Call for Views on the Cyber Security of AI

Contents

Contents 2

Ministerial foreword 3

Executive Summary 5

Chapter 1: Introduction 6

Background 6

Rationale 7

Methodology 8

Chapter 2: Our technology security programme 10

The UK's Approach to Cyber Security 10

Supporting government activity on AI 11

Chapter 3: Voluntary Code of Practice and Global Standard 13

Rationale for a Code of Practice and global standard 13

Chapter 4: How to respond to the Call for Views and our next steps 26

Annex A: Research findings 27

Annex B: Global approach to AI 30

Annex C: Glossary of terms 33

Annex D: Call for Views Survey Questionnaire 34

Annex E: Other interventions considered 39

Annex F: Bibliography of relevant publications mapped to principles by Mindgard 43

Ministerial foreword

Viscount Camrose
Parliamentary Under Secretary of State,
Department for Science, Innovation and Technology

Artificial Intelligence (AI) is a vital technology for the UK economy and for supporting people's everyday lives. AI is enabling organisations to provide better services to customers and offer people quicker access to information. The UK AI market is predicted to grow to over $1 trillion by 2035 and we are well positioned to continue to take advantage of the relentless improvements in AI capabilities. As adoption continues to grow across society, we must ensure that end-users are protected from cyber security risks. This is essential when so many other AI risks stem from an insecure system.

Organisations in the UK face a complex cyber security landscape and we want to ensure that they have the confidence to adopt AI into their infrastructure. Currently, 47% of organisations who use AI do not have any specific AI cyber security practices or processes in place. It is therefore imperative that we ensure that AI is designed, developed, deployed and maintained securely.

AI is identified in the UK's National Cyber Strategy as a technology vital to cyber power and is listed as a critical technology in the UK's Science and Technology Framework. This work on the cyber security of AI sits alongside wider government world-leading efforts to ensure the benefits of AI can be realised safely and securely. This includes hosting the world's first AI Safety Summit in 2023; establishing the AI Safety Institute, the first government-backed organisation of its kind; and being the first government to publish an assessment of advanced AI's capabilities and risks. Successful collaboration with our international partners will continue to underpin everything that we are doing on AI safety and security.

As is the case for other areas of technology, from consumer 'internet of things' products to semiconductors, the UK advocates a secure by design approach. It is imperative that developers of software, including AI, are building security into the design process as well as across the entire lifecycle. In line with our approach, our work in this space is critical to set baseline cyber security requirements and ensure that these are adopted across the software and AI supply chain.

I am therefore pleased to introduce this Call for Views on the Cyber Security of AI and a separate Call for Views on Software Resilience as a continuation of the UK's leadership in this area. The government is proposing to take forward a two-part intervention on AI in the form of a voluntary Code of Practice that will be taken into a global standards development organisation for further development and sets baseline security requirements for stakeholders in the AI supply chain. Outputs from both these Call for Views will drive improved security behaviour across the software supply chain and increase confidence in organisations who adopt software products and services.

Your engagement with this Call for Views will help the government develop and take-forward efforts to build a safer UK, where we can all benefit from AI. It will also support our ongoing work with industry and international partners because we recognise international alignment is crucial for this area. I look forward to continuing discussions on how the government, international partners, industry and civil society should collaborate and prioritise efforts to secure AI. Thank you for your contribution to this critical and generation-defining technology area.

Executive Summary

AI is transforming our daily lives. As the technology continues to evolve and be embedded, it is crucial that we ensure cyber security is a key underpinning of AI safety. This Call for Views sets out specific interventions to help secure AI, so that the many benefits of AI can be realised.

This work has primarily focused specifically on the cyber security risks to AI, rather than wider issues such as safety or the cyber security risks that stem from AI. This work is relevant to all AI technologies, regardless of the sector in which AI is used or the form of AI technology, because security is an essential component and should be considered across the entire AI lifecycle. This work sits alongside wider government activity on AI, much of which is noted in the AI regulation white paper response (see Chapter 2).

The government is proposing to take forward a two-part intervention in the form of a voluntary Code of Practice that will be taken into a global standards development organisation for further development. The proposed voluntary Code sets baseline security requirements for all AI technologies and distinguishes actions that need to be taken by different stakeholders across the AI supply chain.

The voluntary Code of Practice was developed by the Department for Science, Innovation & Technology (DSIT) and is based on the National Cyber Security Centre's (NCSC) Guidelines for secure AI system development which were published in November 2023, alongside the US Cybersecurity and Infrastructure Security Agency and other international cyber partners. The guidelines were co-sealed by agencies from 18 countries. The voluntary Code has also been informed by research we commissioned, including a risk assessment and a mapping of cyber security research in this area. Stakeholder engagement is a key component of our approach and will continue to be embedded throughout this Call for Views process and beyond.

We want to enable AI developers to be able to distinguish themselves from their competitors by highlighting their commitment to security. We also recognise the importance of developing international alignment and ensuring that stakeholders that make up the AI supply chain have a clear understanding of what they need to implement. To that end, we've been engaging closely with international partners and mapped recommendations by industry and other governments to ensure this document sits in support of their efforts. We are also involved in various standards development organisations and multilateral fora to promote the need for security as part of discussions on AI (see Annex B).

This publication is intended as the starting point of a much more extensive dialogue with our stakeholders, including industry and international partners. The cyber security of AI requires a global approach, as the risks cross international borders, and so international engagement has been a key element of our approach. We are now holding a Call for Views for eight weeks until 10th July 2024 to gather feedback on the proposed interventions, including the Code of Practice and the intention to develop a global standard. The feedback will be used to inform UK government policy and our next steps.

Chapter 1: Introduction

Background

1.1 Artificial Intelligence (AI) has become a part of our daily lives and is used by organisations and individuals as a powerful tool to enhance the way we work, interact with data and achieve outcomes. As developers of AI continue to push the boundaries of what models and systems can achieve, it is imperative we address the risks AI presents so that we can continue to unlock the opportunities it offers.[1]

1.2 In the UK, we are already seeing the vast benefits of AI. The UK is home to the third largest number of AI unicorns and start-ups in the world[2] and the AI industry in the UK employs over 50,000 people while contributing £3.7 billion to the economy.[3] AI is also bringing innovation to many sectors, such as transport, agriculture and crime prevention by helping to:

- detect fraud through machine learning algorithms that can identify suspicious transactions in real time;[4]
- optimise public transportation systems by predicting passenger demand; and[5]
- improve organisations cyber security practices through the detection of threats.[6]

1.3 As with any technology, the continued evolution and uptake of AI will also present challenges. To address these challenges, the government set out its pro-innovation and pro-safety regulatory framework, which will ensure that we are able to maximise the opportunities and minimise the risks of this fast-moving technology. The AI regulation white paper outlined five cross-sectoral principles to be applied by existing UK regulators, and a new central function to bring coherence and address regulatory gaps.[7]

1.4 One of the five key principles of the framework is Safety, Security and Robustness. This means that AI systems should function in a robust, secure and safe way throughout the AI lifecycle, and risks should be continually identified, assessed and managed. To achieve this, we need to support developers and deployers of AI systems in addressing cyber security risks to their systems, which in turn will protect users of AI, and strengthen public trust.

1.5 This is vital because cyber security is an essential precondition for the safety of AI systems and is required to ensure, amongst other things, privacy, reliability, and the secure use of models.

1.6 It is imperative that we work towards a global solution for addressing the risks to AI models and systems. This requires a focus on collaborating with international partners to achieve consensus on baseline security requirements (see Annex B for further information on the international landscape).

[1] We define AI developers as those organisations or individuals who design, build, train, adapt, or combine AI models and applications. In the context of the AI Cyber Security Code of Practice, this includes the companies and organisations, development teams, model engineers, data scientists, data engineers and AI designers who are responsible for creating a model and system.

[2] The Global AI Index, Tortoise Media, 2023.

[3] UK unveils world leading approach to innovation in first artificial intelligence white paper to turbocharge growth, Department for Science, Innovation & Technology, 2023.

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

[5] Frontier AI: capabilities and risks, Department for Science, Innovation and Technology, 2023.

[6] The near-term impact of AI on the cyber threat, National Cyber Security Centre, 2024.

[7] A pro-innovation approach to AI regulation, Department for Science, Innovation & Technology, 2023.

1.7 To complement international efforts to drive the adoption of cyber security in AI, we have set out in this document a proposed two-part intervention (Chapter 3) to create, firstly, a voluntary Code of Practice, that will then be used as the basis for the development of a global technical standard. This approach will focus on developing international alignment on baseline security requirements for AI models and systems.

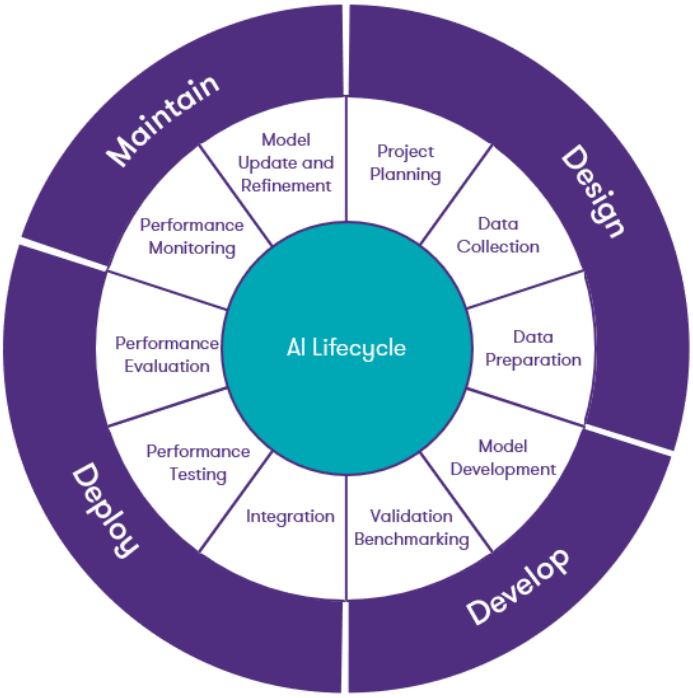
Scope

1.8 This Call for Views is focused on addressing the cyber security risks to AI rather than wider issues relating to AI, such as safety or the cyber security risks that stem from AI. There is specific work on these areas being led by other parts of government (see the AI regulation white paper response for more details).[8]

1.9 Security is an essential component underpinning all types of AI. Therefore, the scope of this Call for Views as well as the voluntary Code of Practice and proposed technical standard, includes all AI technologies, including frontier AI. We have ensured that this work aligns with other parts of government that are examining and addressing the risks associated with frontier AI. This reflects the scope of the AI Safety Summit 2023, the AI regulation white paper and the discussion paper on Frontier AI: Capabilities and Risks.

1.10 We have also focused on the entire AI lifecycle because vulnerabilities in AI systems are not constrained to just one phase of the AI system. We have included below a diagram taken from the risk assessment that DSIT commissioned as an illustration of the different parts of the lifecycle process for creating an AI system.[9]

Diagram of the AI lifecycle



Rationale

[8] A pro-innovation approach to AI regulation: government response, Department for Science, Innovation & Technology, 2024.

[9] Cyber Security Risks to Artificial Intelligence, Department for Science, Innovation & Technology, 2024

1.11 As part of the Government's commitment to deliver the National Cyber Strategy, the Department for Science, Innovation and Technology (DSIT) has undertaken extensive work with the National Cyber Security Centre (NCSC) to make AI systems more secure. This work should also be viewed within the context of the Government's wider ambitions to make the United Kingdom a global AI superpower, as set out in the National AI Strategy.

1.12 We advocate a secure by design approach across all digital technologies which places the responsibility on those that develop technology to build robust security into their systems (further detail on this approach is set out in Chapter 2). Millions of people and organisations are using AI models every day to access better services and information. We must ensure that these models and systems are deployed with security built in by design. Without security, the continued rapid adoption and embedding of AI will result in vulnerabilities being widespread with implications for users and their data, as well as possible national security issues if services relying on AI were compromised or became unavailable.

1.13 We will continue to be actively involved in multilateral initiatives such as the G7, G20, OECD and UN. It is vital that cyber security is a core part of any international agreements and discussions among governments and other stakeholders. Further details on this can be found within Annex B. Our research and work on the voluntary Code of Practice will inform these multilateral discussions.

1.14 We recognise the value of global technical standards in promoting good practices in cyber security. Technical standards developed through a consensus-based multi-stakeholder standards development process promote global interoperability and ensure stakeholders have alignment for baseline cyber security requirements.

1.15 We recognise that work on technical standards relating to AI is underway within multiple global standards development organisations and there is currently a limited focus on baseline security requirements for AI which must be addressed.[10] Further details on our proposed approach involving the development of technical standards can be found in Chapter 3.

1.16 The UK is well positioned to drive forward these crucial discussions within international fora and in global standard development organisations because the UK has consistently demonstrated leadership in championing and developing cyber security for emerging technologies (detailed in Chapter 2). Our approach and proposed interventions have been determined by a comprehensive evidence base. We have also conducted a multi-stakeholder approach to ensure that our proposals are tested thoroughly with experts. This process within the context of AI is set out in more detail below.

Methodology

1.16 Our work has focused on three areas:

- Identifying and evaluating the cyber security vulnerabilities and risks to AI.
- Identifying and evaluating what research and recommendations have been published on addressing the cyber security risks to AI (this informed the development of policy options set out in Annex E).
- Understanding business attitudes towards adopting AI, their views on what is expected from AI developers and their cyber security approach for protecting their infrastructure within the context of AI.

[10] Such discussions are happening within the British Standards Institute (BSI), the European Telecommunications Standards Institute (ETSI), the International Organization for Standardisation (ISO), the European Committee for Standardization (CEN), the European Committee for Electrotechnical Standardisation (CENELEC), the International Telecommunications Union (ITU).

1.17 The content in this Call for Views is supported by various research studies commissioned by DSIT that form part of our robust evidence base. Alongside this Call for Views, we are publishing: a mapping of existing research on the cyber security of AI; a risk assessment of vulnerabilities across the AI lifecycle and how they could be exploited by malicious actors; and a business survey of 350 UK organisations to understand the cyber security processes organisations are implementing for AI technologies. Through this research we are ensuring that our proposed interventions are based on data and can meaningfully address the cyber security risks to AI. Further detail on our research is set out in Annex A.[11]

1.18 In light of the various work being undertaken internationally, we have engaged with international partners, experts within standards development organisations, as well as other stakeholder groups, to ensure our work was thoroughly tested and complements wider international activity. We commissioned a further literature review which mapped both technical and general recommendations made by industry and other governments. This ensured that publications by the international community informed the development of the Code of Practice, and the requirements were written in the context of various initiatives (further detail can be found in Chapter 3 and Annex F).

1.19 This Call for Views is the start of further dialogue with industry and international partners to ensure that users can securely benefit from AI. This work should be viewed in the broader scope of the Government's initiatives in cyber security. The subsequent section offers a summary of various related Government activity linked to AI and cyber security.

[11] Research on the cyber security of AI, Department for Science, Innovation & Technology, 2024.

Chapter 2: Our technology security programme

2.1 All cyber security Codes of Practice produced by DSIT are part of the Government's broader approach to improve baseline cyber security practices and cyber resilience across the UK. They sit alongside various other levers which have been used by the Government to deliver on the objectives of the current, as well as previous, National Cyber Strategies. The Codes of Practice provide guidance ranging from the development of baseline cyber security advice which all organisations should follow, moving progressively towards more product or domain-specific advice due to the increasing risk and evolving threat landscape. A modular approach has been developed to help organisations easily identify which Codes – and within those Codes, which provisions – are relevant to them according to both their business functions, and the types of technologies they either use or manufacture.[12]

2.2 In the case of the voluntary Code contained within this Call for Views, our expectation is that relevant organisations should, at a minimum, also adhere to the provisions in both the Software and Cyber Governance Codes of Practice. While the Cyber Governance Code of Practice sets the baseline expectations for all organisations using digital technologies, the Software Code will also be relevant since software is an integral part of how AI models and systems function. Organisations deemed in scope of this Code would also be expected to assess whether their circumstances warrant consideration of adherence to additional cyber Codes published by the UK Government which may cover specific products or services relevant to them.

2.3 As noted above, this work on AI is closely linked to the Government's recent publication on software resilience due to the inherent overlap between these technology areas. The Government previously held a Call for Views on software resilience and security for businesses and organisations in 2023. The feedback we received on the software resilience Call for Views highlighted the need for a voluntary Code of Practice that set clear expectations with regards to the cyber security responsibilities of software vendors.[13]

The UK's Approach to Cyber Security

2.4 The UK, as a global leader in cyber security, is committed to creating a safe online environment for its citizens. A foundational pillar of this approach is to ensure that both existing and emerging technologies are secure by design. By setting baseline cyber security expectations and incorporating them into the development of digital technologies at inception, we are laying the groundwork for efforts to safeguard users and businesses against evolving cyber threats, and to provide consumers with confidence in the technologies that they use. However, security also needs to be considered throughout the lifecycle of a technology, and the distinct security challenges presented by individual technologies need to be recognised and assessed.

2.5 DSIT has led several initiatives that embed a "secure by design" approach, contributing to the UK's strategic advantage and global cyber security leadership. These include:

- The creation of the world's first mandatory and enforceable security requirements for consumer technology through the Product Security and Telecommunications Infrastructure (PSTI) Act.[14] This work built on the UK's Code of Practice for Consumer IoT, published in 2018.

[12] Cyber security codes of practice, Department for Science, Innovation & Technology, 2024.
[13] Government response to the call for views on software resilience and security for businesses and organisations, Department for Science, Innovation & Technology, 2024.
[14] The UK Product Security and Telecommunications Infrastructure (Product Security) regime, Department for Science, Innovation & Technology, 2023.

- Delivering the world’s first App and App Store Privacy and Security Code of Practice, which is being implemented by all major app store operators.[15]
- Building on the Capability Hardware Enhanced Risk Introduction Set Computer (RISC) Instructions (CHERI) research from the University of Cambridge, we have worked with Arm to develop a processor prototype that integrates CHERI capabilities to enable fine-grained protection of memory. This forms part of our Digital Security by Design programme.[16]

2.6 A secure by design approach is only the first step towards UK-wide cyber resilience. To build on this, we must also focus on cultivating the necessary cyber security skills. This entails aligning our cyber skills development initiatives more closely with the needs of Critical National Infrastructure (CNI) sectors, the specific risks associated with new and emerging technologies which are being adopted, and with the resilience measures that we expect of organisations across the economy. In doing so, we are seeking to foster a skilled workforce capable of deploying the baseline cyber security expectations we set across diverse sectors of the economy.

2.7 Another key part of building a more cyber resilient UK is identifying and mitigating cyber risks as they proliferate across digital supply chains. By providing guidance, we can help businesses and organisations better manage risks associated with the digital products and services on which they rely. Our work in this space includes:

- Cyber Essentials Certification, which is a Government backed scheme to certify that organisations have taken the minimum steps to protect themselves against the most common cyber attacks.
- The Cyber Assessment Framework, which supports organisations as they seek to assess cyber risks to essential functions. This is aimed at critical organisations such as those in CNI sectors.
- The Cyber Governance Code of Practice, which sets the baseline expectations for all organisations using digital technologies.

2.8 Our ability to safeguard businesses and communities from cyber threats hinges upon our ability to both nurture technological and human capabilities, and to recognise and address complex risks on a macro scale. By prioritising secure by design, skill development, and targeted measures which improve resilience across all sectors of our economy, we seek to pave the way for a more secure cyber landscape in the UK.

Supporting government activity on AI

2.9 The work outlined in this Call for Views forms part of the Governments' efforts to enable safe and trustworthy AI. An overview of the various activities being taken on AI can be found in the recently published AI regulation white paper response.[17] The response sets out how Government is implementing the regulatory framework, including by preparing and upskilling the UK’s regulators, setting out the case for targeted binding measures on developers of highly capable general purpose AI systems, and preventing the misuse of AI technologies. We have set out below the key interlinking areas of work that are being taken forward within DSIT.

2.10 The UK hosted the first AI Safety Summit in November 2023. Representatives from 28 nations, including the US, EU, and China endorsed the Bletchley Declaration, an ambitious agreement to support an internationally inclusive network of scientific research on advanced

[15] Code of practice for app store operators and app developers (updated), Department for Science, Innovation & Technology, 2023.

[16] Capability Hardware Enhanced RISC Instructions (CHERI), University of Cambridge, Department of Computer Science and Technology.

[17] A pro-innovation approach to AI regulation: government response, Department for Science, Innovation & Technology, 2024.

AI safety and ensure the benefits of technology can be harnessed for the good of all. Cyber security was an important aspect of the Summit and included in key documents such as the 'Emerging processes for AI safety' policy paper and the Bletchley Declaration itself.

2.11 During the Summit, the UK Government announced the creation of the AI Safety Institute (AISI), which will also be fundamental in informing the UK's regulatory framework. AISI's three core functions are to develop and conduct evaluations on advanced AI systems, drive foundational AI safety research and facilitate information exchange.

2.12 The Summit also emphasised the importance of understanding the risks associated with AI and this work has sought to expand this area in collaboration with the activities being taken forward by the Central AI Risk Function (CAIRF). CAIRF brings together policymakers and AI experts with a mission to identify, assess, report on and prepare for AI associated risks on an ongoing basis. CAIRF will:

- Maintain a holistic view of risks across the AI ecosystem by monitoring current and emerging AI risks facing the UK and assessing their likelihood and impact.
- Report on AI risk to inform government intervention and regularly report assessments of risks and mitigations to Ministers and ensure coordination and interoperability with the National Security Risk Assessment (NSRA).
- Ensure that HMG can effectively respond to AI risks and issues if discovered.

2.13 Data is the beating heart of AI and is the foundation for modelling, training and developing AI systems. Various functions across the UK government are taking a leading role in ensuring that AI and data can be used for good, safely. Our proposed Code of Practice and global standard will complement this work.[18]

2.14 Our work on the cyber security of AI, which is the focus of this Call for Views, builds on existing and developing efforts to tackle risks stemming from the misuse of AI technologies. Taken together, this demonstrates the UK government's ambition to ensure that the benefits of AI innovation can be realised safely and responsibly.

[18] A pro-innovation approach to AI regulation: government response, Department for Science, Innovation & Technology, 2024: this includes the Responsible Technology Adoption Unit, the Information Commissioners Office, States Threat to Data Directorate.

Chapter 3: Voluntary Code of Practice and Global Standard

3.1 Based on the evidence available, set out in Annex A, we have determined that a two-part intervention based on a voluntary Code of Practice and a global standard will help address the cyber security risks to AI. Subject to feedback received, this will require finalising a voluntary Code of Practice and developing a global standard based on this Code.[19]

3.2 It is the Government's intention, subject to the feedback of this consultation, to submit an updated voluntary Code of Practice to the European Telecommunications Standards Institute (ETSI) in September 2024 to help inform the development of this global standard on baseline cyber security for AI systems and models. Taking the voluntary Code of Practice and research (outlined in 1.17) from this work into a standards organisation will ensure the future standard is of a high quality and informed by data and extensive global feedback.

3.3 This Call for Views will only be the start of the process for stakeholders to contribute to this work. In addition to this Call for Views, we will continue to actively seek feedback and engagement as our work progresses. We also encourage stakeholders to participate in the standards development process, and to note that there will be a consultation phase for national standards bodies as part of ETSI's typical standards development process.

3.4 In addition to this Code of Practice, several other options (set out in Annex E) have been considered in support of the programme's objectives. The Government created a criteria to assess the effectiveness of each policy option, including the voluntary Code of Practice and the creation of a global standard. Each intervention was tested to determine if it would address the issues raised from the evidence findings and promote baseline security practices in the development of AI models and systems.

Rationale for a Code of Practice and global standard

3.5 A key aim of the current and previous National Cyber Strategy has been to build international support for a "secure by design" approach as well as baseline security requirements across various areas of technology.[20] A Code of Practice and global standard would help achieve this objective (as set out in Chapter 1). Moreover;

- There are clear risks to AI and it is important that these are addressed so that millions of consumers and organisations can benefit from AI technologies. Baseline security requirements will help reduce the number of cyber attacks and therefore protect users' data and the economy.

- Many organisations that are implementing, or considering adopting, AI do not have a clear understanding of what security expectations they should have from developers. A Code and technical standard will enable cyber security companies and certification firms to help companies with testing and assuring their products and services. This will enable users to more easily verify that the products they use are securely designed and developed, giving them greater confidence.

- We want to create a market ecosystem where AI supply chain stakeholders are looking to use security as a differentiator between their competitors. A technical global

[19] This follows on from National Cyber Security Centre's (NCSC) publication of their Guidelines for Secure AI Development, 2023. The Code is based on this document and the findings of the research.
[20] National Cyber Strategy, Cabinet Office, 2022.

standard will help enhance an organisation's reputation and drive better practices across the industry.

- We want to ensure the UK continues to be a leader in AI and that our market is prepared for further developments in AI. As part of this, we support the ambitions of AI developers to create more sophisticated models and systems and view cyber security as a key enabler of this. A voluntary Code and technical standard will ensure innovation and safety in AI can develop in tandem.

AI Cyber Security Code of Practice

Background

This proposed voluntary Code of Practice was developed by DSIT and is based on NCSC's Guidelines for secure AI system development which were published in November 2023, alongside the US Cybersecurity and Infrastructure Security Agency and other international cyber partners. The Code has also been informed by an assessment of the cyber security risks to AI as well as a literature review that mapped accessible technical recommendations made by industry and other governments. The findings of the literature review have been used to map relevant publications to the Code's principles to offer an indication of where there are crossovers.[21]

The Code sets out practical steps for stakeholders across the AI supply chain, particularly Developers and System Operators, to protect end-users. The Code applies to all AI technologies and will help ensure that security is effectively built into AI models and systems as well as across the AI lifecycle. We have applied this broad scope because a lot of the complexity in an AI system resides out of the model, and there is a significant dependency on data. It is important for stakeholders to note that this voluntary Code sits in support of the UK Government's wider efforts for AI and regulations, such as UK data protection law. Stakeholders across the AI supply chain must ensure that they comply with their regulatory obligations. Considering the direct interlinkage between data and security within the context of an AI model and system; both areas are addressed through the Code's requirements.

Our expectation is that organisations in scope would, at a minimum, also adhere to the provisions in both the Software and Cyber Governance Codes of Practice. While the Cyber Governance Code of Practice sets the baseline expectations for all organisations using digital technologies as part of their business function, the Software Code will also be relevant since software is an integral part of how AI models and systems function. Organisations deemed in scope of this Code would also be expected to assess whether their circumstances warrant consideration of adherence to additional Codes covering more specific products or services depending on their business function.

Millions of businesses and consumers are using AI models and systems, and it is important that they, and the global economy, can benefit from the opportunities provided by AI. It is therefore essential that as new updates are rolled out and new products come to market, security is a core consideration throughout the AI lifecycle. The Code is intended to help inform the policy and practices that organisations currently have in place.

Furthermore, we recognise that several industry and standards bodies, as well as other countries, have compiled recommendations to address the cyber security risks to AI. This voluntary Code of Practice is designed to be complementary to, and supportive of, those efforts. This is particularly important when working groups have been set up in various standards development organisations, including the Secure AI Technical Committee in the European Telecommunications Standards Institute (ETSI).

[21] Research on the cyber security of AI, Department for Science, Innovation & Technology, 2024.

As set out in the Call for Views document, we are encouraging feedback from global stakeholders. This is because the Government's intention, depending on the feedback received during the Call for Views, is to submit the updated voluntary Code to ETSI in September 2024 to help inform the development of a global standard. This Code will be reviewed, and if necessary updated, where there are changes in the technology itself, the risk landscape and the regulatory regimes. We are therefore proposing monitoring and evaluation activities to assess uptake of the Code's principles among key stakeholders (see Call for Views document – points 3.6 and 3.7).

Audience

An indication is given for each principle within this voluntary Code as to which stakeholder is primarily responsible for implementation. The stakeholders are defined as:

Stakeholder	Definitions
Developers	This encompasses any type of business or organisation across any sector as well as individuals that are responsible for creating an AI model and/or system. This applies to all AI technologies and both proprietary and open-source models. For context, a business or organisation that creates an AI model and who is also responsible for embedding/deploying that model/system in their organisation would be defined in this voluntary Code to be both a Developer and a System Operator.
System Operators	This includes any type of business or organisation across any sector that has responsibility for embedding / deploying an AI model and system within their infrastructure. This applies to all AI technologies and both proprietary and open-source models. This term also includes those businesses that provide a contractual service to organisations to embed / deploy an AI model and system for business purposes.
Data controllers	This includes any type of business, organisation or individual that control data permissions and the integrity of data that is used for any AI model or system to function. In the context of an AI system, there could be multiple data controllers involved because some data used to create a model could come from the organisation that is deploying/embedding the system in their infrastructure and other data could be from public databases and other sources.
End-users	This encompasses any employee within an organisation or business and UK consumers who use an AI model and system for any purpose, including to support their work and day-to-day activities. This applies to all AI technologies and both proprietary and open-source models. End-users are not expected or required to implement this Code. This stakeholder group has been created because the voluntary Code has placed expectations on Developers, System Operators and Data controllers to help inform and protect end-users.

The table below gives examples of common cases involving different types of organisations that are relevant to this voluntary Code of Practice as well as the Software Resilience voluntary Code of Practice.

Stakeholder Groups	Guidance
Software vendors who also offer AI services to customers/end-users	These organisations are a Developer and therefore are in scope of this Code and the Software Resilience Code of Practice.
Software vendors who use AI in their own infrastructure which has been created by an external provider	These organisations are a System Operator and therefore are in scope of relevant parts of the Code and the Software Resilience Code of Practice.
Software vendors who create AI in-house and implement it within their infrastructure	These organisations are a Developer and System Operator and therefore are in scope of this Code and the Software Resilience Code of Practice.
Software vendors who only use third-party AI (components) for their in-house use	These organisations are a System Operator and therefore are in scope of relevant parts of the Code and the Software Resilience Code of Practice.
Organisation that creates an AI system for in-house use	These organisations are a Developer and therefore are in scope of this Code.
Organisation that only uses third-party AI components	These organisations are a System Operator and therefore are in scope of relevant parts of the Code.
AI Vendors	Organisations that offer or sell models and components, but do not play a role in developing or deploying them, are not in scope of this Code. These organisations are in scope of the Software Code of Practice and Cyber Governance Code.

What does the terminology mean in the voluntary Code of Practice?

We have used “shall” and “should” terminology for each provision in the voluntary Code to align with the wording used by standards development organisations.[22] The table below sets out the definitions of these words in the context of the voluntary nature of this Code of Practice.

Shall	Indicates a requirement for the voluntary Code
Should	Indicates a recommendation for the voluntary Code
Can/could	Indicates where something is possible, for example, that an organisation or individual is able to do something

Code of Practice Principles

Secure Design

Principle 1: Raise staff awareness of threats and risks

[22] Foreword – supplementary information, ISO and A Guide to Writing World Class Standards, ETSI, 2020.

Primarily applies to: System Operators, Developers, and Data Controllers
[NIST 2022, NIST 2023, ASD 2023, WEF 2024, OWASP 2024, MITRE 2024, Google 2023, ESLA 2023, Cisco 2022, Deloitte 2023, Microsoft 2022]

1.1. Organisations shall establish and maintain a continuous security awareness program to educate their staff about the evolving threat landscape specific to AI systems.

1.1.1. The AI-Security security awareness content shall be reviewed and updated where necessary at least every six months.

1.1.2. AI-specific security awareness training could be incorporated into existing infosec training for staff.

1.2. Developer organisations should provide their staff with regular updates on the latest security threats and vulnerabilities that could impact AI systems

1.2.1. These updates should be communicated through multiple channels, such as security bulletins, newsletters, or internal knowledge-sharing platforms, to ensure broad dissemination and understanding among the staff.

1.3. Developers shall receive training in secure coding techniques specific to AI development, with a focus on preventing and mitigating security vulnerabilities in AI algorithms, models, and associated software.

1.3.1 Developer training should also include content on how developers may leverage AI/LLMs to improve code security

1.4 Developers shall receive awareness training in the characteristics of machine learning and AI systems in general that make them especially complex (and hence particularly vulnerable to technical debt and security issues) – these often include convoluted data dependencies, multi-layered software architectures, and intricate configurations.

Principle 2: Design your system for security as well as functionality and performance[23]
Primarily applies to: System Operator
[OWASP 2024, MITRE 2024, WEF 2024, ENISA 2023, NCSC 2023, BSI1 2023, Cisco 2022, Microsoft 2022, G7 2023, HHS 2021, OpenAI2 2024, ASD 2023, ICO 2020]

2.1 As part of deciding whether to create an AI system, a System Operator shall determine and document the business requirements and/or problem they are seeking to address.

2.1.1 Data controllers shall be part of internal discussions when determining the requirements and data needs of an AI system.

NCSC Guidelines for Secure AI System Development - other areas to consider include:

- The complexity of the model they are using, specifically the chosen architecture and number of parameters.
- The model’s chosen architecture and number of parameters will, among other factors, affect how much training data it requires and how robust it is to changes in input data when in use.

[23] See Guideline on “Design your system for security as well as functionality and performance” in Guidelines for secure AI system development, NCSC, 2023. Additionally, Software Vendors should review Principle 1 on “Secure design and development” for further requirements in the Software Code of Practice (footnotes referencing Software Code to be added when Gov.uk page created).

- The appropriateness of the model for their use case and/or feasibility of adapting it to their specific need (for example by fine-tuning).
- The ability to align, interpret and explain their model's outputs (for example for debugging, audit or regulatory compliance); there may be benefits to using simpler, more transparent models over large and complex ones which are more difficult to interpret.
- The characteristics of training dataset(s), including size, integrity, quality, sensitivity, age, relevance and diversity the value of using model hardening (such as adversarial training), regularisation and/or privacy enhancing techniques.
- The provenance and supply chains of components including the model or foundation model, training data and associated tools.
- See NCSC Machine Learning Principles for more information.

2.2 To support the process of preparing data for an AI system, Developers shall document and audit trail the creation, operation, and life cycle management of models, datasets and prompts incorporated into the system.

2.3 If a Developer and/or System Operator decides to use an external Application Programming Interface (API), they shall apply appropriate controls to data that can be sent to services outside of their organisation's control, such as requiring users to log in and confirm before sending potentially sensitive information.

2.4 Data controllers shall ensure that the intended usage of the system is appropriate with the sensitivity of the data it was controlled on as well as the controls intended to ensure the safety of data.

2.5 Where the AI system will be interacting with other systems, (be they internal or external), Developers and System Operators shall ensure that the permissions used by the system are only provided as required for functionality and are risk assessed.

This includes ensuring identities used by the AI system are constrained in scope and privilege to the access required. This includes external AI and non-AI fail-safes if necessary.

2.6 If a System Operator chooses to work with an external model provider, they shall undertake a due diligence assessment of that provider's security.

This assessment could involve implement scanning and isolation/sandboxing when importing third-party models, serialised weights or untrusted third-party code.

2.7 If a Developer and/or System Operator decides to use an external library, they shall complete a due diligence assessment.[24]

This assessment could consider whether the model can be obtained as a safe model and if not, then doing checks to ensure the library has controls that prevent the system loading untrusted models without immediately exposing itself to arbitrary code execution.

[24] Guidelines for secure AI System Development, NCSC, 2023: This will help ensure the library has controls that prevent the system loading untrusted models without immediately exposing themselves to arbitrary code execution.

Principle 3: Model the threats to your system[25]

Primarily applies to: Developers and System Operators

[OWASP 2024, WEF 2024, Nvidia 2023, ENISA 2023, Google 2023, G7 2023, NCSC 2023, Deloitte 2023]

3.1 Developers and System Operators shall undertake modelling of the threats to a system as part of their risk management process.

NCSC Guidelines for Secure AI System Development: This modelling could include understanding the potential impacts to all AI-responsible stakeholders, end-users, and wider society if an AI component becomes compromised or behaves unexpectedly. Additionally, the modelling could be informed by AI-specific attacks and failure modes, as well as more traditional IT system attacks. The modelling could factor the total range of possible outputs, (including worst case scenarios), from AI components and their impact on the system.

3.1.1 The risk management process shall be conducted to address any security risks that arise when a new setting or configuration option is implemented at any stage of the AI lifecycle.

3.1.2 As part of this process, Developers shall create a document that includes a list of adversarial motivations and possible attack routes in line with those motivations.

The type of attacks could include indirect attacks where attackers poison data which might later be used by, or sent to, the model.

3.1.3 Developers shall manage the risks associated with models that provide multiple functionality, where increased functionality leads to increased risk. For example, where a multi-modal model is being used but only single modality is used for system function.

3.2 Data controllers should conduct a data protection impact assessment when necessary as a measure under UK data protection obligations to determine what controls are needed.

3.3 Where threats are identified that cannot be resolved by Developers, this shall be communicated to System Operators and End-users to allow them to appropriately threat model their systems.

3.4 Where third-party organisations have responsibility for risks identified within an organisations infrastructure, System Operators should attain assurance that these parties are able to address the risk.

3.5 System Operators should seek to apply controls to risks identified through the analysis based on a range of considerations, including the cost of implementation in line with their corporate risk tolerance.

3.6 Developers and System Operators should recognise and accept that a level of risk will remain despite the application of controls to mitigate against them, and continuously monitor and review their system infrastructure according to risk appetite.

[25] See “Model the threats to your system” and “Raise staff awareness of threats and risks” in Guidelines for secure AI system development, NCSC, 2023. Additionally, Software Vendors should review Principle 1 on “Secure design and development” for further requirements in the Software Code of Practice.

Principle 4: Ensure decisions on user interactions are informed by AI-specific risks[26]

Primarily applies to: Developers and System Operators

[OWASP 2024, MITRE 2024, BSI1 2023, Microsoft 2022]

4.1 Developers and System Operators shall ensure that their system provides effective safeguards around model outputs through non-AI components and processes.

This could also include the use of trained human oversight.

4.2 Developers shall take steps to validate that the designed controls specified by the Data Controller have been built into the system.

4.3 Developers should consider placing limits on the rate of model access (e.g. via APIs) to prevent attacks based on experimentation, and limit resource usage for single model inputs to prevent the overuse of resources.

4.4 Developers and System Operators should ensure end-users are aware of prohibited use cases of the AI system.

4.5 Developers and System Operators should be transparent with end-users about known limitations or potential failure modes to protect against overreliance.

4.6 If a Developer offers an API to external customers or collaborators, they shall apply appropriate controls that mitigate attacks on the AI system via the API.

Secure Development

Principle 5: Identify, track and protect your assets[27]

Primarily applies to: Developers, System Operators and Data Controllers

[OWASP 2024, Nvidia 2023, NCSC 2023, BSI1 2023, Cisco 2022, Deloitte 2023, Amazon 2023, G7 2023, ICO 2020]

5.1 Developers, Data Controllers and System Operators shall know where their assets reside and have assessed and accepted any associated risks as they evolve.

These assets could include AI models, data (including user feedback), prompts, software, documentation, logs and assessments (including information about potentially unsafe capabilities and failure modes).

5.2 Developers, Data Controllers and System Operators shall have processes and tools to track, authenticate, manage version control and secure their assets.

5.3 System Operators shall have the ability to restore their systems to a known good state in the event of compromise.

5.4 All responsible stakeholders in this Code shall take steps to protect sensitive data, such as training or test data, against unauthorised access (see 6.2 and 6.2.1 for details on securing your data and other assets).

[26] See “Design your system for security as well as functionality and performance” in Guidelines for secure AI system development, NCSC, 2023. Additionally, Software Vendors should review Principle 1 on “Secure design and development” for further requirements in the Software Code of Practice.

[27] See “Identify, track and protect your assets” in Guidelines for secure AI system development, NCSC, 2023. Additionally, Software Vendors should review Principle 1 on “Secure design and development” for further requirements in the Software Code of Practice.

5.4.1 Developers, Data Controllers and System Operators shall apply checks and sanitisation to data and inputs when designing the model [based on their access to said data and inputs and where those data and inputs are stored]. This shall be repeated when model revisions are made in response to user feedback or continuous learning [See principle 6 for relevant provisions for open source].

Principle 6: Secure your infrastructure[28]

Primarily applies to: Developers and System Operators

[OWASP 2024, MITRE 2024, WEF 2024, NCSC 2023, Microsoft 2022, ICO 2020]

6.1 Alongside implementing essential cyber security practices for securing system infrastructure[29], Developers and System Operators shall adopt appropriate access controls to their APIs, models and data, and to their training and processing pipelines.[30]

6.2 Developers and System Operators shall create segregated environments to enforce sensitivity and threat boundaries.

6.2.1 Developers, System Operators and Data Controllers shall create segregated environments for storing critical data, such as sensitive, training or test data [where this training data is not based on publicly available data – see 7.3.1 and 7.3.2].

6.2.2 Developers shall also create a segregated environment for where research is done and where production models are developed and/or accessed.

Stakeholders could use containerisation and virtualisation as methods for segregation. See NCSC containerisation guidance: https://www.ncsc.gov.uk/collection/using-containerisation

6.3 Developers and System Operators shall implement and publish an effective vulnerability disclosure process to support a transparent and open culture within the organisation.[31]

6.4 Developers and System Operators shall create an incident management plan.

Principle 7: Secure your supply chain[32]

Primarily applies to: Developers, System Operators and Data Controllers

[OWASP 2024, NCSC 2023, Microsoft 2022, ASD 2023]

7.1 Developers and System Operators shall require suppliers to adhere to the same security expectations and requirements that they apply to other software components to develop new software/AI products. This shall align with their risk management policies.

[28] See "Secure your infrastructure" in Guidelines for secure AI system development, NCSC, 2023.

[29] Organisations should review DSIT's Cyber Governance Code of Practice, NCSC's Business Toolkit and Cyber Essentials.

[30] Strategies include: encryption of data at rest, technical access controls for the data to limit access according to least privilege principles, centralised access controls for the data, operational security to protect stored data, logging and monitoring to detect suspicious manipulation of data (e.g. outside of office hours).

[31] Software Vendors should review Principle 3 on Secure deployment and maintenance for further requirements in Software Code of Practice.

[32] See "Secure your supply chain" in Guidelines for secure AI system development, NCSC, 2023. Additionally, Software Vendors should review Principle 1 on "Secure design and development" for further requirements in the Software Code of Practice

7.2 If a component is not produced in-house, Developers and System Operators should acquire and maintain well-secured and well-documented hardware and software components (for example, models, data, software libraries, modules, middleware, frameworks, and external APIs) from verified commercial, open-source, and other third-party developers to ensure robust security in your systems.

7.2.1 Developers that choose to use any models, or components, which are not well-documented or secured shall be able to justify why, (for example if there was no other supplier for said component), and be prepared to share this explanation with end-users, and System Operators if required.

Particular attention should be given to the use of open-source models, where the responsibility of model maintenance and security becomes complex.

7.3 Developers and System Operators should be prepared to failover to alternate solutions for mission-critical systems, if their security criteria are not met.[33]

7.3.1 Where training data has been sourced from publicly available sources, Developers and Data controllers shall need to validate that such training data will not compromise the integrity of their security protocols.

7.3.2 Data controllers should continually monitor the source of publicly available data that could be used for creating a model, such as for changes in the data sources that may risk creating vulnerabilities.

Principle 8: Document your data, models and prompts[34]

Primarily applies to: Developers

[OWASP 2024, WEF 2024, NCSC 2023, Cisco 2022, Microsoft 2022, ICO 2020]

8.1 Developers shall document and maintain a clear audit trail of their model design and post-deployment maintenance plans.

8.1.1 Developers should ensure that the document includes security-relevant information, such as the sources of training data (including fine-tuning data and human or other operational feedback), intended scope and limitations, guardrails, cryptographic hashes or signatures, retention time, suggested review frequency and potential failure modes.

8.1.2 Developers should pay particular attention to document areas of model and system complexity that could lead to unexpected security issues, including details of software dependencies and configurations.

8.2 Developers should ensure that model outputs include only the necessary information for downstream purposes and do not include additional meta-data that might be used for honing attacks against the model.

[33] See Supply chain security guidance, NCSC, 2018, and frameworks such as Safeguarding artifact integrity across any software supply chain, Supply Chain Levels for Software Artifacts (SLSA), for tracking attestations of the supply chain and software development life cycles.

[34] See "Document your data, models and prompts" in Guidelines for secure AI system development, NCSC, 2023. Additionally, Software Vendors should review Principle 1 on "Secure design and development" for further requirements in the Software Code of Practice.

Principle 9: Conduct appropriate testing and evaluation[35]
Primarily applies to: Developers
[OWASP 2024, WEF 2024, Nvidia 2023, NCSC 2023, ENISA 2023, Google 2023, G7 2023]

9.1 Developers shall ensure that no models, applications or systems are released that haven't been tested as part of a security assessment process.

9.2 Developers shall validate that AI models perform as intended through testing.

9.2.1 Developers shall work closely with System Operators for post-deployment testing when maintaining a system. (see 2.1.2 for more details)

9.2.2 Evaluations of AI systems should involve red teaming or other adversarial testing as part of a whole system approach.

9.2.3 Evaluations of AI systems should be undertaken by suitably skilled testers. Where possible, this should be an independent external evaluation.

9.3 Developers should perform benchmarking as part of their risk management process throughout the AI development lifecycle (see principle 2 for more detail).

9.4 Developers should ensure that the findings from the testing and evaluation are shared with System Operators, to inform their own testing and evaluation.

Secure Deployment

Principle 10: Communication and processes associated with end-users[36]
Primarily applies to: Developers and System Operators

10.1 In the context of AI, Developers shall state clearly to end-users (where possible) which aspects of security the end-user is responsible for and are transparent about where and how their data might be used, accessed or stored (for example, if it is used for model retraining, or reviewed by employees or partners).

10.2 Developers should ensure that the organisation proactively supports affected End-users and System Operators during and following a cyber security incident to contain and mitigate the impacts of an incident. The process for undertaking this should be documented and agreed in contracts with end-users.

10.3 Developers should provide end-users with guidance on how to use, manage, integrate, and configure the software product or service securely.

10.3.1 In the context of AI, this should include the appropriate use of your model or system, which includes highlighting limitations and potential failure modes.

[35] See "Release AI responsibly" in Guidelines for secure AI system development, NCSC, 2023. Additionally, Software Vendors should review Principle 1 on "Secure design and development" for further requirements in the Software Code of Practice. See also AI Safety Institute for testing AI models at the Frontier.

[36] See "Make it easy for users to do the right things" and "Develop incident management procedures (for further information)" in Guidelines for secure AI system development, NCSC, 2023. Additionally, Software Vendors should review Principle 4 on "Communication with customers" for further requirements in the Software Code of Practice.

10.3.2 Moreover, Developers shall inform end-users of additional AI model functionality, and allow an opt-out option.

Secure Maintenance

Principle 11: Maintain regular security updates for AI model and systems[37]

Primarily applies to: Developers and System Operators

[ICO 2020]

11.1 Developers and System Operators shall ensure that when documenting their project requirements, their plans include conducting regular security audits and updates and working with external providers (where needed) to achieve this.

11.2 Developers shall provide security updates and patches, where possible, and notify System Operators and End-users of the security updates.

11.2.1 In instances where updates can't be provided, Developers shall have mechanisms for escalating issues to the wider community, particularly customers and other Developers.

To help deliver this, they could publish bulletins responding to vulnerability disclosures, including detailed and complete common vulnerability enumeration.

11.3 Developers should treat major system updates as though a new version of a model has been developed, and therefore undertake a new testing and evaluation process for each to help protect users.

11.4 Developers should support System Operators to evaluate and respond to model changes, (for example by providing preview access via beta-testing and versioned APIs).

Principle 12: Monitor your system's behaviour

Primarily applies to: Developers and System Operators

[OWASP 2024, WEF 2024, Nvidia 2023, ENISA 2023, BSI1 2023, Cisco 2022, Deloitte 2023, G7 2023, Amazon 2023, ICO 2020]

12.1 In line with privacy and data protection requirements, Systems Operators should log all inputs and outputs to/from their AI system to enable auditing, compliance obligations, investigation and remediation in the case of compromise or misuse.

12.2 System Operators and Developers should also consider logging internal states of their AI models where they feel this could better enable them to address security threats, or to enable future security analytics.

12.3 System Operators and Developers should monitor the performance of their models and system over time so that they can detect sudden or gradual changes in behaviour that could affect security.

This can be achieved by using tools that detect anomalous inputs that will skew outputs, without knowing what malicious input looks like. There are specific methods that could be implemented to mitigate input that is out of distribution or invalid, such as outlier detection, anomaly detection, novelty detection, and open set recognition.

[37] See "Follow a secure by design approach to updates" in Guidelines for secure AI system development, NCSC, 2023. Additionally, Software Vendors should review Principle 3 on "Secure deployment and maintenance" for further requirements in the Software Code of Practice.

12.4 System Operators and Developers should analyse their logs to ensure that AI models continue to produce desired outputs over time.

Monitoring and Evaluation

3.6 The Code has been designed in line with our pro-innovation approach. Our intention is to allow flexibility via a principles-based approach when implementing the provisions and adaptation overtime as the area develops to reduce the burden on stakeholders in the AI supply chain. Further details on its scope and the stakeholders that must adhere to it can be found within the Code.

3.7 The proposed Code of Practice would be voluntary; however, we will continue to work closely with interested stakeholders to determine if regulatory action is needed for AI in the future. To help support this, the Government's intention is to undertake monitoring and evaluation uptake of the Code and its effectiveness at encouraging the outcomes that we hope to see in the AI ecosystem.

Chapter 4: How to respond to the Call for Views and our next steps

4.1 DSIT will be holding an eight-week Call for Views on the document from 15 May to 10 July 2024. Stakeholders are invited to provide specific feedback on the interventions and make recommendations regarding other policy options via an online survey form. Please see Annex D for the full Call for Views Survey Questionnaire. We encourage stakeholders, particularly in the AI supply chain, to provide data on the financial and wider impacts associated with the implementation of the Government's proposed interventions. All data will be treated confidentially, and participants will have the opportunity to identify themselves when they submit their responses or to be anonymous.

4.2 During the Call for Views, DSIT will be organising workshops with industry bodies and holding meetings with international counterparts as part of our efforts to promote this work. We are also planning to attend UK and international conferences, including on specific panels so that we can present our approach to a global audience. If you would like to bring any related events/conferences to our attention or if you have any questions on the online survey please contact AIcybersecurity@dsit.gov.uk (by 10 July 2024). You can also submit written comments to the AI Cyber Security Call for Views, Secure Code and Standards, Cyber Security & Digital Identity Directorate, Department for Science, Innovation & Technology, Level 4, 100 Parliament Street, Westminster, London, SW1A 2BQ. DSIT plans to arrange workshops with industry to help gather feedback.

4.3 Following the Call for Views, we will review the feedback provided. We plan to publish a response which provides an overview of the key themes from the Call for Views and DSIT's future direction of travel. If there is support for a global technical standard, we will look to take this forward alongside increasing our participation across global standard development organisations.

Annex A: Research findings

Overview of research studies and objectives

The research field of AI Security is still nascent and has developed for the last 5 years. DSIT has commissioned targeted research to establish an initial evidence base to inform the development of our policy interventions, notably the Code of Practice.[38]

The specific studies involved:

- An assessment of the cyber security risks to AI by Grant Thornton and Manchester Metropolitan University was completed in February 2024.[39]
- A survey of 350 UK businesses to understand how organisations are approaching AI, particularly regarding cyber security. The survey was conducted by IFF Research and ran from January to February 2024.[40]
- A literature review mapping the technical and policy recommendations made by industry and other governments. This was completed in February 2024, conducted by Professor Peter Garraghan of Mindgard, using information published since 2020.[41]
- A literature review of research on the cyber security of AI involving an in-depth analysis of more than 415 publications, completed in February 2024 by Queen's University Belfast.[42]

The key findings from the research are outlined below.

- The vulnerabilities and threats across various AI technologies are broadly similar to each other.
- The exploitation of vulnerabilities in an AI system can have a substantial impact on end-users, such as the loss of sensitive data linked to consumers and employees as well as providing malicious actors with a way of breaching an organisation's infrastructure.
- Organisations generally lack awareness and understanding of what security should be built into models and systems and whether practices/processes should be in place when adopting AI to protect their organisations.
- Key organisations, governments and standards development organisations advocate for security requirements for AI models and systems.
- Vulnerabilities found in AI systems can enable the models and systems to be weaponised which can result in cyber attacks and significant harms on users.[43]
- The majority of the research conducted in the field of the Security of AI is being conducted by academic institutions. Out of the 415 sources on the cyber security of AI fully analysed by Queens University Belfast, only 28% were created by industry organisations.

[38] Research on the cyber security of AI, Department for Science, Innovation & Technology, 2024

[39] Cyber Security Risks to Artificial Intelligence, Department for Science, Innovation & Technology, 2024.

[40] AI Cyber Security Survey, Department for Science, Innovation & Technology, 2024.

[41] A few research papers published from 2014 onwards were also included. See Cyber Security for AI Recommendations, Department for Science, Innovation & Technology, 2024.

[42] Study of Research and Guidance on the Cyber Security of AI, Department for Science, Innovation & Technology, 2024.

[43] A pro-innovation approach to AI regulation: government response, Department for Science, Innovation & Technology, 2024: See work linked to this area which is being led by DSIT's AI Directorate and Cabinet Office.

Threats to AI technologies

The assessment of the cyber security risks to AI was commissioned to provide DSIT with an up-to-date analysis of the risk landscape of AI technologies. The assessment was formed from the findings of two literature reviews and subsequent interviews. The first literature review mapped and evaluated any previous research on the cyber security risks of AI, including known vulnerabilities. The second literature review examined government and industry reports to help frame the study.

The assessment found that there were vulnerabilities throughout the AI lifecycle, across design, development, deployment and maintenance. The assessment also highlighted that there are significant commonalities in the defensive threats that are faced across all AI technologies.

The assessment also examined how each of the vulnerabilities could be exploited and outlined the potential impacts that could result from this. These impacts included risking sensitive user data and breaches of an organisation's network.

The assessment mapped the various cyber attacks that had been conducted against AI technologies based on proof-of-concept publications and real-life case studies. It identified a total of 22 examples. Although only a few incidents were real-life examples, the assessment clearly highlighted the substantial effect that vulnerabilities can have on the safety of end-users.

Organisational awareness of cyber security for AI systems

The IFF Research survey was made up of 350 businesses who were either considering or had already adopted an AI technology within their infrastructure[44]. The survey included various cyber security questions, including whether the businesses had specific cyber security practices or processes in place for AI. Notably, nearly half of respondents (47%) had no specific cyber security practices in place specifically for AI and 13% were unsure. Of those without, or not intending to have, specific AI cyber security practices or processes, there were a few key reasons as to why they had not adopted specific practices. 14% had not considered it or did not know enough about it, and 14% said they do not use AI for anything sensitive. These findings highlight that a significant number of organisations lack cyber security practices for AI.

The survey also asked participants whether there were specific cyber security requirements or features that they expect to be built into AI companies' models and systems. Two fifths (39%) of respondents stated no and a third (33%) were unsure. This further highlights that businesses in the UK lack cyber security knowledge of AI practices and processes.

Security requirements for AI models and systems are needed

Mindgard's literature review sought to identify any recommendations that would help address the cyber security risks to AI. The report identified 67 sources that described 45 unique recommendations. These recommendations were divided into two categories, technical and general. It was evident that security requirements are seen as essential for this policy area, based on the overlapping requirements highlighted by industry and governments. Notably, the author stated "there is evidence that many of the reported cyber security risks within AI strongly justify the need to identify, create, and adopt new recommendations to address them." The assessment of cyber security risks to AI also highlighted this finding, as the report linked vulnerabilities, and their exploitation routes, to poor design and development practices.

[44] The survey focussed on seven key sectors, and therefore findings are not representative of the overall business population.

Based on the findings set out in this Annex, the evidence has suggested there is rationale for the Government to intervene to improve the security of AI.

Annex B: Global approach to AI

The Prime Minister set out the government's approach to managing frontier AI risk in October 2023. He stated: *"My vision, and our ultimate goal, should be to work towards a more international approach to safety, where we collaborate with partners to ensure AI systems are safe before they are released"*. International collaboration is a core element of this work to ensure that the cyber security requirements for AI are internationally aligned where appropriate. This annex details some key multilateral initiatives and international developments informing our thinking during this programme of work. We will continue to engage with a broad range of countries via bilateral dialogues as well as multilateral fora and initiatives.

This international and collaborative approach underpins the UK's effort as an established leader on AI, as demonstrated by hosting the first ever AI Safety Summit in 2023. Following the White House Voluntary AI commitments and building on the AI Safety Summit, the NCSC published their Guidelines for Secure AI System Development.[45] Endorsed by 18 countries, these Guidelines were developed by NCSC and CISA alongside industry experts and 23 international agencies and ministries, with representation from across the world. We will continue to promote international conversations on AI cyber security in the lead up to future AI Safety Summits. We will also continue to engage with a broad range of countries via bilateral dialogues and multilateral fora and initiatives to further inform our own thinking.

Multilateral

The UK is taking a proactive role within multilateral discussions that link to AI cyber security and promote safe and responsible AI development, deployment and use across the world to protect citizens and our democratic values. These multilateral initiatives include:

- **G7:** Under Japan's Presidency, the G7 launched the "Hiroshima AI Process" to address the risks, challenges and opportunities posed by AI. The UK was an active participant in the Hiroshima Process and will look to build on this positive progress under Italy's Presidency.
- **G20**: In September 2023, as part of India's G20 Presidency, the UK Prime Minister agreed to and endorsed the New Delhi Leaders' Declaration, reaffirming the UK's commitment to the 2019 G20 AI Principles and emphasised the importance of a governance approach that balances the benefits and risks of AI and promotes responsible AI for achieving the UN Sustainable Development Goals. The UK will work closely with Brazil on their AI ambitions as part of their 2024 G20 Presidency, which will centre on AI for inclusive sustainable development.[46]
- **Global Partnership on Artificial Intelligence (GPAI)**: The UK continues to actively shape GPAI's multi-stakeholder project-based activities to guide the responsible development and use of AI grounded in human rights, inclusion, diversity, innovation, and economic growth. The UK was pleased to attend the December 2023 GPAI Summit in New Delhi, represented by the Minister for AI, Viscount Camrose, and to both endorse the GPAI New Delhi Ministerial Declaration and host a side-event on outcomes and next steps following the AI Safety Summit. The UK has also begun a two-year mandate as a Steering Committee member and will work with India's Chairmanship to ensure GPAI is reaching its full potential.
- **Council of Europe:** The UK is continuing to work closely with like-minded nations on a Council of Europe Convention on AI to protect and promote human rights,

[45] FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, The White House, 2023.
[46] G20 New Delhi Leaders' Declaration, Ministry of External Affairs, 2023.

democracy, and the rule of law. The Convention offers an opportunity to ensure these important values are codified internationally as one part of a wider approach to effective international governance.[47]

- **OECD:** The UK is an active member of the Working Party on AI Governance (AIGO) and recognises its role in supporting the implementation of the OECD AI Principles and enabling the exchange of experience and best practice across member countries. In 2024, the UK will support the revision of the OECD AI Principles and continue to provide case studies from the UK's Portfolio of AI Assurance Techniques to the OECD's Catalogue of Tools and Metrics of Tools for Trustworthy AI.
- **UN and its associated agencies:** Given the organisation's unique role in convening a wide range of nations, the UK recognises the value of the UN-led discussions on AI and engages regularly to shape global norms on AI. In July 2023, the UK initiated and chaired the first UN Security Council briefing session on AI, and the Deputy Prime Minister chaired a session on frontier AI risks at UN High Level Week in September 2023. The UK continues to collaborate with a range of partners across UN AI initiatives, including negotiations for the Global Digital Compact, which aims to facilitate the Sustainable Development Goals through technologies such as AI, monitoring the implementation of the UNESCO Recommendation on the Ethics of AI, and engaging constructively at the International Telecommunication Union, which hosted the 'AI for Good' Summit in July 2023. The UK will also continue to work closely with the UN AI Advisory Body and is closely reviewing its interim report: Governing AI for Humanity.

Technical Standards

Technical standards are an important tool in supporting global governance of technology, international trade, and technology innovation and can be a tool alongside or in place of regulation. The AI regulation white paper identified a key role for technical standards and assurance techniques to support the implementation of the proposed AI regulatory framework, while enhancing global interoperability. Work on technical standards relating to AI is underway within multiple global standards development organisations, and the UK government is an active participant in these discussions. Cyber security standards, supported by industry and international partners, have an important role to play in driving change and protecting users. The UK actively works to uphold integrity in and reinforce the multi-stakeholder, industry-led global digital standards ecosystem which is open, transparent, and consensus-based. The UK aims to support innovation and strengthen a multi-stakeholder, industry-led model for the development of AI technical standards, including through initiatives such as the UK's AI Standards Hub.[48]

Working groups exploring aspects of AI are active in multiple standards development organisations, including the British Standards Institute (BSI), the European Telecommunications Standards Institute (ETSI), the International Organization for Standardisation (ISO), the European Committee for Standardization (CEN), the European Committee for Electrotechnical Standardisation (CENELEC), the International Telecommunications Union (ITU), the Institute of Electrical and Electronics Engineers Standards Association (IEEE SA), the 3rd Generation Partnership Project (3GPP) and the Internet Engineering Task Force (IETF). We are actively monitoring a number of these working groups and considering how we can effectively support efforts on technical standards relating to AI cyber security. At ETSI, we have led the creation of documents on secure AI principles, including the ETSI GR SAI 002 on Data Supply Chain Security.

We welcome any contributions that AI developers and companies can make towards identifying, and coming to a consensus, on best practice. Several high-profile organisations and AI developers have produced documentation which we have used to inform the Code of

[47] CAI – Committee on Artificial Intelligence, Council of Europe.
[48] About the AI Standards Hub, The AI Standards Hub.

Practice. We want to continue to work with UK industry leaders to ensure that we stay at the forefront of AI security.

We welcome international engagement and dialogue on this topic and will collaborate, support and share information with the international community as we all look to ensure we extract the best from AI and realise its full potential.

Annex C: Glossary of terms

AI or **AI system** or **AI technologies**: products and services that are ‘adaptable’ and ‘autonomous’.

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).

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 risks: The potential negative or harmful outcomes arising from the development or deployment of AI systems.

Application Programming Interface (API): A set of rules and protocols that enables integration and communication between AI systems and other software applications.

Capabilities: The range of tasks or functions that an AI system can perform and the proficiency with which it can perform them.

Evaluations: systematic assessments of an AI system’s performance, capabilities, or safety features. These could include benchmarking.

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.

Frontier AI: For the AI Safety Summit, we defined frontier AI as models that can perform a wide variety of tasks and match or exceed the capabilities present in today’s most advanced models.

Input [to an AI system]: The data or prompt fed into an AI system, often some text or an image, which the AI system processes before producing an output.

Large Language Models (LLMs): Machine learning models trained on large datasets that can recognise, understand, and generate text and other content.

Prompt: an input to an AI system, often a text-based question or query, that the system processes before it produces a response.

Segregated environments: Segregation is a process that separates critical environments and other less sensitive environments.

Weights: parameters in a model are akin to adjustable dials in the algorithm, tweaked during training to help the model make accurate predictions or decisions based on input data, ensuring it learns from patterns and information it has seen before.

Annex D: Call for Views Survey Questionnaire

Demographics

Q1. Are you responding as an individual or on behalf of an organisation?

- Individual
- Organisation

Q2. [if individual] Which of the following statements best describes you?

- Cyber security/IT professional
- Developer of AI components
- Software engineer
- Data scientist
- Data engineer
- Senior leader in a company
- Consumer expert
- Academic
- Interested member of the public
- Government official (including regulator)
- Other (please specify)

Q3. [if organisation/business] Which of the following statements describes your organisation? Select all that apply.

- Organisation/Business that develops AI for internal use only
- Organisation/Business that develops AI for consumer and/or enterprise use
- Organisation/Business that does not develop AI, but has adopted AI
- Organisation/Business that plans to adopt AI in the future
- Organisation/Business that has no plans to adopt AI
- A cyber security provider
- An educational institution
- A consumer organisation
- A charity
- Government
- Other (please specify)

Q4. [if organisation], What is the size of your organisation?

- Micro (fewer than 10 employees)
- Small (10-49 employees)
- Medium (50-499 employees)
- Large (500+ employees)

Q5. [if individual], Where are you based?

- United Kingdom
- Europe (excluding the United Kingdom)
- North America
- South America
- Africa
- Asia
- Oceania

- Other (please specify)

Q6. [if organisation], Where is your organisation headquartered?

- United Kingdom
- Europe (excluding the United Kingdom)
- North America
- South America
- Africa
- Asia
- Oceania
- Other (please specify)

Call for Views Questions

Q7. In the Call for Views document, the Government has set out our rationale for why we advocate for a two-part intervention involving the development of a voluntary Code of Practice as part of our efforts to create a global standard focused on baseline cyber security requirements for AI models and systems. The Government intends to align the wording of the voluntary Code's content with the future standard developed in the European Telecommunications Standards Institute (ETSI).

Do you agree with this proposed approach?

- Yes
- No
- Don't know

[If no], please provide evidence (if possible) and reasons for your answer.

Q8. In the proposed Code of Practice, we refer to and define four stakeholders that are primarily responsible for implementing the Code. These are Developers, System Operators, Data Controllers (and End-users).

Do you agree with this approach?

- Yes
- No
- Don't know

Please outline the reasons for your answer.

Q9. Do the actions for Developers, System Operators and Data Controllers within the Code of Practice provide stakeholders with enough detail to support an increase in the cyber security of AI models and systems?

- Yes
- No
- Don't know

Please outline the reasons for your answer.

The next questions are going to ask you specifically about the Code of Practice that has been designed and proposed by DSIT. There will be a question on whether you support the inclusion of each principle in the Code of Practice and whether you have any feedback on the provisions in each principle.

Q.10 Do you support the inclusion of Principle 1: “Raise staff awareness of threats and risks within the Code of Practice?

- Yes
- No
- Don’t know

[If Yes], please set out any changes you would suggest on the wording of any provisions in the principle.

[If No], please provide the reasons for your answer.

Q11. Do you support the inclusion of Principle 2: “Design your system for security as well as functionality and performance” within the Code of Practice?

- Yes
- No
- Don’t know

[If Yes], please set out any changes you would suggest on the wording of any provisions in the principle.

[If No], please provide the reasons for your answer.

Q12. Do you support the inclusion of Principle 3: “Model the threats to your system” within the Code of Practice?

- Yes
- No
- Don’t know

[If Yes], please set out any changes you would suggest on the wording of any provisions in the principle.

[If No], please provide the reasons for your answer.

Q13. Do you support the inclusion of Principle 4: “Ensure decisions on user interactions are informed by AI-specific risks” within the Code of Practice?

- Yes
- No
- Don’t know

[If Yes], please set out any changes you would suggest on the wording of any provisions in the principle.

[If No], please provide the reasons for your answer.

Q14. Do you support the inclusion of Principle 5: “Identify, track and protect your assets” within the Code of Practice?

- Yes
- No
- Don’t know

[If Yes], please set out any changes you would suggest on the wording of any provisions in the principle.

[If No], please provide the reasons for your answer.

Q15. Do you support the inclusion of Principle 6: "Secure your infrastructure" within the Code of Practice?

- Yes
- No
- Don't know

[If Yes], please set out any changes you would suggest on the wording of any provisions in the principle.

[If No], please provide the reasons for your answer.

Q16. Do you support the inclusion of Principle 7 "Secure your supply chain" within the Code of Practice?

- Yes
- No
- Don't know

[If Yes], please set out any changes you would suggest on the wording of any provisions in the principle.

[If No], please provide the reasons for your answer.

Q17. Do you support the inclusion of Principle 8: "Document your data, models and prompts" within the Code of Practice?

- Yes
- No
- Don't know

[If Yes], please set out any changes you would suggest on the wording of any provisions in the principle.

[If No], please provide the reasons for your answer.

Q18. Do you support the inclusion of Principle 9: "Conduct appropriate testing and evaluation" within the Code of Practice?

- Yes
- No
- Don't know

[If Yes], please set out any changes you would suggest on the wording of any provisions in the principle.

[If No], please provide the reasons for your answer.

Q19. Do you support the inclusion of Principle 10: "Communication and processes associated with end-users" within the Code of Practice?

- Yes
- No
- Don’t know

[If Yes], please set out any changes you would suggest on the wording of any provisions in the principle.

[If No], please provide the reasons for your answer.

Q20. Do you support the inclusion of Principle 11: “Maintain regular security updates for AI models and systems” within the Code of Practice?

- Yes
- No
- Don’t know

[If Yes], please set out any changes you would suggest on the wording of any provisions in the principle.

[If No], please provide the reasons for your answer.

Q21. Do you support the inclusion of Principle 12: “Monitor your system’s behaviour and inputs” within the Code of Practice?

- Yes
- No
- Don’t know

[If Yes], please set out any changes you would suggest on the wording of any provisions in the principle.

[If No], please provide the reasons for your answer.

Q22. Are there any principles and/or provisions that are currently not in the proposed Code of practice that should be included?

- Yes
- No
- Don’t know

[If Yes], please provide details of these principles and/or provisions, alongside your reasoning.

Q23. [If you are responding on behalf of an organisation] Where applicable, would there be any financial implications, as well as other impacts, for your organisation to implement the baseline requirements?

- Yes
- No
- Don’t know

[If yes], please provide any data to explain this. This will help the Government to quantify the impact of the Code and its requirements on different types of organisations.

Q24. Do you agree with DSIT's analysis of alternative actions the Government could take to address the cyber security of AI, which is set out in Annex E within the Call for Views document?

- Yes
- No
- Don't know

[If no], please provide further details to support your answer.

Q25. Are there any other policy interventions not included in the list in Annex E of the Call for Views document that the Government should take forward to address the cyber security risks to AI?

- Yes
- No
- Don't know

[If yes], please provide further details to support your answer.

Q26. Are there any other initiatives or forums, such as in the standards or multilateral landscape, that that the Government should be engaging with as part of its programme of work on the cyber security of AI?

- Yes
- No
- Don't know

[If yes], please provide evidence (if possible) and reasons for your answer.

Q27. Are there any additional cyber security risks to AI, such as those linked to Frontier AI, that you would like to raise separate from those in the Call for Views publication document and DSIT's commissioned risk assessment (see gov.uk)? Risk is defined here as *"The potential for harm or adverse consequences arising from cyber security threats and vulnerabilities associated with AI systems".*

- Yes
- No

[If yes], please provide evidence (if possible) and reasons for your answer.

Q28. Thank you for taking the time to complete the survey. We really appreciate your time. Is there any other feedback that you wish to share?

- Yes
- No

[If yes], Please set out your additional feedback.

Annex E: Other interventions considered

Criteria that informed the development of the policy interventions

The Government has created criteria to assess the effectiveness of each policy option. Each intervention was tested to determine if it would address the issues raised from the evidence findings (see Annex A) and promote the behaviours we want developers of AI models (see above) and those seeking to deploy AI to follow. We have included questions in the Call for

Views questionnaire to gather further information to enhance our knowledge on the potential efficacy and impacts of each proposed intervention.

Benefit vs Cost (to businesses, consumers and government) / potential value for money
We evaluated each intervention within the context of its impact on addressing the cyber risks to AI. We used the business survey results on costs and resources to inform our understanding, which will be supplemented by responses to questions in the Call for Views. We also considered each policy intervention based on if it would still be effective if there were significant changes to the Government's plans or funding arrangements.

Likely effectiveness and measurability
Effectiveness was viewed in terms of whether each intervention would ultimately reduce the threat to users, improve AI security and how it will drive improved security for an organisation that utilises AI. Equally, it also consisted of evaluating whether the proposal could be measured to assess its impact.

Barriers to Implementation
We assessed each intervention in terms of its likelihood of adoption and support from key stakeholders and its ease of implementation. As part of this, we considered the timescale of implementation and whether there were factors faced by different stakeholder groups that needed to be considered.

Unintended consequences
We considered whether our proposed interventions would introduce the right incentives we wish AI developers and organisations to adopt, without introducing negative consequences as a side effect.

Consistency with international approaches
Given the global reach of technology, we have focussed on aligning each intervention with international efforts. This aspect involved evaluating how an action could either foster international cooperation or potentially contradict or interfere with strategies adopted globally.

Equity and Impact (on consumers and businesses in the market)
We evaluated the impact of each intervention in terms of whether it disproportionately affected specific groups. This included examination of whether the intervention lessened competition and contributed to creating monopoly positions within the market. This should be viewed within the context of the government's support for digital innovation (see below).

Pro-innovation approach
We examined whether each policy option reduced the incentives for research and development within AI. We also examined whether the policy would prohibit or restrict products or services offered by AI companies and thus create further barriers of entry into the market. We are taking this pro-innovation approach as AI is already bringing extensive benefits to society, and we want to ensure future opportunities are not impeded by this work.

Interventions that were considered by Government

A number of options have been considered in support of the programme's objectives. The options set out below were considered against the criteria that informed the development of policy interventions.

Measure	Rationale

Business Guidance	The responsibility for securing AI does not fall solely on one stakeholder in the AI Supply Chain. Therefore, we would welcome feedback on whether additional guidance is needed to supplement the proposed voluntary of Code of Practice; for example, this could be guidance targeted at a particular stakeholders or a particular phase of the AI lifecycle. This should be considered within the context of the significant guidance that has already been developed to date by NCSC to support organisations.
Consulting on regulation of proposed security requirements	As set out in the AI white paper response, the Government has committed resources to supporting regulators and, noted that in the future, there will be a need for a highly targeted set of binding measures that apply to the most powerful AI systems. However, it is critical that the Government understands the evidence more fully before we advance regulation. Additionally, based on previous work, it is essential that the UK works continues to work with international partners to build international consensus for baseline security requirements in this area. We will keep regulation under consideration but believe our work to finalise the voluntary Code of Practice and support efforts in global standards bodies should be the priority.
Creating a certification scheme based on the security requirements for AI companies	It is important that any certification scheme is led by industry considering the various companies both in the UK and globally who provide an important service to help businesses assure products and services against specific requirements. Moreover, it is essential that international support is developed for baseline security requirements first so that any future certification scheme is based on principles that have consensus. We are therefore engaging with various countries as well as in standards development organisations and multilateral fora to promote this work and the Code’s requirements.
Create a guide for AI developers to complete and provide for customers	The Government considered whether it would be useful for organisations to be provided with specific information on what steps an AI company had taken to secure their product. Based on engagements with stakeholders, the Government didn’t progress forward with this guide because of the burden it would put on organisations, and it could have brought about potential liability concerns. Additionally, we found that it would not necessarily drive the adoption of better security practices in comparison to a Code of Practice. The Government remains committed to increasing transparency in this field and welcome thoughts from stakeholders on how this could be addressed in the future.

Guidance for AI developers	In November 2023, NCSC published their Secure AI Guidelines for AI developers which provided useful information to help inform the development of models and systems. Based on our evidence and stakeholder engagement, the Code will be the most suitable document to lead on from this because it sets out the specific actions that stakeholders across the AI supply chain need to implement to help protect users. We considered whether guidance targeted at a particular phase of the lifecycle may be useful, but based on the findings of DSIT's risk assessment, it was clear that there were risks in each phase that needed to be addressed. The Government welcomes feedback from AI developers if they believe further information on a particular area would be useful.
Guidance for consumers	While we advocate every user to take action to ensure their own digital safety and security, the Government is not progressing with this intervention because the burden for taking action to ensure a user is safe should not fall on a consumer in the first instance in the supply chain. The Code as well as the need for models and systems to be secure by design is essential if we want to ensure that consumers, businesses and the wider economy can continue to benefit from AI.
Awareness campaign targeted at organisations and users to increase understanding of security in context of AI	We recognise that there are many businesses in the UK who have not implemented AI within their infrastructure, and this may be because of a lack of understanding of what security requirements they should expect from AI developers as well as if they should have specific cyber security practices for AI models. However, our engagement and previous work has shown that the Code is a more effective lever to drive change because if we can ensure that the market is adopting the requirements then the burden on taking action will be significantly reduced on users. However, we would recommend that businesses take stock of DSIT's Cyber Governance Code, NCSC's Business Toolkit as well as NCSC's specific AI guidance for businesses to help inform their commercial decisions.

Annex F: Bibliography of relevant publications mapped to principles by Mindgard

[Amazon, 2023] Amazon, AWS Cloud Adoption Framework for Artificial Intelligence, Machine Learning, and Generative AI, Amazon White Paper, 2023.

[ASD, 2023] Australian Signals Directorate, An introduction to Artificial Intelligence, 2023.

[BSI1, 2023] Federal Office in Information Security, AI Security Concerns in a Nutshell, 2023.

[Cisco, 2022]Cisco, The Cisco Responsible AI Framework, 2022. Online: https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/docs/cisco-responsible-artificial-intelligence-framework.pdf

[Deloitte, 2023] Deloitte, Safeguarding Generative Artificial Intelligence with Cybersecurity Measures, 2023.

[ESLA, 2023] ESLA, European Lighthouse on Secure and Safe AI, 2023.

[ENISA, 2023] ENISA, Multilayer Framework for Good Cybersecurity Practices for AI, ENISA, 2023.

[Google, 2023] Google, Google Secure AI Approach Framework (SAIF), 2023. Online: https://services.google.com/fh/files/blogs/google_secure_ai_framework_approach.pdf

[G7, 2023] G7 Hiroshima Summit, Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, 2023.

[HHS, 2021] United States Department of Health and Human Services, Trustworthy AI (TAS) Playbook, 2021.

[ICO, 2020] Information Commissioner’s Office, Guidance on AI Auditing Frameworks, ICO, 2020.

[Microsoft, 2022] Microsoft, Artificial Intelligence and Machine Learning Security, Microsoft Learn, 2022.

[MITRE, 2024] MITRE, MITRE ATLAS: Mitigations. Accessed: January 10th 2024. Online: https://atlas.mitre.org/mitigations/

[NCSC, 2023] NCSC, Guidelines for Secure AI Systems Development, 2023. Online: https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

[NIST, 2022] NIST, AI Risk Management Framework: Second Draft, 2022. Online: https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf

[NIST, 2023] A. Vassilev, A. Oprea, A. Fordyce, H. Anderson, Adversarial Machine Learning – A Taxonomy and Terminology of Attacks and Mitigations, NIST, 2023

[OpenAI2, 2024] OpenAI, Introducing ChatGPT, Online: https://openai.com/blog/chatgpt, Accessed: January 9th 2024.

[OWASP, 2024] OWASP, OWASP AI Exchange. Accessed January 12th 2024. Online: https://owaspai.org/

[WEF, 2024] World Economic Forum, IBM, The Presidio AI Framework, 2024. Online: https://www3.weforum.org/docs/WEF_Presidio_AI%20Framework_2024.pdf