Department for
Science, Innovation
& Technology

# Implementing the UK’s AI Regulatory Principles

## Initial Guidance for Regulators

February 2024

## Contents

# Summary

This guidance sets out the considerations that regulators may wish to have when developing tools and guidance to implement the UK’s approach to AI regulation. It builds on our White Paper commitment to produce guidance for regulators to support the implementation of the UK’s five pro-innovation regulatory principles.

It is not intended to be a prescriptive guide on implementation as the principles are voluntary and how they are considered is ultimately at regulators’ discretion. Elements of this guidance may not be applicable to regulators who adopt a ‘technology agnostic’ approach to regulation as long as these regulators are satisfied that their regulatory framework adequately covers issues relating to AI adoption. This is the first phase of guidance – we have been working actively with regulators to develop this and will continue to do so as we refine and further develop it.

# Background

In March 2023, the UK Government published its AI Regulation White Paper setting out its pro-innovation approach to AI regulation. The approach sets out five principles for regulators to interpret and apply within their remit. These are:

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

The White Paper emphasised that this framework is intended to support and supplement the work of our expert, independent regulators. This guidance is primarily aimed at regulators whose remits are most immediately impacted by the use of AI, and who may just be beginning to consider the impacts of AI. It is also intended to be relevant to any regulator whose remit could be impacted by AI in the future.[1] We note that different regulators are at different stages of their consideration of AI risks.

[1] We recognise that there are a wide range of organisations who perform regulatory functions within the UK, with different structures, systems of accountability and degrees of independence. As such we have not attempted to provide a standardised definition of ‘regulator’ in this guidance, as organisations are being asked to consider this guidance on a voluntary (non-statutory) basis. In keeping with our regulator-led approach to implementing the principles, we instead encourage all bodies with any kind of regulatory responsibility for AI to consider how this guidance might support them to apply the principles – noting that guidance may be more relevant to some regulators than others e.g. where regulators take a technology agnostic approach to regulation.

To ensure a coherent and streamlined AI regulatory landscape, DSIT has started establishing a central function. The central function supports UK regulators’ understanding of the AI risk landscape and will support them to conduct risk assessments by providing expert risk analysis, which is already underway within DSIT. This allows us to monitor risks holistically and identify any potential gaps in our approach that leave risk not adequately mitigated.

The central function also catalyses the development of regulators’ skills and expertise in AI. In its White Paper consultation response, DSIT announced a £10 million package to boost regulators’ AI capabilities. The central function will work closely with regulators in the coming months to identify the most promising opportunities to leverage this funding and will continue to support regulators to future-proof their AI capabilities, as AI technologies and the broader context in which they are used continue to change.

Further key roles of the central function include supporting increased coherence across regulators, promoting information sharing and working with regulators to analyse and review potential gaps in existing regulatory powers and remits. However, we also note that some of these principles may not be relevant for specific regulators. We are working closely with regulators to further develop the central coordination function and will set its role out further in phase two of guidance (see below).

To enable greater regulatory coherence, we committed to publishing guidance for regulators to support the implementation of the five regulatory principles. This guidance is not intended to duplicate, replace or contradict regulators’ initiatives or existing statutory definitions of principles – they know their remits best. It remains the responsibility of regulators to develop their own tools and guidance as they deem appropriate to support AI developers, deployers and end users within their remit. We also note that some of these principles may not be relevant for specific regulators.

Instead, this guidance supports the work that many regulators are already undertaking to produce their own remit-specific tools and guidance. This could include issuing outcomes-based guidance that is not specific to AI, yet sufficiently covers relevant AI risks. We understand that there is not a one size fits all approach to AI regulation and to considering the principles - this guidance intends to promote and enable greater coherence in AI outcomes across different regulatory remits where possible.

**We are taking a phased approach to issuing this guidance. We will issue guidance in the following three stages:**

- **Phase one:** this initial guidance supports regulators by enabling them to properly consider the principles and to start considering developing tools and guidance for their regulatory remits if they have not done so already. It sets out considerations for regulators as they develop regulatory activities to support AI developers and deployers within their regulatory remit to comply with their obligations, follow good practice and mitigate risks. **This document is the phase one guidance.**

- **Phase two:** will iterate and expand on this initial phase one guidance to provide further detail, informed by feedback from regulators and other stakeholders. We plan to issue this by summer 2024. We are working with regulators to understand and identify which (additional) resources or points of clarification are needed. In addition, phase two will set out further information on what mechanisms and resources the central function offers.
- **Phase three:** will involve collaborative working with regulators to identify areas for potential joint tools and guidance across regulatory remits. Through this process, we will aim to work with regulators to identify where additional information, resources are needed, and where appropriate, collaborate on joint solutions – for example, encouraging multi-regulator guidance.

Throughout this document, we highlight important considerations for regulators when developing 'tools and guidance' to support the implementation of the principles in their regulatory remit. 'Tools and guidance' refer to any product[i] that regulators could issue to support AI developers, deployers and end users and to promote greater coherence on AI risk management across regulatory remits. It is inclusive of all products aimed at promoting understanding in different contexts, including non-AI specific products.

To promote innovation and competition, regulators are encouraged to develop tools and guidance that promote knowledge and understanding as relevant in the context of their remit, rather than setting out step-by-step processes. It is recognised that the varying status and remits of regulators means that this may not always be relevant or appropriate. Some regulators adopt a technology-agnostic approach to regulation as long as they are satisfied this framework adequately covers issues relating to AI adoption. Where this is not the case, regulators may either update their existing framework or issue AI-specific guidance as necessary to address gaps.

Regulators can establish published policy material, in respect of AI, that is consistent with their respective regulatory objectives, setting out clearly and concisely the outcomes regulators expect, so that regulated firms can meet these expectations through their actions. A broad range of products – including non-AI specific products that sufficiently address AI-related risks - can help improve awareness in regulatory remits (i.e., policy publications, leaflets, webpages, public awareness campaigns) and the choice of product should consider the nature and the severity of AI risk in that use case or context and the audience(s) they are targeting. This document also refers to 'joint tools and guidance'- again this is inclusive of all products that could be produced across two or more regulators.

## Overview of voluntary guidance for implementing the regulatory principles

This tables summarises the key considerations that regulators could have when issuing tools and guidance in their remit. These considerations are set out in more detail in their respective sections within the main body of this guidance document.

| | |
|---|---|
| Section 1: Guidance on interpreting and applying the AI regulatory framework | • Promoting transparency by putting information in the public domain on the actions regulators are taking to assess and manage AI risks and opportunities within a regulators’ sector(s)<br>• Consider relevant guidance published by other regulators, the benefits of issuing joint tools or guidance and potential forums to facilitate collaboration<br>• Note that different principles may have more obvious relevance in certain regulatory remits, but regulators are encouraged to give consideration to all the principles as a first step<br>• Use technical standards to support AI developers and deployers to implement the principles |
| Section 2: Applying individual principles | **Safety, security, robustness:**<br>• Understand and communicate the level of safety related risk in their regulatory remit<br>• Stress the importance of AI developers and deployers (within regulators’ remits) undertaking safety risk assessments and implementing appropriate mitigations to identified risks<br>• Consider how AI developers and deployers should mitigate and build resilience to cybersecurity related risks throughout the AI life cycle<br>**Appropriate transparency and explainability:**<br>• Explain that appropriate levels of transparency and explainability help to foster trust in AI and increase AI use<br>• Encourage AI developers and deployers to implement appropriate transparency and explainability measures<br>• Understand that this principle is necessary for the proper implementation of the other four principles<br>**Fairness:** |

| | |
|---|---|
| | • Continue to develop, publish descriptions or signpost to existing descriptions of fairness that apply to AI systems' outcomes within their regulatory remit<br>• Consider how AI systems that are used in their regulatory remit are designed, developed, deployed and used considering this description of fairness<br>• Note that aligning descriptions of fairness and developing joint tools and guidance is particularly important in cross-cutting regulatory remits<br>**Accountability and governance:**<br>• Place clear expectations for compliance and good practice on appropriate actors in the AI supply chain (within regulators’ remits), including expectations for what appropriate internal accountability and governance frameworks might look like<br>• Consider whether existing powers that place accountability on decision makers are applicable in the context of AI and to AI developers and AI deployers<br>• Seek to foster accountability through promoting appropriate transparency and explainability<br>**Contestability and redress:**<br>• Where appropriate, encourage AI developers and AI deployers (within regulators’ remits) to provide clarity to users on which routes they can use to contest AI outcomes or decisions<br>• Highlight that appropriate transparency and explainability is key to ensuring that AI deployers or end users can contest outcomes and are aware of routes to redress |
| Section 3: how to communicate progress on engagement with AI principles | Regulators are encouraged to ensure that AI developers, AI deployers and end users understand how the principles are being implemented in context to boost clarity and trust in the use of AI, for example by publishing AI strategies. |

# Key terms

**AI agents:** Autonomous AI systems that perform multiple sequential steps – sometimes including actions like browsing the internet, sending emails, or sending instructions to physical equipment – to try and complete a high-level task or goal.

**AI or AI system:** Products and services that are 'adaptable' and 'autonomous' in the sense outlined in our definition in section 3.2.1 of the AI White Paper.  This definition is inclusive of AI agents, frontier AI, and narrow AI.

**AI deployers:** Any individual or organisation that supplies or uses an AI application to provide a product or service to an end user.

**AI developers:** Organisations or individuals who design, build, train, adapt or combine AI models and applications.

**AI end user:** Any intended or actual individual or organisation that uses or consumes an AI-based product or service as it is deployed.

**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.

**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.

**Narrow AI:** An AI system that performs well on a single task or narrow set of tasks, like sentiment analysis or playing chess.

*We note that definitions of AI are often challenging due to the quick advancements in the technology. These definitions of key terms in this guidance, and wider relevant definitions to AI, are set out in more detail in our consultation response and our discussion paper on frontier AI capabilities and risks.*

*Through tools and guidance, regulators could encourage AI developers and AI deployers within their remits to implement the five regulatory principles. However, where legal responsibility can be assigned is dependent on the law in a specific case. A regulators' role in holding AI developers and AI deployers to account legally is also contingent on whether they are a regulated entity within that regulators' remit. This is explained further in section two, accountability and governance, and we are undertaking further internal policy work to clarify this area and intend to provide more information in future phases of guidance.*

# Guidance on interpreting and applying the AI regulatory framework

As laid out in our White Paper, we have created a principles-based AI regulatory framework that outlines the key outcomes AI systems should be aligned with regardless of the context in which that system is deployed. The framework provides regulators with the flexibility to interpret and apply these principles to the outcomes of AI use cases that fall within their remit.

To deliver on the AI regulatory framework regulators could consider:

- **Promoting appropriate transparency by putting information in the public domain on the action's regulators are taking in relation to AI.** For example, this could include developing, publishing and maintaining AI plans – whether standalone or as part of wider corporate strategies. Considering the pace at which AI is transforming our economy and impacting regulatory remits, regulators are encouraged to keep such information up to date. This is set out in more detail in section three of this guidance.

**Examples of regulator collaboration:**

- DRCF members (Ofcom, ICO, FCA and CMA) are continuing to build consensus on the application of AI and are examining the emerging risks and opportunities it creates in their remits, particularly generative AI technologies.
- Established by the HRA, MHRA, NICE and CQC, the AI and Digital Regulations Service provides a single source of advice and guidance on health sector regulatory requirements for both AI developers and deployers. Since its launch, the service's developer guidance has recorded over 2800 visits and the service's adopter guidance has recorded over 3,800 visits.
- The FCA, the PRA and the Bank of England have collaborated on and published a Discussion Paper to deepen dialogue on the impact that artificial intelligence and machine learning might have on the supervision of financial firms. The paper acknowledges the potential that these technologies have to make financial services and markets more efficient, accessible and tailored. However, it also notes the regulatory risks the technologies could pose to markets and the need for the three organisations to collaborate to mitigate these risks.

- **Examining opportunities for collaboration and knowledge exchange through existing mechanisms, as well as new ones.** Regulators are encouraged to proactively collaborate with each other to regulate AI where possible. This could mean using existing forums such as the Digital Regulation Cooperation Forum (DRCF) or establishing new cooperation forums. Such forums improve coherence in the way laws are interpreted in relevant areas and can provide soft channels for communication, enabling more agile information sharing. This could be particularly useful for regulators not in the DRCF. In other instances, other mechanisms such as a memorandum of understanding may be appropriate. When fully established, the new central function will also support knowledge exchange.

- **Relevant guidance published by other regulators**. When issuing tools and guidance on how principles interact with existing legislation, regulators could consider relevant guidance produced by other regulators through:
    - Assessing what other regulatory guidance exists.
    - Consulting with regulators that have produced this guidance where appropriate to align approaches. This includes seeking coherence on the definition of each principle when considering it in cross-cutting cases or setting-out differences in definitions and the rationale for these differences.
    - Where appropriate, taking an iterative approach to revising guidance supported by the work of other regulators.
- **The benefits of issuing joint tools or guidance where cross-cutting risks or issues are identified.** Joint tools and guidance are particularly important regarding areas in which regulatory remits overlap. Regulators with horizontal remits are therefore most likely to need to issue joint tools and guidance. Regulators are encouraged to communicate their progress publicly to assist others in implementing the framework.
- **That principles do not supersede existing legislation and that existing regulatory frameworks may already be managing similar risks to those that stem from AI but are not unique to it.** The principles are intended to supplement existing work and to support a coherent approach to AI regulation across regulators.
- **Noting that different principles may have more obvious relevance in certain regulatory remits, but all principles should be considered where feasible.** Given that the principles are cross-cutting and outcome-focused, some may feel less applicable than others based on the regulatory remit. However, regulators could consider all principles where feasible and seek to make them clear in tools and guidance related to AI. Doing so could support the identification of gaps in regulating AI that will be used to develop the UK's regulatory framework (more detail will be provided on this process in subsequent phases of guidance).
- **Maintaining an understanding and overview of how the principles are being interpreted and acted on by organisations that fall within their regulatory remit - noting that not all principles will be relevant in all cases.** Information and insights from regulators on this will help with the identification of regulatory/capability gaps, to help inform policy on how these can be addressed.
- Map which technical standards could help AI developers and deployers understand the principles in practice and cite these standards in tools and guidance. This is not to say that regulators should prescribe the use of specific standards to AI developers and

deployers, but that standards may be a helpful tool to illustrate how AI developers and deployers could comply and give them a tool to support their understanding.

**Referencing technical standards to support AI developers and AI deployers**

- In addition to technical standards specific to a regulatory remit, in tools and guidance a regulator may wish to cite horizontal AI standards produced by organisations like BSI, ISO and IEC. In section two of this guidance, we reference examples of horizontal standards that could help to illustrate the application of a specific pro-innovation principle. However, to support AI developers and deployers alignment with all five principles, regulators may find it helpful to reference these standards in tools and guidance:
- **ISO/IEC 42001 – AI Management System Standard** – this standard guides organisations to continuously improve and iterate responsible processes customised for AI systems.
- **ISO/IEC 22989 – Artificial intelligence concepts and terminology** – this standard establishes terminology for AI and describes key concepts in the field of AI. The standard is applicable to all organisations that use AI.

Please note that this guidance is not an endorsement of any specific standard. It is for regulators to consider standards and their suitability in a given situation (and/or encourage those they regulate to do so likewise).

# Applying individual principles

Regulators should also examine how to support the realisation of the principles individually where possible. This section sets out these considerations for each principle, notes potential key questions that regulators may want to consider for each principle and suggests technical standards and existing best practice to review. Whilst the principles are presented individually, it is also important to think about their interdependence and how they support each other. Where references are made to aligning of definitions of principles, it is important to note that this suggestion is not meant to supersede existing regulatory definitions that are already described in statue.

## Safety, security & robustness

**AI systems should function in a robust, secure and safe way throughout the AI life cycle, and risks should be continually identified, assessed and managed.**

**When considering this principle in tools and guidance regulators could:**

- **Communicate the level of safety related risk** in their remit by appropriately identifying, monitoring, communicating and acting upon risks. **This will require defining what safety, security and robustness mean** in the context of AI systems in a particular regulatory remit. Having clear and consistent definitions is particularly important for regulators with cross-cutting remits to ensure regulatory coherence.
- **Provide tools and guidance for undertaking AI-related safety risk assessments and implementing appropriate mitigations.** For example, by providing examples of risk identification processes, risks to consider and mitigations that can be implemented. Tools and guidance could stipulate that risks need to be assessed regularly and mitigating actions updated accordingly.
- Enable **AI deployers (within their remit) and end users to make informed decisions about the safety of AI products and services.** Where appropriate, regulators could think about the role of AI developers in enabling AI deployers and end users to make informed safety assessments, and additionally AI deployers in enabling end users to make informed safety assessments. Regulators could also consider issuing tools and guidance aimed directly at end users to increase their ability to make informed decisions. This links closely to the implementation of the transparency and explainability principle.
- Consider **how the associated actors on the AI supply chain can regularly test or carry out due diligence** on the functioning, resilience and security of a system. Regulators could also encourage actors to share this information throughout the supply chain for others to use in their procurement, purchase or developer decisions.

- **Encourage AI developers and deployers (within their remit) to mitigate and build resilience to cybersecurity related risks throughout the AI life cycle.** For example, this could include the National Cyber Security Centre principles for secure machine learning models[2] when providing tools and guidance on this.
- **Encourage AI developers and deployers to consider and mitigate where possible potential malicious or criminal use of AI products and services**. For example, this could be the use of generative AI to create illegal content, undertake fraud, or transmit false communications under the Online Safety Act.

**Key questions to consider when developing tools and guidance:**

- How do you describe AI safety, security and robustness in your remit?
- What is the level of safety related risk from AI in your regulatory remit?
- How will you encourage AI developers and deployers to regularly test or carry out due diligence and risk assessments on their systems?
- How will you encourage AI developers and deployers to assess security threats and build security resilience into their systems?

**Technical standards to consider:**

- ISO/IEC 23894:2023 – Information Technology - Artificial Intelligence - Guidance on Risk Management
    - This standard provides guidance on managing risks related to AI. It assists organisations to integrate risk management to their functions and activities.
- ISO/IEC CD TS 8200 - Controllability of automated artificial intelligence systems
    - Irrespective of a system's safety and level of autonomy, it is vital to be able to intervene in its operations and potentially interrupt them to avoid negative consequences. This standard is currently under development and will provide principles and approaches for realising and enhancing this controllability to ultimately ensure safety.
- ISO/IEC TR 5469:2024 - Artificial Intelligence - Functional safety and AI systems
    - This standard covers matters related to both AI systems and the use of AI in safety related systems. It provides guidance on the use of AI to fulfil safety functions. It also details safety risks that AI systems could pose and suggestions on mitigation solutions.
- ISO/IEC TR 29119-11:2020 - Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems
    - This standard introduces important concepts and means to test models. It identifies challenges in testing AI systems and proposes mitigations to those

[2] https://www.ncsc.gov.uk/collection/machine-learning

challenges.

**Existing examples of guidance or best practice that focus on this principle:**

- ICO - Guidance on AI and data protection
- MHRA - Software and Artificial Intelligence (AI) as a Medical Device

## Appropriate transparency and explainability

**AI systems should be appropriately transparent and explainable.**

**When considering this principle in tools and guidance regulators could:**

- **Emphasise that transparency and explainability help to foster trust in AI and can increase appropriate innovation and adoption**. Insufficient transparency and explainability also increases the risk of inadvertently breaking laws, infringing rights, or causing harms, ultimately compromising the use and uptake of AI systems. However, the tools and guidance could note that the degree of transparency could also be responsive to risk(s) identified through risk assessments. In certain cases, high levels of, or certain formats of transparency could increase security risks.
- **Encourage AI developers and deployers (within their remit) to implement appropriate transparency and explainability measures.** This could include encouraging developers and deployers to notify end users when they are affected by or engaging with an AI system, explaining as simply as possible the purpose of AI systems, how the AI system makes decisions and how outputs may be used. End users should also be able to access sufficient information about those systems to be able to enforce their rights and to understand what would constitute legal and illegal uses of their product.
- **AI developers within regulators' remits could also be encouraged to provide appropriate transparency and explainability measures to AI deployers about the system they are using to deliver a product or service.** To enable this, AI developers could provide clear information on how their AI system works and suggest how AI deployers could explain this to end users.
- Consider **asking or requiring (under existing powers) AI developers and deployers within regulators' remits to provide information to show how they are adhering to this principle** in their organisational processes, such as through product labelling and identifiers of AI generated content,
- Note that **this principle is necessary for the proper implementation of the other four principles**. It enables AI deployers and end users to make informed decisions about the systems that they interact with and the outcomes that those systems reach.

**Key questions to consider when developing tools and guidance:**

- What is an appropriate level of transparency and explainability for different AI systems in your regulatory remit given potential security risks identified? Could certain forms of transparency exacerbate security risks?
- How will you encourage AI developers and deployers within your remit to adopt appropriate levels of transparency and explainability?
- How will you encourage AI developers within your remit to appropriately explain their products to AI deployers and end users within your remit? Additionally, how will you encourage AI deployers to explain their use of AI systems to end users?
- How will you communicate the importance of this principle in implementing the other four regulatory principles?

**Technical standards to consider:**

- ISO/IEC TS 6254 - Information technology — Artificial intelligence — Objectives and approaches for explainability of ML models and AI systems
  - This standard is currently under development and will describe existing methods and approaches for improving the explainability of AI systems. It will review the different forms that explanation can take dependent on the target audience.
- IEEE 7001 - Standard for Transparency of Autonomous Systems
  - This standard provides examples of appropriate levels of transparency for different stakeholder groups. It distinguishes between transparency and explainability to support their implementation.
- ISO/IEC CD 12792 – Information Technology – Artificial Intelligence – Transparency taxonomy of AI systems
  - This standard is currently under development and will provide horizontal guidance to define a taxonomy of information elements to identify and address transparency in AI systems.

**Existing examples of guidance or best practice that focus on this principle:**

- CMA - AI Foundation Models: Initial report
- CDDO & CDEI – Algorithmic transparency recording standard
- ICO and The Alan Turing Institute – Explaining decisions made with AI

# Fairness

**AI systems should not undermine the legal rights of individuals or organisations, discriminate unfairly against individuals or create unfair market outcomes. Actors involved in all stages of the AI life cycle should consider descriptions of fairness that are appropriate to a system’s use, outcomes and the application of relevant law.**

**When considering this principle in tools and guidance regulators could:**

- **Provide descriptions of fairness that can be applied to outcomes of AI systems used within the sector(s) they regulate**. This will require regulators to describe what a fair outcome from AI is and where appropriate in cross-cutting remits shared descriptions could be established with other regulators. There is not one description of fairness and regulators are encouraged to consider fairness in context specific outcomes, consulting end users and ethics researchers where appropriate. Fairness includes concepts such as negative bias mitigation and the need to treat people fairly but will also cover wider considerations such as issues around procedural fairness. The development of joint tools and guidance may be particularly important in cross-cutting remits given the need for alignment on context specific descriptions of fairness.
- **Tools and guidance could also consider relevant law, regulation, technical standards and assurance techniques.** These should be applied and interpreted similarly by different regulators where possible. For example, regulators need to consider their responsibilities under the 2010 Equality Act and the 1998 Human Rights Act. Regulators may also need to understand how AI might exacerbate vulnerabilities or create new ones and provide tools and guidance accordingly.
- **Consider how AI systems in their remit are designed, developed, deployed and used considering such descriptions of fairness** and issue tools and guidance to promote this. These could support AI developers within regulators’ remit to assess and mitigate the potential impact of negative bias in systems they create, and AI deployers within their remit to assess and mitigate negative impact caused by their use of AI systems. This may require more robust inspection than when decisions are made by humans or non-AI software - one AI system could make many decisions, exacerbating the potential for negative impacts stemming from biases held by an AI system.

**Key questions to consider when developing tools and guidance:**

- How would you describe a fair outcome of AI use in your regulatory remit? How can you clearly communicate this description?
- What evidence or information is required to assess whether AI systems are being used fairly in your regulatory remit?
- How can you support regulated entities to assess and mitigate bias?

- How might evaluations of fairness change in the context of AI technologies compared to when decisions are made by humans, or non-AI software?

**Technical standards to consider:**

- IEEE P7003 – Algorithmic Bias Considerations
    - This standard describes processes and methodologies to help users address issues of bias in the creation of algorithms.
- ISO/IEC TR 24027:2021 – Information technology — Artificial intelligence (AI) — Bias in AI systems
    - This report offers a comprehensive overview of bias and fairness issues in AI systems, analyses the sources of bias and describes potential mitigation techniques to treat unwanted bias.
- ISO/IEC 17866 PWI Artificial intelligence — Best practice guidance for mitigating ethical and societal concerns
    - This standard is still being developed but will provide guidance on identifying and treating ethical and societal issues throughout an AI lifecycle.
- IEEE 7000 – IEEE Standard Model Process for Addressing Ethical Concerns During System Design
    - This standard aims to help innovators include ethical considerations throughout different stages of a systems' design and development. The standard can be used by AI developers of any size to ensure they're building systems aligned with certain ethical values.

**Existing examples of guidance or best practice that focus on this principle:**

- EHRC – EHRC guidance on use of AI by public bodies
- Best practice initiatives: In October 2023, the CDEI and Innovate UK launched the 'Fairness Innovation Challenge', a grant challenge to drive the development of new sociotechnical solutions to address bias and discrimination in AI systems. The challenge is being delivered in partnership with UK regulators, the EHRC and the ICO, who will engage with winners to ensure that their solutions are compliant with relevant legal frameworks, including data protection and equalities law.

## Accountability and governance

**Governance measures could be put in place to ensure effective oversight of the supply and use of AI systems, with clear lines of accountability established across the AI life cycle.**

**When considering this principle in tools and guidance regulators could:**

- Consider whether regulators’ **regulatory powers or remits allow them to place legal responsibility** on actors in the supply chain that are best placed to mitigate the risks. Issue tools and guidance to AI developers and deployers to communicate clearly how these laws apply in the context of AI and whom those laws can hold to account.
- **Where legal responsibility cannot be assigned to an actor in the supply chain that operates in a regulatory remit, encourage the AI actors within the remit to ensure good governance in who they outsource to.** For some regulators with vertical remits, it is acknowledged that it may only be possible to assign legal responsibility on AI deployers whilst fostering accountability to AI developers through non-statutory measures. We are undertaking further internal policy work to clarify this area and will provide more information on it in future phases of guidance.
- Be clear that **‘accountability’ refers to the expectation that AI developers could adopt appropriate governance measures to ensure the proper functioning, throughout the life cycle,** of the AI systems that they research, design, develop, train, operate, deploy or otherwise use and decommission.
- **Place clear expectations for compliance, good practice and internal governance structures on AI developers and deployers within regulators’ remits.** There is a wealth of existing precedents, standards and best-practices across regulated sectors to learn from. Examples include encouraging the use of good governance practices to ensure that expectations like robust risk management are met.
- **Clarify the responsibilities of AI developers and deployers within regulators’ remits to demonstrate proper accountability and governance.** This will help to create legal certainty whilst ensuring regulatory compliance and could include activities such as sharing documentation on key decisions and allowing audits where appropriate.
- **Foster accountability through promoting appropriate transparency and explainability**. By making clear when AI systems are being used and explaining how those systems are being used, end users can be empowered to hold those systems to account.

**Key questions to consider when developing tools and guidance:**

- How could you clarify expectations for regulatory compliance and good practice in the use of AI in your regulatory remit?
- Where does responsibility for outcomes caused by AI systems sit in your remit and do you have any existing powers to legally assign this, or are there other non-statutory mechanisms you can use to assign accountability?
- What good governance practices could you promote in your regulatory remit to ensure accountability and how will you promote actors to demonstrate that they are following these practices*?*

**Technical standards to consider:**

- ISO/IEC 38507:2002 – Governance of AI

  - This standard provides guidance for members of a governing body of an organisation to enable the effective, efficient and acceptable use of AI within their organisation.
- ISO/IEC 25059 – Quality model for AI systems
  - This standard outlines a quality model for AI systems and details a consistent terminology for specifying, measuring, and evaluating AI system quality.
- ISO/IEC DIS 42006 – Requirements for bodies providing audit and certification of artificial intelligence management systems
  - This standard is currently under development and will detail the requirements needed to be an accredited certification body to reliably audit the management system for organisations that develop or use AI systems according to ISO/IEC 42001 (AI Management System).
- ISO/IEC CD 42005 - AI system impact assessment
  - This standard is currently under development and will provide guidance for AI developers and deployers to undertake impact assessments on how end users and society can be impacted by AI systems.

**Existing examples of guidance or best practice that focus on this principle:**

- ASA - Generative AI & Advertising: Decoding AI Regulation[3]
- CQC - Using machine learning in diagnostic service: A report with recommendations from CQC's regulatory sandbox
- OPSS - Study on the impact of artificial intelligence on product safety

## Contestability and redress

**Where appropriate, users, impacted third parties and actors in the AI life cycle should be able to contest an AI decision or outcome that is harmful or creates**

**When considering this principle in tools and guidance regulators could:**

- **Ensure that AI developers and deployers within their remit are consistent with their statutory objectives, to provide clarity to users on which existing routes to contestability apply.** This will encourage them to implement proportionate measures so that AI-informed decisions can be contested where appropriate. This is ultimately the responsibility of AI developers and deployers, but regulators could promote this practice. This will enable AI deployers to contest outcomes with AI developers, and will enable AI users to contest outcomes with AI deployers, AI developers or both.

[3] The ASA is the self-regulatory body for the UK's advertising industry.

- Highlight that appropriate **transparency and explainability are relevant to the good implementation of this principle**. Transparency is key in making routes to redress clear.

**Key questions to consider when developing tools and guidance:**

- What existing routes do end users, AI deployers, or anyone impacted by AI use, have to contest outcomes? Are these routes to contestability appropriate in the context of AI? How will you ensure that these routes are clearly communicated and used in the context of AI?
- Do impacted parties have the information needed to contest outcomes with AI deployers and AI developers using these routes, e.g. information that a decision was informed by an AI system? How could you communicate the importance of transparency and explainability in enabling contestability.

# How to communicate progress on engagement with AI principles

This guidance provides suggestions on how regulators could implement the five principles outlined above in the way that works best in their specific regulatory remits. This could include technology-agnostic approaches to regulation as long as these regulators are satisfied that this framework adequately covers issues relating to AI adoption. We want to ensure that AI developers, deployers and end users within regulators' remits understand how regulators may implement these principles and how the UK is addressing the regulatory challenges posed by AI. This is key to ensuring clarity and confidence to drive innovation and boost AI usage in the UK.

Regulators are encouraged to publish an update, outlining their strategic approach to AI and the steps they are taking in line with the principles. Regulators are best placed to determine the form and substance of this, but the update could include:

- Their current assessment of how AI applies within the scope of their regulatory responsibilities including an explanation of their enabling legislation and its relevance in the context of AI.
- The steps they are already taking to adopt the AI principles set out in the white paper – where possible this should include concrete examples of the actions they have taken over the preceding 12 months. We are aware that some principles will have limited applicability for certain regulators.
- A summary of guidance they have issued or plan to issue on how the principles interact with existing legislation and the steps industry should take in line with the principles.
- The work they are doing to understand, assess and manage the current and emerging risks posed by AI as relevant to their sector and remit. This could range from social harms such as bias and discrimination, to broader harms such as cyber security, privacy risks and potential for AI misuse from bad actors (to be informed in due course by the government's central AI risk assessment).
- Consider interactions and overlap between their areas of responsibility and that of other regulators. They could also cover any assessments on AI risks and opportunities that they have made and the regulatory, supportive and enforcement approaches they will seek to tackle them.
- The steps they have taken to collaborate with other regulators to identify and tackle AI-related issues that cut across regulatory remits.
- An explanation of their current capability to address AI risks within their regulatory remit - and how this compares with their assessment of the capabilities they need. This should set out th structures and resources they currently have in place including an assessment – e.g. quantified if possible – a) the number of people working partly or fully

on AI-related issues, b) the budget they have allocated to AI-related issues, c) specific skills and expertise they require in order to effectively regulate AI within their sector.

- A forward look of their plans and activities over the coming 12 months, this should include the actions they are taking to address any capability gaps identified above and could also include – but need not be limited to – risk assessment work they plan to undertake, tools and/or guidance they are preparing, planned stakeholder engagement activity, and international engagement. It would be useful to understand how they may prioritise their organisation’s resources to support the work within this forward look.

This activity provides regulators with an additional opportunity to communicate to key stakeholders a robust and detailed understanding of how they assess AI within their domain, what they are currently doing and what they propose to do in the future to manage AI risk. Detailing what enabling legislation exists within their sector will also help communicate how the UK’s regulatory approach is being implemented in specific regulatory remits, as well as make clear where potential gaps are that might inform future regulatory reform.

## Next steps

We plan to publish phase two guidance by summer 2024 and are working with regulators and wider key stakeholders on this. The aim is to ensure an improved flow of information between government and regulators, as well as supporting knowledge sharing across regulators themselves.

The central function is better equipping government to identify gaps within and across regulatory remits and will support regulators to identify solutions to these risks. It will also ensure that regulators are able to draw on the government’s central risk analysis when assessing risks and opportunities posed by AI within their own remits and build a common understanding of cross-cutting issues.

# Annex 1: Existing guidance, updates and other information sources purely focused on AI published by UK regulators

We recognise that further tools and guidance have been published by UK regulators that reference AI, or are applicable to AI, but are not solely focused on it.

| Regulator | Guidance and other information sources | Description | WP principle(s) this primarily relates to |
|---|---|---|---|
| Advertising Standards Authority[4] | Generative AI & Advertising: Decoding AI Regulation | High-level summary making clear that existing regulation on how ads are made ultimately applies to ads made using AI. | Accountability and governance |
| CMA | AI Foundation Models: Initial report | The guidance details how AI developers, deployers and end users need to be informed of the limitations of FMs to enable healthy competition. | Appropriate transparency and explainability |
| CQC | Using machine learning in diagnostic service: A report with recommendations from CQC's regulatory sandbox | Summary of findings from sandbox undertaken by the CQC on machine learning in diagnostic services. Notes responsibilities of developers and deployers in engaging with CQC. | Accountability and governance |
| EHRC | EHRC guidance on | Provides guidance to public bodies on how they should ensure they're | Fairness |

[4] The ASA is the self-regulatory body for the UK's advertising industry.

| | | | |
|---|---|---|---|
| | use of AI by public bodies | meeting PSED obligations when using AI. | |
| ICO | Guidance on AI and data protection | Best practice for data protection-compliant AI, as well as how to interpret data protection law as it applies to AI systems that process personal data. This includes how AI systems can exacerbate known security risks and make them more difficult to manage. It also presents the challenges for compliance with the data minimisation principle. Several techniques are presented to help both data minimisation and effective AI development and deployment. | Safety, security, robustness<br>Appropriate transparency and explainability<br>Fairness<br>Accountability and governance |
| ICO (and Turing) | Explaining decisions made with AI | Co-badged guidance by the ICO and The Alan Turing Institute aims to give organisations practical advice to help explain the processes, services and decisions delivered or assisted by AI, to the individuals affected by them. | Appropriate transparency and explainability<br>Fairness |
| MHRA | Software and Artificial Intelligence (AI) as a Medical Device | Guidance on how AI and software assisted medical devices interacts with existing legislation to protect patient safety and obligations on business to ensure their software/ AI used in the medical space complies with existing regulation. | Safety, security, robustness<br>Accountability and governance |
| Ofcom | Synthetic media (including deepfakes) in broadcast programming | Guidance for broadcasters on managing the risks related to synthetic media content that's often generated by AI. | Fairness<br>Accountability and governance |

| | | | |
|---|---|---|---|
| OPSS | Study on the impact of artificial intelligence on product safety | Considers where liability for consumers safety should sit in the context of products that are developed by AI. | Accountability and governance |

[i] A ‘product’ refers to any external facing document or communication that could assist in the implementation of AI principles in a regulatory remit.

This publication is available from: www.gov.uk/dsit

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