
- Collaborative: by convening partners including frontier researchers, industry, civil society, international partners and regulators, to identify emerging trends.
- Trustworthy: by ensuring that our regulatory framework is able to adapt in the face of emerging trends.

Ensure interoperability with international regulatory frameworks

Activities

- Monitor alignment between UK principles and international approaches to regulation, assurance and/or risk management, and technical standards.
- Support cross-border coordination and collaboration by identifying opportunities for regulatory interoperability.

Rationale

To be effective, this function must be performed centrally. The whole regulatory landscape needs to be considered to understand how well the UK framework aligns with international jurisdictions. The impact of international alignment on innovation and adoption of AI in the UK is a key concern for businesses. The central oversight and monitoring of the global alignment of the framework will support UK engagement with like-minded international partners on AI regulation, building our influence in AI.

This function will play a central part in delivering a regulatory regime that is:

- Pro-innovation: by ensuring that UK innovators can trade internationally and UK companies can attract overseas investment.
- Collaborative: by fostering close cooperation with international partners.

- Proportionate: by making sure the framework is sufficiently aligned with international approaches to maximise market access and business opportunities without imposing unnecessary burdens that could stifle innovation or otherwise negatively impact on international trade and/or investment in AI in the UK.
- Adaptable: as it forms an important part of the feedback loop to understand how well the regime is functioning and how it should iterate.

Consultation questions:

9. Do you agree that the functions outlined in Box 3.1 would benefit our AI regulation framework if delivered centrally?

10. What, if anything, is missing from the central functions?

11. Do you know of any existing organisations who should deliver one or more of our proposed central functions?

12. Are there additional activities that would help **businesses** confidently innovate and use AI technologies?

12.1. If so, should these activities be delivered by government, regulators or a different organisation?

13. Are there additional activities that would help **individuals and consumers** confidently use AI technologies?

13.1. If so, should these activities be delivered by government, regulators or a different organisation?

14. How can we avoid overlapping, duplicative or contradictory guidance on AI issued by different regulators?

Box 3.2: Supporting coherence in risk assessment

Why?

Many AI risks do not fall neatly into the remit of one individual regulator and they could go unaddressed if not monitored at a cross-sector level. A central, cross-economy risk function will also enable government to monitor future risks in a rigorous, coherent and balanced way. This will include 'high impact but low probability' risks such as existential risks posed by artificial general intelligence or AI biosecurity risks.

A pro-innovation approach to regulation involves tolerating a certain degree of risk rather than intervening in all cases. Government needs the ability to assess and

prioritise AI risks, ensuring that any intervention is proportionate and consistent with levels of risk mitigation activity elsewhere across the economy or AI life cycle.

Establishing a central risk function will bring coherence to the way regulators and industry think about AI risk. It will also foster collaboration between government, regulators, industry and civil society to provide clarity for businesses managing AI risk across sectors.

What?

The central risk function will identify, assess, prioritise and monitor cross-cutting AI risks that may require government intervention.

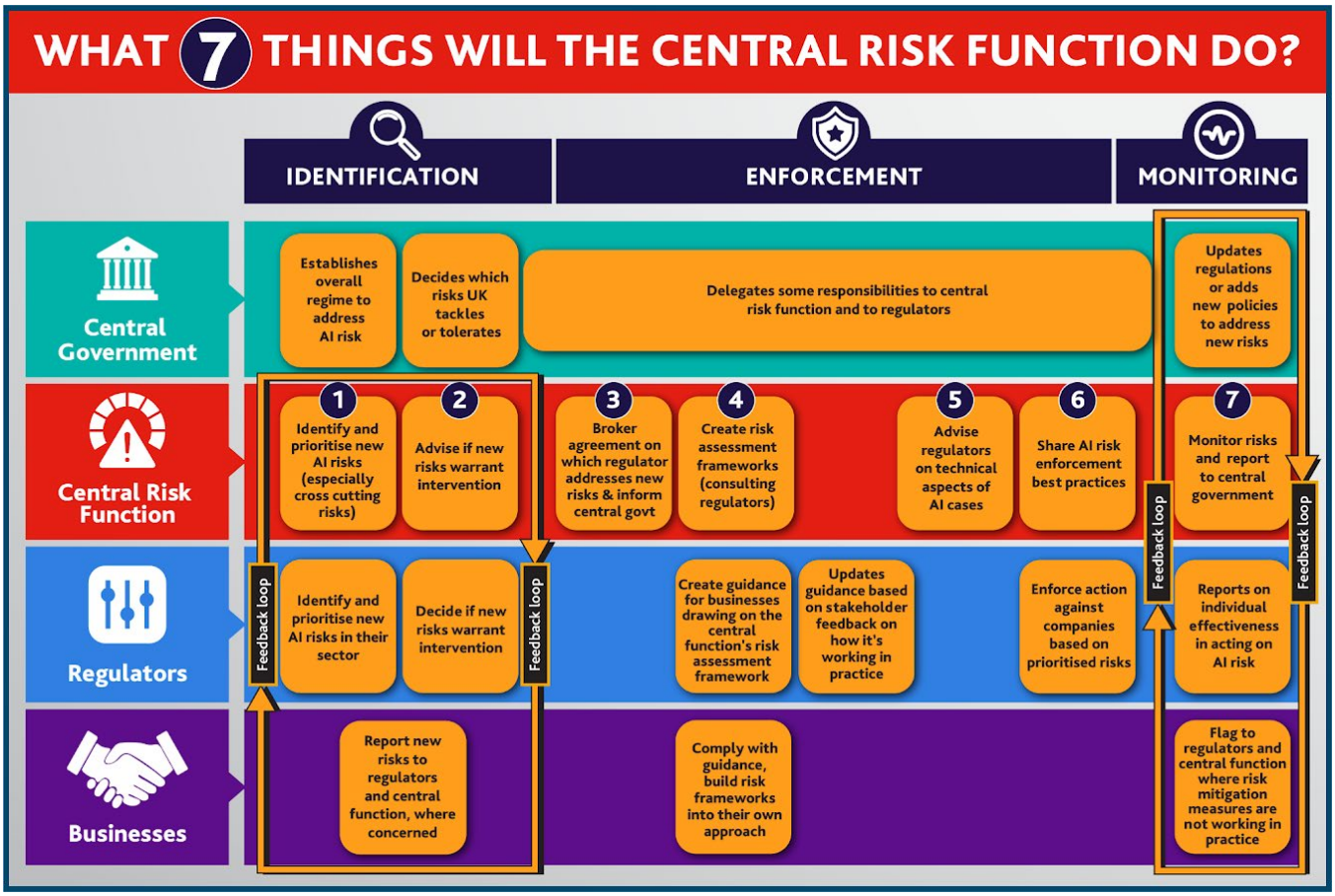
How?

The central risk function will bring together cutting-edge knowledge from industry, regulators, academia and civil society – including skilled computer scientists with a deep technical understanding of AI.

Given the importance of risk management expertise, we will seek inspiration and learning from sectors where operational risk management is highly developed. This will include looking for examples of how failures and near misses can be recorded and used to inform good practice.

Regulators will have a key role in designing the central risk framework and ensuring alignment with their existing practices. Where a risk that has been prioritised for intervention falls outside of any existing regulator's remit, the central risk function will identify measures that could be taken to address the gap (for example, updates to regulatory remits). The central risk function will also support smaller regulators that lack technical AI expertise to better understand AI risks.

Figure 2: Central risks function activities



Box 3.3: How a central monitoring and evaluation (M&E) function enables a proportionate, adaptable approach

Why?

We will need to monitor the implementation of the framework closely to make sure that it is working as designed. We will monitor the regime to ensure it is pro-innovation, proportionate, adaptable, trustworthy, clear and collaborative – our desired characteristics.

What?

The central M&E function will gather evidence and feedback from a range of sources and actors in the ecosystem. For example, effective M&E of the whole framework is likely to require input and data from industry, regulators, civil society, academia, international partners and the general public. Insights from regulatory sandboxes and testbeds as well as wider monitoring of the AI ecosystem as a whole will also be valuable.[122]

Government's ability to access reliable, comprehensive data and insights for the purposes of monitoring the AI regulatory framework will be closely related to our work raising awareness and educating businesses and consumers on AI-related issues. It is important for our M&E data to be drawn from a wide range of sources, reflecting the full spectrum of views and including seldom heard voices from the general public. Raising awareness and educating stakeholder groups will help to ensure that the broader conversation is inclusive, informed and rigorous.

We will develop and monitor metrics that demonstrate whether the framework is working as intended. For example, the central M&E function will look at the effectiveness of the framework in mitigating unacceptable risks and assess whether the implementation of the principles by regulators is disproportionate or negatively affecting innovation.

Insights from the M&E function will contribute to the adaptability of our framework by enabling government to identify opportunities for improvement so we can benefit fully from the flexibility we have built into our approach. Such iteration could include removing or amending existing regulation as well as updating the AI regulatory framework itself.

How?

The range, sources and quality of the data that informs our monitoring and evaluation of the framework will be critical.

The M&E function will identify the metrics and data sources to help us measure how well the regime is working, both in terms of the framework's ability to mitigate

[122] For example, stakeholders have outlined proposals for governments' roles in monitoring the wider AI ecosystem as a means of addressing challenging policy issues. See Why and how governments should monitor AI development, Whittlestone and Clark, 2021.

risk but also to ensure that it is supporting innovation. It will bring together a wide range of views including industry, civil society groups and academia.

Crucially, we will work with regulators to identify how their work – including data collected from their own regulatory activities – can support our central M&E function in order to ensure the best outcomes for the whole economy.

Consultation questions:

15. Do you agree with our overall approach to monitoring and evaluation?

16. What is the best way to measure the impact of our framework?

17. Do you agree that our approach strikes the right balance between supporting AI innovation; addressing known, prioritised risks; and future-proofing the AI regulation framework?

74. It is important to have the right architecture in place to oversee the delivery of the central functions described above. The AI ecosystem already benefits from a range of organisations with extensive expertise in regulatory issues. Ground-breaking coordination initiatives like the Digital Regulation Cooperation Forum (DRCF) play a valuable role in enhancing regulatory alignment and fostering dialogue on digital issues across regulators. However, the DRCF was not created to support the delivery of all the functions we have identified or the implementation of our proposed regulatory framework for AI.

75. Government will initially be responsible for delivering the central functions described above, working in partnership with regulators and other key actors in the AI ecosystem to leverage existing activities where possible. This is aligned with our overall iterative approach and enables system-wide review of the framework. We recognise that there may be value in a more independent delivery of the central functions in the longer term.

76. Where relevant activities are already undertaken by organisations either within or outside of government, the primary role of the central functions will be to leverage these activities and assess their effectiveness. Where this is not the case – for example, where new bespoke capabilities are needed to monitor and evaluate the operation of the framework as a whole – these functions will initially be established in government.

Case study 3.8: Building on a strong foundation of regulatory coordination

The growth of digital technologies requires regulators to coordinate and act cohesively. The Digital Regulation Cooperation Forum (DRCF) has published its vision for a joined-up approach to digital regulation. It conducts cross-regulator horizon scanning for future technology and has issued detailed discussion papers on the benefits, harms and auditing of algorithms.

Regulators are also exploring ways to provide simpler 'shop fronts' for those they regulate, with the NHS AI and Digital Regulations Service offer already robustly tested with end users and now widely available.[123] DRCF regulators have a multi-agency advice service for digital innovators pilot underway, supported by government's Regulators' Pioneer Fund,[124] which aims to make it easier for firms operating across digital regulatory boundaries to do business.

Existing regulatory forums may need to be supplemented or adapted to successfully implement the cross-cutting principles. We will work in partnership with existing bodies as well as industry to improve and enhance regulatory coordination.

77. We are deliberately taking an iterative approach to the delivery of the regulatory framework and we anticipate that the model for providing the central functions will develop over time. We will identify where existing structures may need to be supplemented or adapted. In particular, we are focused on understanding:

    - Whether existing regulatory forums could be expanded to include the full range of regulators involved in the regulation of AI or whether additional mechanisms are needed.
    - What additional expertise government may need to support the implementation and monitoring of the principles, including the potential role that could be played by established advisory bodies.
    - The most effective way to convene input from across industry and consumers to ensure a broad range of opinions.

78. Government, in fulfilling the regulatory central functions and overseeing the framework, will benefit from engaging external expertise to gather insights and advice from experts in industry, academia and civil society. The AI Council has been an important source of expertise over the last three years, advising government on the development of the National AI Strategy as well as our approach to AI governance. As we enter a new phase we will review the role of the AI Council and consider how best to engage expertise to support the implementation of the regulatory framework.

79. As the regulatory framework evolves and we develop a clearer understanding of the system-level functions that are needed, we will review the operational model outlined above. In particular, we will consider if a government unit is the most appropriate mechanism for delivering the central functions in the longer term or if an independent body would be more effective.

Consultation questions:

18. Do you agree that regulators are best placed to apply the principles and government is best placed to provide oversight and deliver central functions?

[123] AI and Digital Regulations Service, Care Quality Commission, Health Research Authority, Medicines and Healthcare Products Regulatory Agency, National Institute for Health and Care Excellence, 2023.

[124] Enabling innovation – piloting a multi-agency advice service for digital innovators, Regulators' Pioneer Fund, 2022 (an ICO-led project).

3.3.2 Government's role in addressing accountability across the life cycle

80. The clear allocation of accountability and legal responsibility is important for effective AI governance. Legal responsibility for compliance with the principles should be allocated to the actors in the AI life cycle best able to identify, assess and mitigate AI risks effectively. Incoherent or misplaced allocation of legal responsibility could hinder innovation or adoption of AI.

81. However, AI supply chains can be complex and opaque, making effective governance of AI and supply chain risk management difficult. Inappropriate allocation of AI risk, liability, and responsibility for AI governance throughout the AI life cycle and within AI supply chains could impact negatively on innovation. For example, inappropriate allocation of liability to a business using, but not developing, AI could stifle AI adoption. Similarly, allocating too much responsibility to businesses developing foundation models, on the grounds that these models could be used by third parties in a range of contexts, would hamper innovation.

82. We recognise the need to consider which actors should be responsible and liable for complying with the principles, which may not be the same actors who bear the burden under current legal frameworks. For example, data protection law differentiates between data controllers and data processors. Similarly, product safety laws include the concepts of producers and distributors. In the context of those specific legal frameworks, liability for compliance with various existing legal obligations is allocated by law to those identified supply chain actors. It is not yet clear how responsibility and liability for demonstrating compliance with the AI regulatory principles will be or should ideally be, allocated to existing supply chain actors within the AI life cycle.

83. We are not proposing to intervene and make changes to life cycle accountability at this stage. It is too soon to make decisions about liability as it is a complex, rapidly evolving issue which must be handled properly to ensure the success of our wider AI ecosystem. However, to further our understanding of this topic we will engage a range of experts, including technicians and lawyers. It may become apparent that current legal frameworks, when combined with implementation of our AI principles by regulators, will allocate legal responsibility and liability across the supply chain in a way that is not fair or effective. We would consider proportionate interventions to address such issues which could otherwise undermine our pro-innovation approach to AI regulation. Our agile approach benefits our sovereign parliamentary system's reliable ability to introduce targeted measures – for example by amending existing legislation if necessary – in response to new evidence.[125]

84. Tools for trustworthy AI like assurance techniques and technical standards can support supply chain risk management. These tools can also drive the uptake and adoption of AI by building justified trust in these systems, giving users confidence that key AI-related risks have been identified, addressed and mitigated across the supply chain. For example, by describing measures that manufacturers should take to ensure the safety of AI systems, technical standards can provide reassurance to purchasers and users of AI systems that appropriate safety-focused measures have been adopted, ultimately encouraging adoption of AI.

85. Our evaluation of the framework will assess whether the legal responsibility for AI is effectively and fairly distributed. As we implement the framework, we will continue our extensive engagement to gather evidence from regulators, industry, academia, and civil society on its impact on different actors across the AI life cycle. This will allow us to monitor the effects of our

[125] What is the UK constitution? The Constitution Unit, University College London, 2023.

framework on actors across the AI supply chain on an ongoing basis. We will need a particular focus on foundation models given the potential challenges they pose to life cycle accountability, especially when available as open-source. By centrally evaluating whether there are adequate measures for AI accountability, we can assess the need for further interventions into AI liability across the whole economy and AI life cycle.

Consultation questions:

L1. What challenges might arise when regulators apply the principles across different AI applications and systems? How could we address these challenges through our proposed AI regulatory framework?

L2.1. Do you agree that the implementation of our principles through existing legal frameworks will fairly and effectively allocate legal responsibility for AI across the life cycle?

L.2.2. How could it be improved, if at all?

L3. If you work for a business that develops, uses, or sells AI, how do you currently manage AI risk including through the wider supply chain? How could government support effective AI-related risk management?

3.3.3 Foundation models and the regulatory framework

86. Foundation models are an emerging type of general purpose AI that are trained on vast quantities of data and can be adapted to a wide range of tasks. The fast-paced development of foundation models brings novel challenges for governments seeking to regulate AI. Despite high levels of interest in the topic, the research community has not found a consensus on how foundation models work, the risks they pose or even the extent of their capabilities.[126]

87. Foundation models have been described as paradigm shifting and could have significant impacts on society and the economy.[127] They can be used for a wide variety of purposes and deployed in many already complex ecosystems. Given the widely acknowledged transformative potential of foundation models, we must give careful attention to how they might interact with our proposed regulatory framework. Our commitment to an adaptable, proportionate approach presents a clear opportunity for the UK to lead the global conversation and set global norms for the future-proof regulation of foundation models.

88. There is a relatively small number of organisations developing foundation models. Some organisations exercise close control over the development and distribution of their foundation models. Other organisations take an open-source approach to the development and distribution of the technology. Open-source models can improve access to the transformational power of foundation models, but can cause harm without adequate guardrails.[128] The variation in organisational approaches to developing and supplying foundation models introduces a wide range of complexities for the regulation of AI. The potential opacity of foundation models means

[126] On the opportunities and risks of foundation models, Bommasani et al., 2022; Expert opinion: Regulating AI in Europe, Edwards, 2022.
[127] Taxonomy of Risks posed by Language Models, Weidinger et al., 2022.
[128] The value chain of general-purpose AI, Ada Lovelace Institute, 2023.

that it can also be challenging to identify and allocate accountability for outcomes generated by AI systems that rely on or integrate them.

89. Our proposed framework considers the issues raised by foundation models in light of our life cycle accountability analysis, outlined in section 3.3.2 above. Given the small number of organisations supplying foundation models and a proportionately larger number of businesses integrating or otherwise deploying foundation models elsewhere in the AI ecosystem, we recognise the important role of tools for trustworthy AI, including assurance techniques and technical standards.

90. The proposed central functions described in section 3.3.1 will play an important role in informing our approach to regulating foundation models. The central risk function's proactive, rigorous monitoring of risks associated with foundation models and the horizon scanning function's identification of related opportunities will be critical to ensuring that we strike the balance needed as part of our proportionate, pro-innovation regulatory approach. It will be crucial to ensure that the proposed monitoring and evaluation function has access to the technical skills and capabilities needed to assess the impact that our framework has on the opportunities and risks presented by foundation models.

91. We recognise that industry, academia, research organisations and global partners are looking for ways to address the challenges related to the regulation of foundation models.[129] For example, we know that developers of foundation models are exploring ways to embed alignment theory into their models. This is an important area of research, and government will need to work closely with the AI research community to leverage insights and inform our iteration of the regulatory framework. Our collaborative, adaptable framework will draw on the expertise of those researchers and other stakeholders as we continue to develop policy in this evolving area.

92. The UK is committed to building its capabilities in foundation models. Our Foundation Model Taskforce announced in the Integrated Review Refresh 2023[130] will support government to build UK capability and ensure the UK harnesses the benefits presented by this emerging technology. Our proposed framework will ensure we create the right regulatory environment as we move to maximise the transformative potential of foundation models.

Case-study 3.9: Life cycle accountability for large language models

> Large language models (LLMs) are a type of foundation model.[131] The potential of LLMs goes beyond reproducing or translating natural language: LLMs also have the power to write software,[132] generate stories[133] through films and virtual reality,[134] and more.[135]

[129] See for example, The value chain of general-purpose AI, Ada Lovelace Institute, 2023; An overview of AI alignment, Conjecture, 2023; Make safe systems and deploy them reliably, Anthropic, 2023.

[130] Integrated Review Refresh 2023, Prime Minister's Office, 10 Downing Street, Foreign, Commonwealth and Development Office, Ministry of Defence 2023.

[131] On the opportunities and risks of foundation models, Bommasani et al., 2022.

[132] Jigsaw: Large Language Models meet Program Synthesis, Jain et al., 2021.

[133] Huge "foundation models" are turbo-charging AI progress, The Economist, 2022.

[134] GPT-3 Powers the Next Generation of Apps, Open AI, 2021.

[135] Large language models broaden AI's reach in industry and enterprise, Venture Beat, 2022.

LLMs fall within the scope of our regulatory framework as they are autonomous and adaptable.

We are mindful of the rapid technological change in the development of foundation models such as LLMs and the new opportunities that they bring to applications including search engines, medical devices, and financial and legal services. However, LLMs also have limitations, for example, the models are not trained on a sense of truth,[136] so they can reproduce inconsistent or false outputs that seem highly credible.[137] Because they can be adapted to a wide variety of tasks downstream within an AI supply chain, any improvements or defects in a foundation model could quickly affect all adapted products.

Under the UK's pro-innovation AI regulatory framework, regulators may decide to issue specific guidance and requirements for LLM developers and deployers to address risks and implement the cross-cutting principles. This could include guidance on appropriate transparency measures to inform users when AI is being used and the data used to train the model.

The wide-reaching impact of LLMs through the AI supply chain – together with their general purpose and potential wide ranging application – means they are unlikely to be directly 'caught' within the remit of any single regulator. This makes effective governance and supply chain risk-management challenging where LLMs are involved. The AI regulatory framework's monitoring and evaluation function will therefore need to assess the impacts of LLMs. The cross-cutting *accountability and governance* principle will encourage regulators and businesses to find ways to demonstrate accountability and good governance in responsible LLM development and use.

At this point it would be premature to take specific regulatory action in response to foundation models including LLMs. To do so would risk stifling innovation, preventing AI adoption, and distorting the UK's thriving AI ecosystem.

However, we are mindful of the rapid rate of advances in the power and application of LLMs, and the potential creation of new or previously unforeseen risks. As such, LLMs will be a core focus of our monitoring and risk assessment functions and we will work with the wider AI community to ensure our adaptive framework is capable of identifying and responding to developments relating to LLMs.

For example, one way of monitoring the potential impact of LLMs could be by monitoring the amount of compute used to train them, which is much easier to assess and govern than other inputs such as data, or talent. This could involve statutory reporting requirements for models over a certain size. This metric could become less useful as a way of establishing who has access to powerful models if machine learning development becomes increasingly open-source.[138]

[136] The Creator of ChatGPT Thinks AI Should Be Regulated, Time, 2023.

[137] ChatGPT by OpenAI; The Chatbot Problem, The New Yorker, 2023.

[138] See Future of Compute Review: Submission of Evidence, Centre for Long Term Resilience, 2022.

Life cycle accountability – including the allocation of responsibility and liability for risks arising from the use of foundation models including LLMs – is a priority area for ongoing research and policy development. We will explore the ways in which technical standards and other tools for trustworthy AI can support good practices for responsible innovation across the life cycle and supply chain. We will also work with regulators to ensure they are appropriately equipped to engage with actors across the AI supply chain and allocate legal liability appropriately.

Consultation questions:

F1. What specific challenges will foundation models such as large language models (LLMs) or open-source models pose for regulators trying to determine legal responsibility for AI outcomes?

F2. Do you agree that measuring compute provides a potential tool that could be considered as part of the governance of foundation models?

F3. Are there other approaches to governing foundation models that would be more effective?

3.3.4 Artificial intelligence sandboxes and testbeds

93. Government is committed to supporting innovators by addressing regulatory challenges that prevent new, cutting-edge products from getting to market. Barriers can be particularly high when a path to market requires interaction with multiple regulators or regulatory guidance is nascent. Sir Patrick Vallance's Digital Report recommends that government works with regulators to develop an AI sandbox to support innovators. At the Budget, government confirmed our commitment to taking forward this recommendation.[139]

94. The Information Commissioner's Office (ICO) and the Financial Conduct Authority (FCA) have already successfully piloted digital sandboxes in their sectors.[140] The FCA sandbox has worked with over 800 businesses and accelerated their speed to market by an estimated 40% on average.[141] Sandbox participation has also been found to have significant financial benefits, particularly for smaller organisations.[142] We have heard from regulators, including those with less experience of taking part in previous initiatives, that they are keen to participate in new AI sandboxes to support their regulated sectors.

95. Regulatory sandboxes and testbeds will play an important role in our proposed regulatory regime**.** Such initiatives enable government and regulators to:

[139] HM Government Response to Sir Patrick Vallance's Pro-Innovation Regulation of Technologies Review, HM Treasury, 2023.
[140] Regulatory Sandbox, ICO, 2022; Regulatory Sandbox, FCA, 2022.
[141] Innovation Hub: Market Insights, FCA. 2023; A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications, Tuby et al, 2021.
[142] Inside the regulatory sandbox: effects on fintech funding, Cornelli et al, 2020.

- Support innovators to get novel products and services to market faster, so they can start generating economic and social benefits.
- Test how the regulatory framework is operating in practice and illuminate unnecessary barriers to innovation that need to be addressed.
- Identify emerging technology and market trends to which our regulatory framework may need to adapt.

96. To deliver an effective sandbox, we would like to understand more deeply what service focus would be most useful to industry. We are considering four options:

    - Single sector, single regulator: support innovators to bring AI products to the market in collaboration with a single regulator, focusing on only one chosen industry sector.[143]
    - Multiple sectors, single regulator: support AI innovators in collaboration with a single regulator that is capable of working across multiple industry sectors.[144]
    - Single sector, multiple regulator: establish a sandbox that only operates in one industry sector but is capable of supporting AI innovators whose path to market requires interaction with one or more regulators operating in that sector.[145]
    - Multiple sectors, multiple regulators: a sandbox capable of operating with one or more regulators in one or more industry sectors to help AI innovators reach their target market. The DRCF is piloting a version of this model.[146]

97. We intend to focus an initial pilot on a single sector, multiple regulator sandbox. Recognising the importance of AI innovations that have implications in multiple sectors (like generative AI models), we would look to expand this capability to cover multiple industry sectors over time.

98. Initially, we envisage focusing the sandbox on a sector where there is a high degree of AI investment, industry demand for a sandbox, and appetite for improved collaboration between regulators to help AI innovators take their products to market. We invite consultation feedback on this proposal as well as suggestions for industry sectors that meet these criteria.

99. We would also like to build a deeper understanding of what service offering would be most helpful to industry. Some sandboxes offer supervised real-life or simulated test environments where innovators can trial new products, often under relaxed regulatory requirements.[147] In other scenarios, a team of technologists and regulation experts give customised advice and support to participating innovators over a number of months to help them understand and overcome regulatory barriers so they can reach their target market.[148] Our current preference is for the customised advice and support model, as we think this is where we can deliver benefits

[143] For an existing example of this type of model see Regulatory Sandbox, FCA, 2022.

[144] For an existing example of this type of model see Regulatory Sandbox, ICO, 2023.

[145] For a report on a pilot of this type of model see: Using machine learning in diagnostic services, CQC, 2020.

[146] Enabling innovation – piloting a multi-agency advice service for digital innovators, Regulators' Pioneer Fund, 2022.

[147] The MHRA's 'airlock process' is an example of this kind of service, designed for AI products meeting certain criteria. See: Software and AI as a medical device change programme, MHRA, 2022.

[148] For an example, see: NHS Innovation Service, Accelerated Access Collaborative, 2023. For AI projects, see: AI and Digital Regulations Service, Care Quality Commission, Health Research Authority, Medicines and Healthcare Products Regulatory Agency, National Institute for Health and Care Excellence, 2023.

most effectively in the short term. We will explore options for developing a safe test environment capability at a later date, informed by our initial pilot work.

100. The implementation of an AI regulatory sandbox will also be closely informed by Sir Patrick Vallance's review into digital regulation and his recommendation to establish a multi-regulator sandbox.[149] The review sets out a number of design principles, which we will build into our pilot approach. This includes targeting such initiatives at start-ups and small to medium-sized businesses. As a matter of priority, we will engage with businesses to understand how such an approach should be designed and delivered to best support their needs.

Consultation questions:

S1. To what extent would the sandbox models described in section 3.3.4 support innovation?

S2. What could government do to maximise the benefit of sandboxes to AI innovators?

S3. What could government do to facilitate participation in an AI regulatory sandbox?

S4. Which industry sectors or classes of product would most benefit from an AI sandbox?

3.3.5 Regulator capabilities

101. Government has prioritised the ongoing assessment of the different capability needs across the regulatory landscape. We will keep this under close review as part of our ongoing monitoring and evaluation activity.

102. While our approach does not currently involve or anticipate extending any regulator's remit,[150] regulating AI uses effectively will require many of our regulators to acquire new skills and expertise. Our research[151] has highlighted different levels of capability among regulators when it comes to understanding AI and addressing its unique characteristics. Our engagement has also elicited a wide range of views on the capabilities regulators require to address AI risks and on the best way for regulators to acquire these.

103. We identified potential capability gaps among many, but not all, regulators, primarily in relation to:

AI expertise. Particularly:

149 Pro-innovation Regulation of Technologies Review: Digital Technologies, HM Treasury, 2023.

150 Any attempt by a regulator to enforce a principle beyond its existing remit and powers may be legally challenged on the basis of going beyond its legal authority.

151 Including but not limited to Common Regulatory Capacity for AI, The Alan Turing Institute, 2022.

- Technical expertise in AI technology.[152] For example, on how AI is being used to deliver products and services and on the development, use and applicability of technical standards.[153]
- Expertise on how AI use cases interact across multiple regulatory regimes.
- Market intelligence on how AI technologies are being used to disrupt existing business models, both in terms of the potential opportunities and risks that can impact regulatory objectives.

Organisational capacity. A regulator's ability to:

- Effectively adapt to the emergence of AI use cases and applications, and assimilate and share this knowledge throughout the organisation.
- Work with organisations that provide assurance techniques (e.g. assurance service providers) and develop technical standards (i.e. standards development organisations), to identify relevant tools and embed them into the regulatory framework and best practice.
- Work across regulators to share knowledge and cooperate in the regulation of AI use cases that interact across multiple regulatory regimes.
- Establish relationships and communicate effectively with organisations and groups not normally within their remit.

104. In the initial phases of implementation, government will work collaboratively with key partners to leverage existing work on this topic. For example, the Digital Regulation Cooperation Forum (DRCF) is already exploring ways of addressing capability gaps within its members.

105. There are options for addressing capability gaps within individual regulators and across the wider regulatory landscape, which we will continue to explore. It may, for example, be appropriate to establish a common pool of expertise that could establish best practice for supporting innovation through regulatory approaches and make it easier for regulators to work with each other on common issues. An alternative approach would be to explore and facilitate collaborative initiatives between regulators – including, where appropriate, further supporting existing initiatives such as the DRCF – to share skills and expertise.

Consultation questions:

19. As a regulator, what support would you need in order to apply the principles in a proportionate and pro-innovation way?

20. Do you agree that a pooled team of AI experts would be the most effective way to address capability gaps and help regulators apply the principles?

[152] There is evidence that this is predominantly a recruitment problem. Regulators are trying to recruit but often cannot find the right candidates as they are competing for a limited supply of suitable candidates.

[153] Evidence showed that technical standards expertise varies across regulators. MHRA regularly uses and designates standards to clarify legal requirements, provide presumptive conformity and demonstrate the state of the art. Other regulators recognise their potential to support regulatory guidance but their thinking is nascent.

Part Four: Tools for trustworthy AI to support implementation

4.1 AI assurance techniques

106. Tools for trustworthy AI including assurance techniques and technical standards will play a critical role in enabling the responsible adoption of AI and supporting the proposed regulatory framework. Industry and civil society were keen to see a range of practical tools to aid compliance. Government is already supporting the development of these tools by publishing a Roadmap to an effective AI assurance ecosystem in the UK[154] and establishing the UK AI Standards Hub[155] to champion the use of technical standards.[156]

107. To assure AI systems effectively, we need a toolbox of assurance techniques to measure, evaluate and communicate the trustworthiness of AI systems across the development and deployment life cycle. These techniques include impact assessment, audit, and performance testing along with formal verification methods.

108. It is unlikely that demand for AI assurance can be entirely met through organisations building in-house capability. The emerging market for AI assurance services and expertise will have an important role to play in providing a range of assurance techniques to actors within the AI supply chain. There is an opportunity for the UK to become a global leader in this market as the AI assurance industry develops. This will enable organisations to determine whether AI technologies are aligned with relevant regulatory requirements.

109. To help innovators understand how AI assurance techniques can support wider AI governance, the government will launch a Portfolio of AI assurance techniques in Spring 2023. The Portfolio is a collaboration with industry to showcase how these tools are already being applied by businesses to real-world use cases and how they align with the AI regulatory principles.

4.2 AI technical standards

110. Assurance techniques need to be underpinned by available technical standards, which provide common understanding across assurance providers. Technical standards and assurance techniques will also enable organisations to demonstrate that their systems are in line with the UK's AI regulatory principles.

111. Multiple international and regional standards development organisations are developing, or have already released, AI-specific technical standards, addressing topics such as risk management, transparency, bias, safety and robustness. Accordingly, technical standards can be used by regulators to complement sector-specific approaches to AI regulation by providing

[154] Roadmap to an effective AI assurance ecosystem in the UK, DSIT (formerly DCMS), 2021.

[155] The AI Standards Hub is led by The Alan Turing Institute in partnership with the British Standards Institution (BSI) and the National Physical Laboratory (NPL) and supported by the UK Government.

[156] Technical standards are generally voluntary and developed through an industry-led process in global standards development organisations (SDOs), based on the principles of consensus, openness, and transparency, and benefiting from global technical expertise and best practice. In this paper, when referring to "technical standards", we are referring to standards developed in standards development organisations.

common benchmarks and practical guidance to organisations.[157] Overall, technical standards can embed flexibility[158] into regulatory regimes and drive responsible innovation by helping organisations to address AI-related risks.[159]

112. The UK plays a leading role in the development of international technical standards, working with industry, international and UK partners.[160] The government will continue to support the role of technical standards in complementing our approach to AI regulation, including through the UK AI Standards Hub.

Box 4.1: Supporting a layered approach to AI technical standards

The government will complement its context-specific approach to AI regulation by proposing a proportionate 'layered approach' to applying available AI technical standards. This involves regulators identifying relevant technical standards and encouraging their adoption by actors in the AI life cycle to support the integration of the AI regulation principles into technical and operational business processes:

Layer 1: To provide consistency and common foundations across regulatory remits, in the first instance regulators could seek to encourage adoption of sector-agnostic standards which can be applied across AI use cases to support the implementation of cross-sectoral principles. For example, management systems, risk management, and quality standards[161] can provide industry with good practices for the responsible development of AI systems. The adoption of these standards should be encouraged by multiple regulators as tools for regulated entities to establish common good practices for AI governance.

Layer 2: To adapt these governance practices to the specific risks raised by AI in a particular context, regulators could look at encouraging adoption of additional standards addressing specific issues such as bias and transparency.[162] Such standards would act as tools for industry to operationalise compliance with specific AI regulation principles. As these standards will provide good practices for AI governance applicable to multiple sectors, regulators could complement these with sector-specific guidance.

For example, standards for bias mitigation could be promoted by the Financial Conduct Authority (FCA) and the Equality and Human Rights Commission (EHRC)

[157] AI-specific standards addressing trustworthiness characteristics such as safety, transparency and robustness, amongst others, have been developed or are currently being developed ("*" indicates standards which are under development at the time of writing) in SDOs such as ISO/IEC and IEEE (e.g. IEEE 7001, ISO/IEC TS 6254*, ISO/IEC TR 5469*, ISO/IEC 24029-2*).

[158] Technical standards can be updated as good practices and the technology develop, allowing flexibility for requirements to adapt to technological change.

[159] Standards help organisations to manage and mitigate risks, as well as helping to unlock and scale the benefits of their products and services. In doing so, standards play a role in responsible innovation both as tools supporting good governance and as mechanisms for enabling and accelerating innovation.

[160] The UK government established a strategic coordination initiative with the British Standards Institution (BSI) and the National Physical Laboratory (NPL) to step up UK's engagement in the global development of standards.

[161] For example, these include ISO/IEC DIS 42001* , ISO/IEC FDIS 23894* and ISO/IEC DIS 25059.

[162] For example, transparency standards include ISO/IEC AWI 12792*, IEEE P7001-2021 and ISO/IEC AWI TS 6254*. Bias mitigation standards include ISO/IEC TR 24027:2021 and ISO/IEC AWI TS 12791*.

as practical tools for providers of AI scoring models to identify and mitigate relevant sources of bias to ensure the fairness of the outcomes when the AI model is applied to financial services (credit scoring) and HR practices (candidate scoring) respectively.

Layer 3: Where relevant, regulators could encourage adoption of sector-specific technical standards to support compliance with specific regulatory requirements and performance measures.[163]

Consultation questions:

21. Which non-regulatory tools for trustworthy AI would most help organisations to embed the AI regulation principles into existing business processes?

[163] For example, safety in healthcare can be addressed by the joint application of management system, risk management and quality standards along with horizontal thematic safety standards (e.g., ISO 5469*) and sector specific standards (e.g. BS 30440*). Accordingly, regulators such as MHRA might decide to reference sector-specific standards in their regulatory guidance as tools for AI providers to demonstrate compliance with regulatory requirements for AI as a medical device.

Part Five: Territorial application

5.1 Territorial application of the regulatory framework

113. Our AI regulation framework applies to the whole of the UK. AI is used in various sectors and impacts on a wide range of policy areas, some of which are reserved and some of which are devolved. We will continue to consider any devolution impacts of AI regulation as the policy develops and in advance of any legislative action. Some regulators share remits with their counterparts in the devolved administrations. Our framework, to be initially set out on a non-statutory basis, will not alter the current territorial arrangement of AI policy. We will rely on the interactions with existing legislation on reserved matters, such as the Data Protection Act 2018 and the Equality Act 2010, to implement our framework.

114. We will continue to engage devolved administrations, businesses, and members of the public from across the UK to ensure that every part of the country benefits from our pro-innovation approach. We will, for example, convene the devolved administrations for views on the functions we expect the government to perform and on the potential implications of introducing a statutory duty on regulators to have due regard to the principles.

5.2 Extraterritorial application of the regulatory framework

115. While we expect our principles-based approach to influence the global conversation on AI governance, we are not currently proposing the introduction of new *legal* requirements. Our framework will not therefore change the territorial applicability of existing legislation relevant to AI (including, for example, data protection legislation).

Part Six: Global interoperability and international engagement

6.1 Our regulatory framework on the world stage

116. Countries and jurisdictions around the world are moving quickly to set the rules that govern AI. The UK is a global leader in AI with a strategic advantage that places us at the forefront of these developments. The UK is ranked third in the world for AI publications and also has the third highest number of AI companies.[164] We want to build on this position, making the UK the best place to research AI and to create and build innovative AI companies. At the same time, we recognise the importance of working closely with international partners. As such, the UK's approach to both our domestic regulation and international discussions will continue to be guided by our ambition to develop AI frameworks that champion our democratic values and economic priorities.

117. In line with our domestic approach, we will focus on supporting the positive global opportunities AI can bring while protecting citizens against the potential harms and risks that can emanate across borders. We will work closely with international partners to both learn from, and influence, regulatory and non-regulatory developments (see examples in box 6.1). Given the complex and cross-border nature of AI supply chains, with many AI businesses operating across multiple jurisdictions, close international cooperation will strengthen the impact of our proposed framework.

118. We will promote interoperability and coherence between different approaches, challenging barriers which may stand in the way of businesses operating internationally. We will ensure that the UK's regulatory framework encourages the development of a responsive and compatible system of global AI governance. We will build our international influence, allowing the UK to engage meaningfully with like-minded partners on issues such as cross-border AI risks and opportunities.

119. The UK will continue to pursue an inclusive, multi-stakeholder approach, from negotiating new global norms to helping partner countries build their awareness and capacity in relation to the benefits and risks of AI technology. We will, for example, support other nations to implement regulation and technical standards that support inclusive, responsible and sustainable artificial intelligence. More widely, the International Tech Strategy reiterates how we will shape global AI activities in line with UK values and priorities, protecting against efforts to adopt and apply AI technologies in the service of authoritarianism and repression. We will work with UK industry leaders to ensure that we stay at the forefront of AI and share our best practice with like-minded nations. Similarly, we will learn from our international partners, encouraging them to share lessons we can integrate into our framework.

120. Our international approach will include ensuring that proven, effective, and agreed upon assurance techniques and international technical standards play a role in the wider regulatory ecosystem. Such measures will also support cross-border trade by setting out risk management and AI governance practices that are globally recognised by trading partners, reducing technical barriers to trade and increasing market access. We will also use our world-leading innovation

[164] Global AI Index, Tortoise Media, 2022; AI rankings by country, AI Rankings, 2023.

provisions in Free Trade Agreements to address the challenges innovators in AI may face and ensure that businesses are able to take advantage of the opportunities it presents.

121. In multilateral engagements, we will work to leverage each forum's strengths, expertise and membership to ensure they are adding maximum value to global AI governance discussions and are relevant to our democratic values and economic priorities.

Box 6.1: Examples of international engagement and collaboration

The UK has played an active and leading role on the international AI stage and will continue to do so. Some (non-exhaustive) examples of activities are:

Multilateral AI engagement

- **OECD AI Governance Working Party (AI-GO):** The UK is an active member of the OECD's Working Party on AI Governance (AIGO), which supports the implementation of the OECD's AI principles and enables the exchange of experience and best practice to advance the responsible stewardship of AI.[165]
- **Global Partnership on AI (GPAI):** The UK is a key contributor to – and founding member of – the Global Partnership on AI (GPAI), which is an independent organisation consisting of 29 countries and a range of international experts. GPAI was launched in 2020 as the first international multilateral forum to focus solely on AI and the UK has played a significant role in shaping its development and influencing its agenda.[166]
    - At the 2022 GPAI Ministerial Summit in Japan, we demonstrated the scale of the UK's AI ambitions by announcing £1.2m of funding to develop a Net Zero Data Space for AI Applications (which will also support our Net Zero policy objectives).[167] This is in addition to the previous £1m investment to advance GPAI research on data justice (collaborating with The Alan Turing Institute and 12 pilot partners in low and medium income countries).
- **G7:** The UK is actively engaged in the G7's work on AI and we are working closely with Japan – which holds the G7 Presidency for 2023 – to encourage greater international collaboration, support the development of consistent, proportionate and interoperable regulatory interventions, and champion the role of tools for trustworthy AI where appropriate.
- **Council of Europe Committee on AI (CAI):** The UK holds a Bureau position and we are working closely with like-minded nations on the proposed Convention on AI, to help protect human rights, democracy and rule of law.[168]

[165] OECD Working Party and Network of Experts on Artificial Intelligence Governance (AIGO), OECD, 2023.

[166] Global Partnership on Artificial Intelligence, GPAI, 2023.

[167] Climate Change and AI: Recommendations for Government Action, Global Partnership on AI GPAI, Climate Change AI and the Centre for AI & Climate, 2021.

[168] Committee on Artificial Intelligence (CAI), Council of Europe, 2023.

- **UNESCO:** The UK was actively involved in the development of the UNESCO Ethics of AI Recommendations and UK organisations have been supporting the development of implementation tools.[169]
- **Global standards development organisations:** The UK will continue to work with international partners and global standards development organisations to develop and promote global technical standards for AI, including through the UK AI Standards Hub.[170] For example, the UK is playing a leading role in the International Organisation for Standardisation and International Electrotechnical Commission[171] (ISO/IEC) on four active AI projects.[172] Through the British Standards Institution (BSI),[173] we are also a member of the Open Community for Ethics in Autonomous and Intelligent Systems (OCEANIS).[174]

Bilateral AI engagement

- The UK is engaging with individual nations and jurisdictions as they develop regulatory and governance approaches to AI. These include the European Union (and its Member States), US, Canada, Singapore, Japan, Australia, Israel, Norway, and Switzerland, amongst many others. We will continue to maintain close dialogues to share information and knowledge, learn from and adapt our approach in collaboration with others, and work together to shape the international landscape.

169 Artificial Intelligence, UNESCO, 2023.

170 AI Standards Hub, 2023.

171 International Organisation for Standardisation and International Electrotechnical Commission, ISO, 2023.

172 The ISO/IEC work programme, which the UK is contributing to alongside our partners, includes the development of an AI Management System Standard (MSS), which intends to help solve some of the implementation challenges relating to AI governance. This standard will be known as ISO/IEC 42001 and will help organisations develop or use artificial intelligence responsibly when pursuing their objectives, and fulfil their obligations to interested parties. Additionally, through BSI, the UK is leading the development of AI international standards in concepts and terminology at ISO/IEC, including those on data, bias, governance implications, and data life cycles. At the European Telecommunications Standards Institute (ETSI) we have led the creation of documents including the ETSI GR SAI 002 on Data Supply Chain Security, out of the UK's National Cyber Security Centre.

173 British Standards Institution, 2023.

174 The Open Community for Ethics in Autonomous and Intelligent Systems (OCEANIS), 2023.

Part Seven: Conclusion and next steps

7.1 Conclusion and next steps

122. Our proportionate approach to regulating AI is designed to strengthen the UK's position as a global leader in artificial intelligence, harness AI's ability to drive growth and prosperity,[175] and increase public trust in these technologies. The approach we set out is proportionate, adaptable, and context-sensitive to strike the right balance between responding to risks and maximising opportunities.

123. The proposals set out in this document have been informed by the feedback we received from over 130 respondents as part of our call for views on our 2022 policy paper. We will continue to work closely with businesses and regulators as we start to establish the central functions we have identified. Ongoing engagement with industry will be key to our monitoring and evaluation. Feedback will ensure the framework can adapt to new evidence, future-proofing the UK's role as a leader in AI innovation and ensuring that we can take a leading role in shaping the global narrative on AI regulation.

124. Given the pace at which AI technologies and risks emerge, and the scale of the opportunities at stake, we know that there is no time to waste if we are to strengthen the UK's position as one of the best places in the world to start an AI company. In collaboration with regulators, we are already exploring approaches to implementing the framework and will scale up this activity over the coming months. We are committed to an adaptable, iterative approach that allows us to learn and improve the framework. Our sovereign parliamentary system enables us to deliver targeted and proportionate measures – including by adapting existing legislation if necessary – based on emerging evidence.[176] There are therefore aspects of our implementation work that will be delivered in parallel with the wider consultation set out in this white paper.

125. In the first six months following publication we will:

- Engage with industry, the public sector, regulators, academia and civil society through the consultation period.
- Publish the government's response to this consultation.
- Issue the cross-sectoral principles to regulators, together with initial guidance to regulators for their implementation. We will work with regulators to understand how the description of AI's characteristics can be applied within different regulatory remits and the impact this will have on the application of the cross-sectoral principles.
- Design and publish an AI Regulation Roadmap with plans for establishing the central functions (detailed in section 3.3.1), including monitoring and coordinating implementation of the principles. This roadmap will set out key partner organisations and identify existing initiatives that will be scaled-up or leveraged to deliver the central functions. It will also include plans to pilot a new AI sandbox or testbed.

[175] The AI sector is estimated to contribute £3.7bn in GVA (Gross Value Added) to the UK economy. AI Sector Study 2022, DSIT, 2023.
[176] What is the constitution? The Constitution Unit, University College London, 2023.

- Analyse findings from commissioned research projects and improve our understanding of:
  - Potential barriers faced by businesses seeking to comply with our framework and ways to overcome these.
  - How accountability for regulatory compliance is currently assigned throughout the AI life cycle in real-world scenarios.
  - The ability of key regulators to implement our regulatory framework, and how we can best support them.
  - Best practice in measuring and reporting on AI-related risks across regulatory frameworks.

126. In the six to twelve months after publication we will:

- Agree partnership arrangements with leading organisations and existing initiatives to deliver the first central functions.
- Encourage key regulators to publish guidance on how the cross-sectoral principles apply within their remit.
- Publish proposals for the design of a central M&E framework including identified metrics, data sources, and any identified thresholds or triggers for further intervention or iteration of the framework. This will be published for consultation.
- Continue to develop a regulatory sandbox or testbed with innovators and regulators.

127. In the longer-term, twelve months or more after publication, we will:

- Deliver a first iteration of all the central functions required to ensure the framework is effective.
- Work with key regulators that have not published guidance on how the cross-sectoral principles apply within their remit to encourage and support them to do so.
- Publish a draft central, cross-economy AI risk register for consultation.
- Develop the regulatory sandbox or testbed drawing on insights from the pilot.
- Publish the first monitoring and evaluation report. This will evaluate how well the cross-sectoral principles are functioning and the delivery of the central functions. Performance will be measured against our framework characteristics: pro-innovation, proportionate, trustworthy, adaptable, clear and collaborative. The report will also consider existing regulatory activity and the role of government in supporting this, including whether appropriate guidance (including joint guidance) has been issued. In the report, we will include considerations on the need for any iteration of the framework, including the need for statutory interventions.
- Publish an updated AI Regulation Roadmap which will set out plans for the future delivery of the central functions. In particular, it will assess whether a central government team is the most appropriate mechanism for overseeing the central functions in the longer term or if a more independent body would be more effective.

Consultation questions:

22. Do you have any other thoughts on our overall approach? Please include any missed opportunities, flaws, and gaps in our framework.

Annex A: Implementation of the principles by regulators

A.1 Factors that government believes regulators may wish to consider when providing guidance/implementing each principle

Principle	Implementation considerations
Safety, security and robustness	We anticipate that regulators will need to: 1. Provide guidance about this principle including: • considerations of good cybersecurity practices, such as the NCSC principles for the security of machine learning,[177] as a secured system should be capable of maintaining the integrity of information. • considerations of privacy practices such as accessibility only to authorised users and safeguards against bad actors. 2. Refer to a risk management framework that AI life cycle actors should apply. Models should be regularly reviewed over time as a mitigation strategy. 3. Consider the role of available technical standards, for example addressing AI safety, security, testing, data quality, and robustness (e.g. ISO/IEC 24029-2*, ISO/IEC 5259-1*, ISO/IEC 5259-3* , ISO/IEC 5259-4* , ISO/IEC TR 5469*) to clarify regulatory guidance and support the implementation of risk treatment measures.
Appropriate transparency and explainability	We anticipate that regulators will need to: 1. Set expectations for AI life cycle actors to proactively or retrospectively provide information relating to: • the nature and purpose of the AI in question including information relating to any specific outcome, • the data being used and information relating to training data, • the logic and process used and where relevant information to support explainability of decision-making and outcomes,

[177] Principles for the security of machine learning, National Cyber Security Centre, 2022.

	• accountability for the AI and any specific outcomes. 2. Set explainability requirements, particularly of higher risk systems, to ensure appropriate balance between information needs for regulatory enforcement (e.g. around safety) and technical tradeoffs with system robustness. 3. Consider the role of available technical standards addressing AI transparency and explainability (e.g. IEEE 7001, ISO/IEC TS 6254*, ISO/IEC 12792*)[178] to clarify regulatory guidance and support the implementation of risk treatment measures.
Fairness	We anticipate that regulators will need to: 1. Interpret and articulate 'fairness' as relevant to their sector or domain. 2. Decide in which contexts and specific instances fairness is important and relevant (which it may not always be). 3. Design, implement and enforce appropriate governance requirements for 'fairness' as applicable to the entities that they regulate. 4. Where a decision involving use of an AI system has a legal or similarly significant effect on an individual, regulators will need to consider the suitability of requiring AI system operators to provide an appropriate justification for that decision to affected parties. 5. AI systems should comply with regulatory requirements relating to vulnerability of individuals within specific regulatory domains. Regulators will need to consider how use of AI systems may alter individuals' vulnerability, pursuant to their existing powers and remits. 6. Consider the role of available technical standards addressing AI fairness, bias mitigation and ethical considerations (e.g. ISO/IEC TR 24027:2021, ISO/IEC 12791*, ISO/IEC TR 24368:2022) to clarify regulatory guidance and support the implementation of risk treatment measures.
Accountability and governance	We anticipate that regulators will need to: 1. Determine who is accountable for compliance with existing regulation and the principles. In the initial stages of implementation, regulators might provide guidance on how to demonstrate accountability. In the medium to long term, government may issue additional guidance on how accountability applies to specific actors within the ecosystem.

[178] Technical standards marked with "*" are under development.

	2. Provide guidance on governance mechanisms including, potentially, activities in scope of appropriate risk management and governance processes (including reporting duties). 3. Consider how available technical standards addressing AI governance, risk management, transparency and other issues can support responsible behaviour and maintain accountability within an organisation (e.g. ISO/IEC 23894*, ISO/IEC 42001*, ISO/IEC TS 6254*, ISO/IEC 5469* , ISO/IEC 25059*).
Contestability and redress	We anticipate that regulators will need to: 1. Create or update guidance with relevant information on where to direct a complaint or dispute for those affected by AI harms. Guidance should clarify existing 'formal' routes of redress offered by regulators in certain scenarios. 2. Clarify interactions with requirements of appropriate transparency and explainability, acting as pre-conditions of effective redress and contestability.

Annex B: Stakeholder engagement

B.1 Summary

In July 2022, we published a policy paper outlining our proposals for Establishing a pro-innovation approach to regulating AI.[179] We proposed a non-statutory framework underpinned by a set of cross-sectoral principles including transparency, safety, and security. The principles were intended to guide how regulators approach AI risks. We outlined our intention for the framework to be coherent, proportionate and adaptable, with regulatory coordination to reduce burdens on business and agility to keep pace with rapid technological advancements. Our proposals were designed to strengthen the UK's position as a global leader in AI by ensuring the UK is the best place to develop and use AI technologies.

We launched a call for views on the proposals outlined in our policy paper to capture feedback from stakeholders between July and September 2022. We received responses from over 130 different stakeholders. There were some clear themes amongst the responses, with stakeholders noting the importance of regulatory coordination and asking for further details on how this will be achieved.

The 2023 AI regulation white paper sets out our latest position based on the feedback we received. In particular, we have considered the need for new central functions to undertake activities such as system-wide risk monitoring and evaluation of the AI regulation framework.

We welcome feedback on our latest proposals and will actively engage stakeholders as part of a consultation running to 21 June. See Annex C for more details on how to contribute to this consultation.

B.2 Background

In July 2022, we opened a public call for views on our policy paper: Establishing a pro-innovation approach to regulating AI. We invited stakeholder views on how the UK can best set the rules for regulating AI in a way that drives innovation and growth while also protecting our fundamental values. Feedback was collected to inform the development of the white paper.

We welcomed reflections on our proposed approach and specifically invited views and supporting evidence on the following questions:

1. What are the most important challenges with our existing approach to regulating AI? Do you have views on the most important gaps, overlaps or contradictions?
2. Do you agree with the context-driven approach delivered through the UK's established regulators set out in this paper? What do you see as the benefits of this approach? What are the disadvantages?
3. Do you agree that we should establish a set of cross-sectoral principles to guide our overall approach? Do the proposed cross-sectoral principles cover the common issues and risks posed by AI technologies? What, if anything, is missing?
4. Do you have any early views on how we best implement our approach? In your view, what are some of the key practical considerations? What will the regulatory system need to deliver on our approach? How can we best streamline and coordinate guidance on AI from regulators?

[179] Establishing a pro-innovation approach to regulating AI, Office for Artificial Intelligence, 2022.

5. Do you anticipate any challenges for businesses operating across multiple jurisdictions? Do you have any early views on how our approach could help support cross-border trade and international cooperation in the most effective way?
6. Are you aware of any robust data sources to support monitoring the effectiveness of our approach, both at an individual regulator and system level?

The call for views and evidence was open for 10 weeks, closing on 26 September 2022. In this period we met with 39 stakeholders to capture detailed feedback on our proposals. In total, we received responses from over 130 stakeholders. Stakeholders represented a range of perspectives, from start-ups to Big Tech, and included developers, deployers, and funders from across the AI life cycle. We also heard from researchers, regulators, lawyers, trade bodies and unions as well as representatives from the devolved administrations, local government, and wider public sector.

We have carefully analysed all the views and evidence submitted. We are grateful for the time and effort our stakeholders committed during this process, which has informed and strengthened our policy position as outlined in the white paper.

B.3 Responses

Overall, there was strong support for context specific regulation implemented by existing regulators and many noted that this approach would drive innovation. Stakeholders felt our proposals were a proportionate way to establish regulatory best practice in a fast-changing landscape. However, responses also asked for more practical detail, particularly around risk tolerance, compliance measures, and the overall coherence of the framework.

Our analysis found six overarching themes raised by stakeholders:

1. Articulating the intended societal benefits of AI is key to a future-proofed regulatory vision that works for citizens as well as businesses.

Stakeholders were keen to see a long-term vision that set out our ambition to unlock societal benefits alongside economic opportunities. Stakeholders broadly agreed that the principles addressed the key risks posed by AI. A number of stakeholders commented that our approach should explicitly reference human rights. While stakeholders welcomed our alignment with the OECD framework, many felt further use of international approaches by organisations such as the OECD or UNESCO would add more human focused benefits and aid companies working across jurisdictions. A small number of stakeholders noted that environmental sustainability was missing from our principles. Some suggested that it should be included as a core principle, while others recommended that environmental outcomes should be measured through impact assessments.

Government response: We have analysed our principles in consideration of both stakeholder feedback and our risk assessment work. The white paper clarifies the substance of the principles in section 3.2.3. Human rights and environmental sustainability are not explicitly named in the revised principles as we expect regulators to adhere to existing law when implementing the principles. We have emphasised the social benefits alongside the economic opportunities we intend to unlock with our pro-innovation approach to AI regulation.

2. Offering greater central clarity around the scope of the regime is critical to ensuring business confidence

A number of stakeholders praised our description of AI for capturing the distinct regulatory challenges that AI poses and our proposed characteristics were largely considered to be fit for purpose. There were some concerns that the definition was not 'user-friendly' on its own. While many felt that creating a more specific definition of AI would be difficult and some noted it could be unhelpful, there was clear appetite for further detail on how regulators will maintain a coherent definition of AI within and across sectors. Use cases were suggested as a means of illustrating AI technologies within scope.

Many stakeholders, especially from industry, were keen to see a clear and transparent risk management framework with assessment criteria. In particular, multiple stakeholders felt that it would be beneficial for central government or a central body to provide a clear description of what constitutes 'unacceptable risk'. Some suggested this could complement more detailed risk analysis by regulators to ensure a coordinated and coherent approach – as well as effectively identifying any gaps. Stakeholders indicated that greater clarity on risk would support business development and could also promote high standards, public trust, and the adoption of AI.

Government response: We stress-tested our proposed characteristics of AI against stakeholder feedback and found that concerns centred on how we would ensure coherence across sectors and regulators. We recognise a trade-off between the certainty provided by a blanket approach, such as a singular definition and central risk framework, and the agility enabled by sector-specific expertise, including regulator-refined definitions. Given the fast pace of technological development and stakeholder praise for a future-proofed approach, we have retained our core, defining characteristics for AI, see section 3.2.1. We have considered how regulators can be given the technical capability necessary to create clear definitions for AI in and across their sectors, see section 3.2.1. In section 3.3.1 of the white paper, we outline how new central functions will help identify conflicts or gaps in regulator definitions of AI. Acknowledging feedback that a central steer on 'acceptable' risk would provide business confidence and investment, we have proposed that centralised risk monitoring and horizon scanning would be key central functions.

3. A principles-based approach will enable regulation to keep pace with a fast-evolving technology

Stakeholders generally agreed that a principles-based approach implemented by regulators would offer a proportionate way to build best practice. Stakeholders felt the principles address the key risks that AI poses while allowing regulators to tailor approaches to their sectors. Stakeholders welcomed our use of the OECD principles as a means of promoting international alignment and interoperability.

While stakeholders recognised the benefits that a flexible non-statutory approach offers, some stakeholders were concerned that a non-statutory approach would be unenforceable. A few stakeholders suggested clarifying how AI regulation dovetails with existing legislation and defining thresholds for when our regime may shift to statutory implementation.

Government response: We appreciate the praise of our adaptation of the multilaterally agreed OECD principles. We further outline our international approach in the white paper, recognising that interoperability will help ensure that UK businesses can continue to innovate. While we continue with a non-statutory approach for initial implementation, reflecting on stakeholder concerns around enforceability, we anticipate that introducing a statutory duty to have due regard on regulators might be

needed to strengthen the framework. A duty to have due regard to our cross-sector principles will provide a legislative incentive while maintaining flexibility for the framework to adapt to technological changes. We will monitor the implementation of the framework to assess whether it is effective without the need to implement a statutory duty and will also review responses to the white paper consultation.

4. Providing centralised coordination and oversight will be essential to regulatory coherence and horizon scanning

Stakeholders voiced concerns that regulators did not have the capability to ensure a coherent compliance process, especially for businesses operating across or between industry sectors or regulatory remits. Stakeholders reported expensive, time-consuming confusion when there was not clear regulatory ownership of a technology or issue. Some criticised communication and knowledge-sharing between regulators. One stakeholder explained that joint guidance had previously been very useful. Others suggested that regulators should have more stringent duties to collaborate to ensure consistency and shared best practice.

A number of stakeholders were supportive of a central coordination function for existing regulators, as opposed to a new regulator for AI. Many stressed the importance of a coordination function to aid navigation of trade-offs and conflicts (such as between the need to collect data to minimise bias and the need to refrain from collecting data in the interest of privacy). While many stakeholders stated the need for central coordination, many were solution-agnostic. Proposals included:

- An expanded role for the DRCF. Some stakeholders suggested the DRCF was well-positioned to take on a coordination function but others questioned the DRCF's suitability. In particular, it was felt that the DRCF would require more capacity to fulfil a coordination role.
- A new central body to undertake coordination. Stakeholders suggested establishing a new body, such as a 'Centre for AI Governance', to undertake functions such as: conducting cross-sector risk-mapping; conducting regulatory gap analyses and horizon scanning; monitoring the applicability of emerging AI standards; supplying training; and monitoring international approaches.
- Appointing an existing regulator as 'lead regulator' for AI. Some stakeholders felt that regulators should have more incentives to work together and the entire regulatory landscape could learn from more advanced regulators.

Stakeholders stated the importance of clarifying regulator remits and addressing gaps, noting the fast pace of change for AI technologies. Some suggested that a coordination body should be responsible for a horizon scanning function that monitors and evaluates risks.

Government response: Building on reflections from stakeholders, we identified a small range of regulators with remits that are likely to be significantly affected by AI and conducted analysis of their capability to implement our policy paper proposals. We found varied readiness, with some regulators already demonstrating world-leading approaches to regulating AI and others asking for further support. Similarly, knowledge and information sharing mechanisms were not uniform across regulators and we identified a need for coordination mechanisms to streamline compliance processes for business and ensure regulation provides system-wide coverage of current and future opportunities. We considered multiple options for coordination functions, in line with stakeholder suggestions, and incorporated feedback into our analysis. We outline our proposals for central functions in section 3.3.1 in the white paper.

5. Streamlining liability and tailoring reporting obligations will be key to enabling responsible innovation

While stakeholders were strongly supportive of compliance and assurance as a means of facilitating public trust and the wider adoption of AI technologies, many were keen to limit the burden of reporting obligations, particularly for startups and SMEs. Industry stakeholders noted that the costs of reporting burdens would be passed onto consumers. Some stakeholders emphasised that the government should have a role in providing education and support for small businesses.

There was interest in regulatory sandboxes as a way to enable investment and establish best practice. Generally there was a strong appetite for industry-led solutions and a less burdensome or 'tick box' approach to compliance. Stakeholders were strongly supportive of standards as a way to drive accountability, adoption, and good consumer outcomes. Stakeholders suggested sector compliance templates and voluntary industry forums as ways to share knowledge and reduce the burden of establishing best practice.

Some stakeholders felt the paper lacked a position on liability and argued a clear allocation of legal responsibility would enable effective enforcement and unlock investment. More specifically, some stakeholders suggested that, when appropriate, targeting foundation models (often developed by larger organisations) would increase innovation and competition by reducing liability burdens on smaller companies. Stakeholders often suggested impact assessments could be used to help address liability issues at all stages of the AI life cycle.

Government response: We welcomed the thoughtful suggestions from respondents regarding innovative compliance measures. We noted the significant appetite for regulatory sandboxes and have outlined our proposals in the white paper, see section 3.3.4. We agree that reporting burdens should be proportionate and give detail on how we will continue to work with regulators to ensure compliance measures are streamlined. We acknowledge that regulation measures can affect competition and innovation by creating undue burdens on start-ups and SMEs. We are confident that regulators will oversee proportionate and innovation friendly measures in their remits, with a central function undertaking activity to streamline and ensure coherence. We recognise that liability is complicated by a complex AI value-chain that can incorporate many different actors in different roles. As such, we believe that regulators are best positioned to begin allocating liability in their sectors, adopting a context-based approach that builds on best practice. Our proposal setting out activities to be undertaken centrally will ensure that regulators' approaches to liability are proportionate, coherent across sectors, and supportive of innovation.

6. Establishing interoperability will be critical to ensuring an internationally competitive approach

Stakeholders welcomed the UK's relatively flexible approach but many were concerned that the need for interoperability across jurisdictions would result in businesses conforming to the strictest regulation. Stakeholders warned that international divergence could create more burdens than advantages for businesses. Many stakeholders wanted friction minimised to ensure export prospects for British businesses, with support for an international agreement on AI regulation equivalence, where AI systems authorised on key international markets would be permitted for trade in the UK. Many stakeholders also wanted to see the UK maintain its position as a global leader in AI discussions.

Stakeholders emphasised the importance of alignment with international partners such as the EU and US to ensure global AI governance supports our common democratic values.

Government response: In the white paper, we set out our vision for AI regulation to ensure that the UK is the best place to start and grow an AI business. We share stakeholder concerns on interoperability and plan to continue using our leading role in international forums such as the OECD, G7, and Council of Europe to promote pro-innovation approaches to regulation that capitalise on the potential social and economic benefits of AI while addressing the new risks the technology can pose. Our plan for international engagement, detailed in part six, clarifies our approach with an emphasis on interoperability.

Annex C: How to respond to this consultation

We are inviting individuals and organisations to provide their views by responding to the questions set out in this consultation. The questions are listed below.

The consultation will be open for 12 weeks, until 21 June.

You can respond online via the following link: https://dcms.eu.qualtrics.com/jfe/form/SV_cBDeiMplOHExtYO. Our privacy statement is set out at the following link here.

If for exceptional reasons, you are unable to use the online system, for example because you use specialist accessibility software that is not compatible with the system, you may request and complete a word document version of the form.

By email
evidence@officeforai.gov.uk

By post
Office for Artificial Intelligence
Department for Science, Innovation and Technology
100 Parliament Street
London
SW1A 2BQ

Questions:

The revised cross-sectoral AI principles

1. Do you agree that requiring organisations to make it clear when they are using AI would improve transparency?
2. Are there other measures we could require of organisations to improve transparency for AI?
3. Do you agree that current routes to contest or get redress for AI-related harms are adequate?
4. How could current routes to contest or seek redress for AI-related harms be improved, if at all?
5. Do you agree that, when implemented effectively, the revised cross-sectoral principles will cover the risks posed by AI technologies?
6. What, if anything, is missing from the revised principles?

A statutory duty to regard

7. Do you agree that introducing a statutory duty on regulators to have due regard to the principles would clarify and strengthen regulators' mandates to implement our principles, while retaining a flexible approach to implementation?
8. Is there an alternative statutory intervention that would be more effective?

New central functions to support the framework

9. Do you agree that the functions outlined in section 3.3.1 would benefit our AI regulation framework if delivered centrally?

10. What, if anything, is missing from the central functions?
11. Do you know of any existing organisations who should deliver one or more of our proposed central functions?
12. Are there additional activities that would help businesses confidently innovate and use AI technologies?
    12.1. If so, should these activities be delivered by government, regulators or a different organisation?
13. Are there additional activities that would help individuals and consumers confidently use AI technologies?
    13.1. If so, should these activities be delivered by government, regulators or a different organisation?
14. How can we avoid overlapping, duplicative or contradictory guidance on AI issued by different regulators?

Monitoring and evaluation of the framework

15. Do you agree with our overall approach to monitoring and evaluation?
16. What is the best way to measure the impact of our framework?
17. Do you agree that our approach strikes the right balance between supporting AI innovation; addressing known, prioritised risks; and future-proofing the AI regulation framework?
18. Do you agree that regulators are best placed to apply the principles and government is best placed to provide oversight and deliver central functions?

Regulator capabilities

19. As a regulator, what support would you need in order to apply the principles in a proportionate and pro-innovation way?
20. Do you agree that a pooled team of AI experts would be the most effective way to address capability gaps and help regulators apply the principles?

Tools for trustworthy AI

21. Which non-regulatory tools for trustworthy AI would most help organisations to embed the AI regulation principles into existing business processes?

Final thoughts

22. Do you have any other thoughts on our overall approach? Please include any missed opportunities, flaws, and gaps in our framework.

Legal responsibility for AI

L1. What challenges might arise when regulators apply the principles across different AI applications and systems? How could we address these challenges through our proposed AI regulatory framework?

L2.i. Do you agree that the implementation of our principles through existing legal frameworks will fairly and effectively allocate legal responsibility for AI across the life cycle?

L.2.ii. How could it be improved, if at all?

L3. If you work for a business that develops, uses, or sells AI, how do you currently manage AI risk including through the wider supply chain? How could government support effective AI-related risk management?

Foundation models and the regulatory framework

F1. What specific challenges will foundation models such as large language models (LLMs) or open-source models pose for regulators trying to determine legal responsibility for AI outcomes?

F2. Do you agree that measuring compute provides a potential tool that could be considered as part of the governance of foundation models?

F3. Are there other approaches to governing foundation models that would be more effective?

AI sandboxes and testbeds

S1. To what extent would the sandbox models described in section 3.3.4 support innovation?

S2. What could government do to maximise the benefit of sandboxes to AI innovators?

S3. What could government do to facilitate participation in an AI regulatory sandbox?

S4. Which industry sectors or classes of product would most benefit from an AI sandbox?

If you need a version of this document in a more accessible format, please email alt.formats@beis.gov.uk. Please tell us what format you need. It will help us if you say what assistive technology you use.