

What's new?	2
About this guidance	5
What are the accountability and governance implications of AI?	12
How do we ensure transparency in AI?	28
How do we ensure lawfulness in AI?	29
What do we need to know about accuracy and statistical accuracy?	38
How do we ensure fairness in AI?	44
What about fairness, bias and discrimination?	53
What is the impact of Article 22 of the UK GDPR on fairness?	69
How should we assess security and data minimisation in AI?	72
How do we ensure individual rights in our AI systems?	89
Annex A: Fairness in the AI lifecycle	105
Glossary	135

What's new?

This guidance was updated on 15 March 2023.

The Guidance on AI and Data Protection has been updated after requests from UK industry to clarify requirements for fairness in AI. It also delivers on a key [ICO25 commitment](#), which is to help organisations adopt new technologies while protecting people and vulnerable groups.

This update [supports the UK government's vision](#) of a pro-innovation approach to AI regulation and more specifically its intention to [embed considerations of fairness into AI](#).

We continue to engage with the UK government, along with our partners within the [Digital Regulation Cooperation Forum \(DRCF\)](#), on its broader proposals on regulatory reform.

The ICO supports the government's mission to ensure that the UK's regulatory regime keeps pace with and responds to new challenges and opportunities presented by AI. We look forward to supporting the implementation of its forthcoming White Paper on AI Regulation.

We will continue to ensure ICO's AI guidance is user friendly, reduces the burden of compliance for organisations and reflects upcoming changes in relation to AI regulation and data protection.

For ease of use and given the foundational nature of data protection principles we decided to restructure the guidance moving some of the existing content into new chapters. Acknowledging the fast pace of technological development the ICO believes more updates will be required in the future so using data protection's principles as the core of this expanding work makes editorial and operational sense.

We outlined below where new content resides so past readers of the Guidance on AI and Data Protection can navigate the changes at speed.

What are the accountability and governance implications of AI?

Change overview: This is an old chapter with new additions

What you need to know:

- [New content on things to consider as part of your DPIA](#)

How do we ensure transparency in AI?

Change overview: This is a new chapter with new content

What you need to know:

- We have created a standalone chapter with new high-level content on the transparency principle as it applies to AI. The main guidance on transparency and explainability resides within our existing [Explaining Decisions Made with AI](#) product.

How do we ensure lawfulness in AI?

Change overview: This is a new chapter with old content - moved from the previous chapter titled 'What do we need to do to ensure lawfulness, fairness, and transparency in AI systems?' - and two added new sections.

What you need to know:

- New content added on [AI and inferences](#), [affinity groups](#) and [special category data](#)

What do we need to know about accuracy and statistical accuracy?

Change overview: This is new chapter with old content.

What you need to know:

- Following the restructuring under the data protection principles, the statistical accuracy content – that used to reside with the chapter 'What do we need to do to ensure lawfulness, fairness, and transparency in AI systems?' - has moved into a new chapter that will focus on the accuracy principle. Statistical accuracy continues to remain key for fairness but we felt it was more appropriate to host it under a chapter that focuses on the accuracy principle.

Fairness in AI

Change overview: This is a new chapter with new and old content.

What you need to know:

The old content was extracted from the former chapter titled 'What do we need to do to ensure lawfulness, fairness, and transparency in AI systems?'. The new content includes information on:

- Data protection's approach to fairness, how it applies to AI and a non-exhaustive list of legal provisions to consider.
- The difference between fairness, algorithmic fairness, bias and discrimination.
- High level considerations when thinking about evaluating fairness and inherent trade-offs.
- Processing personal data for bias mitigation.
- Technical approaches to mitigate algorithmic bias.
- How are solely automated decision-making and relevant safeguards linked to fairness, and key questions to ask when considering Article 22 of the UK GDPR.

Annex A: Fairness in the AI lifecycle

Change overview: This is a new chapter with new content

- This section is about data protection fairness considerations across the AI lifecycle, from problem formulation to decommissioning. It sets out why fundamental aspects of building AI such as underlying assumptions, abstractions used to model a problem, the selection of target variables or the tendency to over-rely on quantifiable proxies may have an impact on fairness. This chapter also explains the different sources of bias that can lead to unfairness and possible mitigation measures. Technical terms are also explained in the updated glossary.

Glossary

Change overview: This is an old chapter with old and new content.

What you need to know:

New additions include definitions of:

- Affinity groups, algorithmic fairness, algorithmic fairness constraints, bias mitigation algorithm, causality, confidence interval, correlation, cost function, dataset labellers, decision, construct and observed space, decision boundary, decision tree, downstream effects, ground truth, inductive bias, in-processing, hyperparameters, multi-criteria optimisation, objective function, post-processing bias mitigation, regularisation, redundant encodings, reward function, use case, target variable, variance.

About this guidance

At a glance

This guidance covers what we think is best practice for data protection-compliant AI, as well as how we interpret data protection law as it applies to AI systems that process personal data. The guidance is not a statutory code. It contains advice on how to interpret relevant data protection law as it applies to AI, and recommendations on good practice for organisational and technical measures to mitigate the risks to individuals that AI may cause or exacerbate.

In detail

- [Why have you produced this guidance?](#)
- [What do you mean by 'AI'?](#)
- [How does this guidance relate to other ICO work on AI?](#)
- [What is a risk-based approach to AI?](#)
- [Is this guidance a set of AI principles?](#)
- [What legislation applies?](#)
- [How is this guidance structured?](#)
- [Who is this guidance for?](#)
- [How should we use this guidance?](#)

Why have you produced this guidance?

We see new uses of artificial intelligence (AI) everyday, from healthcare to recruitment, to commerce and beyond.

We understand the benefits that AI can bring to organisations and individuals, but there are risks too. We have set out some of these risks, such as AI-driven discrimination in ICO25, our strategic plan for the next two years. Enabling good practice in AI has been one of our [regulatory priorities](#) for some time, and we developed this guidance on AI and data protection to help organisations comply with their data protection obligations.

The guidance:

- gives us a clear methodology to audit AI applications and ensure they process personal data fairly, lawfully and transparently;
- ensures that the necessary measures are in place to assess and manage risks to rights and freedoms that arise from AI; and
- supports the work of our investigation and assurance teams when assessing the compliance of organisations using AI.

As well as using the guidance to support our own audit and enforcement activity, we also wanted to share

our thinking behind it. The framework therefore has three distinct outputs:

1. Auditing tools and procedures which our investigation and assurance teams will use when assessing the compliance of organisations using AI. The specific auditing and investigation activities they undertake vary, but can include off-site checks, on-site tests and interviews, and in some cases the recovery and analysis of evidence, including AI systems themselves.
2. This detailed guidance on AI and data protection for organisations, which outlines our thinking.
3. A [toolkit designed to provide further practical support to organisations auditing the compliance of their own AI systems](#).

This guidance covers what we think is best practice for data protection-compliant AI, as well as how we interpret data protection law as it applies to AI systems that process personal data.

This guidance is not a statutory code. It contains advice on how to interpret relevant data protection law as it applies to AI, and recommendations on good practice for organisational and technical measures to mitigate the risks to individuals that AI may cause or exacerbate. There is no penalty if you fail to adopt good practice recommendations, as long as you find another way to comply with the law.

This guidance is restricted to data protection law. There are other legal frameworks and obligations relevant to organisations developing and deploying AI that will need to be considered, including the Equality Act 2010 as well as sector specific law and regulations. You will need to consider these obligations in addition to this guidance.

What do you mean by ‘AI’?

Data protection law does not use the term ‘AI’, so none of your legal obligations depend on exactly how it is defined. However, it is useful to understand broadly what we mean by AI in the context of this guidance. AI has a variety of meanings, including:

- In the AI research community, it refers to various methods ‘[for using a non-human system to learn from experience and imitate human intelligent behaviour](#)’; or
- in the data protection context, ‘[the theory and development of computer systems able to perform tasks normally requiring human intelligence](#)’.

We use the umbrella term ‘AI’ because it has become a standard industry term for a range of technologies. One prominent area of AI is ‘machine learning’ (ML), which is the use of computational techniques to create (often complex) statistical models using (typically) large quantities of data. Those models can be used to make classifications or predictions about new data points.

While not all AI involves ML, most of the recent interest in AI is driven by ML in some way, whether in image recognition, speech-to-text, or classifying credit risk. This guidance therefore focuses on the data protection challenges that ML-based AI may present, while acknowledging that other kinds of AI may give rise to other data protection challenges.

You may already process personal data in the context of creating statistical models, and using those models to make predictions about people. Much of this guidance will still be relevant to you even if you do not class these activities as ML or AI. Where there are important differences between types of AI, for example, simple regression models and deep neural networks, we will refer to these explicitly.

Further reading outside this guidance

[International Working Group on Data Protection in Telecommunications' working paper on privacy and artificial intelligence](#)

How does this guidance relate to other ICO work on AI?

This guidance is designed to complement existing ICO resources, including:

- the [Big Data, AI, and Machine Learning report](#), published in 2014 and updated in 2017; and
- our guidance on Explaining decisions made with AI, produced in collaboration with The Alan Turing Institute.

The Big Data report provided a strong foundation for understanding the data protection implications of these technologies. As noted in the Commissioner's foreword to the 2017 edition, this is a complicated and fast-developing area. New considerations have arisen since, both in terms of the risks AI poses to individuals, and the organisational and technical measures that can be taken to address those risks. Through our engagement with stakeholders, we gained additional insights into how organisations are using AI on the ground, which go beyond those presented in the 2017 report.

Another significant challenge raised by AI is **explainability**. As part of the government's AI Sector Deal, in collaboration with the Alan Turing Institute (The Turing) we have produced guidance on how organisations can best explain their use of AI to individuals. This resulted in the '[Explaining decisions made with AI](#)' guidance, which was published in May 2020.

While the Explaining decisions made with AI guidance already covers the challenge of AI explainability for individuals in substantial detail, this guidance includes some additional considerations about AI explainability **within** the organisation, eg for internal oversight and compliance. The two pieces of guidance are complementary, and we recommend reading them together.

Further reading outside this guidance

- See our report on '[Big data, artificial intelligence, machine learning and data protection](#)'
- [ICO and The Alan Turing Institute guidance on 'Explaining decisions made with AI'](#)
- See our [AI related work](#).

What is a risk-based approach to AI?

Taking a risk-based approach means:

- assessing the risks to the rights and freedoms of individuals that may arise when you use AI; and
- implementing appropriate and proportionate technical and organisational measures to mitigate these risks.

These are general requirements in data protection law. They do not mean you can ignore the law if the

risks are low, and they may mean you have to stop a planned AI project if you cannot sufficiently mitigate those risks.

To help you integrate this guidance into your existing risk management process, we have organised it into several major risk areas. For each risk area, we describe:

- the risks involved;
- how AI may increase their likelihood and/or impact; and
- some possible measures which you could use to identify, evaluate, minimise, monitor and control those risks.

The technical and organisational measures included are those we consider good practice in a wide variety of contexts. However, since many of the risk controls that you may need to adopt are context-specific, we cannot include an exhaustive or definitive list.

This guidance covers both the AI-and-data-protection-specific risks, **and** the implications of those risks for governance and accountability. Regardless of whether you are using AI, you should have accountability measures in place.

However, adopting AI applications may require you to re-assess your existing governance and risk management practices. AI applications can exacerbate existing risks, introduce new ones, or generally make risks more difficult to assess or manage. Decision-makers in your organisation should therefore reconsider your organisation's risk appetite in light of any existing or proposed AI applications.

Each of the sections of this guidance deep-dives into one of the AI challenge areas and explores the associated risks, processes, and controls.

Is this guidance a set of AI principles?

This guidance does not provide generic ethical or design principles for the use of AI. While there may be overlaps between 'AI ethics' and data protection (with some proposed ethics principles already reflected in data protection law), this guidance is focused on data protection compliance.

Although data protection does not dictate how AI developers should do their jobs, if you use AI to process personal data, you need to comply with the principles of data protection by design and by default.

Certain design choices are more likely to result in AI systems which infringe data protection in one way or other. This guidance will help developers and engineers understand those choices better, so you can design high-performing systems whilst still protecting the rights and freedoms of individuals.

It is worth noting that our work focuses exclusively on the data protection challenges introduced or heightened by AI. Therefore, more general data protection considerations, are not addressed in this guidance, except in so far as they relate to and are challenged by AI. Neither does it cover AI-related challenges which are outside the remit of data protection.

Further reading outside this guidance

[Global Privacy Assembly's 'Declaration on ethics and data protection in artificial intelligence'](#)

What legislation applies?

This guidance deals with the challenges that AI raises for data protection. The most relevant piece of UK legislation is the Data Protection Act 2018 (DPA 2018).

The DPA 2018, together with the UK General Data Protection Regulation (UK GDPR), set out the UK's data protection regime. Please note that from January 2021, you should read references to the GDPR as references to the equivalent articles in the UK GDPR. The DPA 2018 comprises the following data protection regimes:

- Part 2 – covers general processing and supplements and tailors the UK GDPR;
- Part 3 – sets out a separate regime for law enforcement authorities; and
- Part 4 – sets out a separate regime for the three intelligence services.

Most of this guidance will apply regardless of which part of the DPA applies to your processing. However, where there are relevant differences between the requirements of the regimes, these are explained in the text.

You should also review our guidance on how the end of the transition period impacts data protection law.

The impacts of AI on areas of ICO competence other than data protection, notably Freedom of Information, are not considered here.

Further reading outside this guidance

- See our [guide to data protection](#).
- If you need more detail on data protection and Brexit, see our [FAQs](#).

How is this guidance structured?

This guidance is divided into several parts covering different data protection principles and rights.

The general structure is based on the foundational principles of data protection:

- lawfulness, fairness, and transparency;
- purpose limitation;
- data minimisation;
- accuracy;
- storage limitation; and
- security and accountability.

It also provides more in-depth analysis of measures to comply with people's individual rights.

In order to provide more guidance to AI developers and AI risk managers, we have also created an AI and data protection risk toolkit and technical guidance on how data protection approaches fairness in AI that follows the AI lifecycle in Annex A. A glossary at the end of this guidance provides more background information for non-technical professionals who want to understand the more technical aspects of the

existing guidance.

As the technology evolves and legislation changes, we are likely to update this guidance.

Who is this guidance for?

This guidance covers best practices for data protection-compliant AI. There are two broad audiences.

First, those with a compliance focus, including:

- data protection officers (DPOs);
- general counsel;
- risk managers;
- senior management; and
- the ICO's own auditors – in other words, we will use this guidance as a basis to inform our audit functions under the data protection legislation.

Second, technology specialists, including:

- machine learning developers and data scientists;
- software developers / engineers; and
- cybersecurity and IT risk managers.

The guidance is split into four sections that cover areas of data protection legislation that you need to consider.

While this guidance is written to be accessible to both audiences, some parts are aimed primarily at those in either compliance or technology roles and are signposted accordingly at the start of each section as well as in the text.

How should we use this guidance?

In each section, we discuss what you **must** do to comply with data protection law as well as what you **should** do as good practice. This distinction is generally marked using 'must' when it relates to compliance with data protection law and using 'should' where we consider it good practice but not essential to comply with the law. Discussion of good practice is designed to help you if you are not sure what to do, but it is not prescriptive. It should give you enough flexibility to develop AI systems which conform to data protection law in your own way, taking a proportionate and risk-based approach.

The guidance assumes familiarity with key data protection terms and concepts. We also discuss in more detail data protection-related terms and concepts where it helps to explain the risks that AI creates and exacerbates.

The guidance also assumes familiarity with AI-related terms and concepts. These are further explained in the glossary at the end of this guidance.

The guidance focuses on specific risks and controls to ensure your AI system is compliant with data protection law and provides safeguards for individuals' rights and freedoms. It is not intended as an exhaustive guide to data protection compliance. You need to make sure you are aware of all your

obligations and you should read this guidance alongside our other guidance. Your DPIA process should incorporate measures to comply with your data protection obligations generally, as well as conform to the specific standards in this guidance.

What are the accountability and governance implications of AI?

■ [Latest updates](#)

15 March 2023 - This is an old chapter with new additions, including what we need to include in our DPIA.

At a glance

This section is about the accountability principle, which makes you responsible for complying with data protection law and for demonstrating that compliance in any AI system that processes personal data. A data protection impact assessment (DPIA) is an ideal way to demonstrate your compliance. The section will also explain the importance of identifying and understanding controller/ processor relationships. Finally, it covers striking the required balance between the right to data protection and other fundamental rights in the context of your AI system.

Who is this section for?

This section is aimed at senior management and those in compliance-focused roles, including DPOs, who are accountable for the governance and data protection risk management of an AI system. There are some terms and techniques described that may require the input of a technical specialist.

In detail

- [How should we approach AI governance and risk management in the data protection context?](#)
- [How should we set a meaningful risk appetite?](#)
- [What do we need to consider when undertaking data protection impact assessments for AI?](#)
- [How should we understand controller/processor relationships in AI?](#)
- [How should we manage competing interests when assessing AI-related risks?](#)

How should we approach AI governance and risk management?

If used well, AI has the potential to make organisations more efficient, effective and innovative. However, AI also raises significant risks for the rights and freedoms of individuals, as well as compliance challenges for organisations.

Different technological approaches will either exacerbate or mitigate some of these issues, but many others are much broader than the specific technology. As the rest of this guidance suggests, the data protection implications of AI are heavily dependent on the specific use cases, the population they are deployed on, other overlapping regulatory requirements, as well as social, cultural and political considerations.

While AI increases the importance of embedding data protection by design and default into an organisation's culture and processes, the technical complexities of AI systems can make this more difficult. Demonstrating how you have addressed these complexities is an important element of accountability.

You cannot delegate these issues to data scientists or engineering teams. Your senior management, including DPOs, are also accountable for understanding and addressing them appropriately and promptly (although overall accountability for data protection compliance lies with the controller, ie your organisation).

To do so, in addition to their own upskilling, your senior management will need diverse, well-resourced teams to support them in carrying out their responsibilities. You also need to align your internal structures, roles and responsibilities maps, training requirements, policies and incentives to your overall AI governance and risk management strategy.

It is important that you do not underestimate the initial and ongoing level of investment of resources and effort that is required. You must be able to demonstrate, on an ongoing basis, how you have addressed data protection by design and default obligations. Your governance and risk management capabilities need to be proportionate to your use of AI. This is particularly true now while AI adoption is still in its initial stages, and the technology itself, as well as the associated laws, regulations, governance and risk management best practices are developing quickly.

We [have also developed a more general accountability framework](#). This is not specific to AI, but provides a baseline for demonstrating your accountability under the UK GDPR, on which you could build your approach to AI accountability.

[Annex A: Fairness in the AI lifecycle also includes aspects you should consider.](#)

How should we set a meaningful risk appetite?

The risk-based approach of data protection law requires you to comply with your obligations and implement appropriate measures in the context of your particular circumstances – the nature, scope, context and purposes of the processing you intend to do, and the risks this poses to individuals' rights and freedoms. That is to say, you need to identify the risks to people's data protection rights associated with your processing activities. This will help you to determine the measures you need to put in place to ensure your processing complies with your data protection obligations.

Your compliance considerations therefore involve assessing the risks to the rights and freedoms of individuals and judging what is appropriate in those circumstances. In all cases, you need to ensure you comply with data protection requirements.

This applies to the use of AI just as to other technologies that process personal data. In the context of AI, the specific nature of the risks posed and the circumstances of your processing will require you to strike an appropriate balance between competing interests as you go about ensuring data protection compliance. This may in turn impact the outcome of your processing. It is unrealistic to adopt a 'zero tolerance' approach to risks to rights and freedoms, and indeed the law does not require you to do so. It is about ensuring that these risks are identified, managed and mitigated. We talk about trade-offs and how you should manage them below and provide examples of some trade-offs throughout the guidance.

To manage the risks to individuals that arise from processing personal data in your AI systems, it is important that you develop a mature understanding of fundamental rights, risks, and how to balance these

and other interests. Ultimately, it is necessary for you to:

- assess the risks to individual rights that your use of AI poses;
- determine how you will address these; and
- establish the impact this has on your use of AI.

You should ensure your approach fits both your organisation and the circumstances of your processing. Where appropriate, you should also use risk assessment frameworks.

This is a complex task, which can take time to get right. However, it will give you, as well as the ICO, a fuller and more meaningful view of your risk positions and the adequacy of your compliance and risk management approaches.

The following sections deal with the AI-specific implications of accountability including:

- how you should undertake data protection impact assessments for AI systems;
- how you can identify whether you are a controller or processor for specific processing operations involved in the development and deployment of AI systems and the resulting implications for your responsibilities;
- how you should assess the risks to the rights and freedoms of individuals, and how you should address them when you design, or decide to use, an AI system; and
- how you should justify, document and demonstrate the approach you take, including your decision to use AI for the processing in question.

What do we need to consider when undertaking data protection impact assessments for AI?

DPIAs are a key part of data protection law's focus on accountability and data protection by design.

You should not see DPIAs as simply a box ticking compliance exercise. They can effectively act as roadmaps for you to identify and control the risks to rights and freedoms that using AI can pose. They are also an ideal opportunity for you to consider and demonstrate your accountability for the decisions you make in the design or procurement of AI systems.

Why are DPIAs required under the data protection law?

In the vast majority of cases, the use of AI will involve a type of processing likely to result in a high risk to individuals' rights and freedoms, and will therefore trigger the legal requirement for you to undertake a DPIA. You will need to make this assessment on a case by case basis. In those cases where you assess that a particular use of AI does not involve high risk processing, you still need to document how you have made this assessment.

If the result of an assessment indicates residual high risk to individuals that you cannot sufficiently reduce, you must consult with the ICO prior to starting the processing.

In addition to conducting a DPIA, you may also be required to undertake other kinds of impact assessments or do so voluntarily. For example, public sector organisations are required to undertake equality impact assessments, while other organisations voluntarily undertake 'algorithm impact assessments'. Similarly, the machine learning community has proposed '[model cards](#)' and '[datasheets](#)' which describe how ML models may perform under different conditions, and the context behind the datasets they are trained on, which

may help inform an impact assessment. There is no reason why you cannot combine these exercises, so long as the assessment encompasses all the requirements of a DPIA.

The ICO has produced [detailed guidance on DPIAs](#) that explains when they are required and how to complete them. This section sets out some of the things you should think about when carrying out a DPIA for the processing of personal data in AI systems.

Further Reading

 [Relevant provisions in the legislation - see Articles 35 and 36 and Recitals 74-77, 84, 89-92, 94 and 95 of the UK GDPR](#) 

External link

 [See Sections 64 and 65 of the DPA 2018](#) 

External link

How do we decide whether to do a DPIA?

We acknowledge that not all uses of AI will involve types of processing that are likely to result in a high risk to rights and freedoms. However, you should note that Article 35(3)(a) of the UK GDPR requires you to undertake a DPIA if your use of AI involves:

- systematic and extensive evaluation of personal aspects based on automated processing, including profiling, on which decisions are made that produce legal or similarly significant effects;
- large-scale processing of special categories of personal data; or
- systematic monitoring of publicly-accessible areas on a large scale.

Beyond this, AI can also involve several processing operations that are themselves likely to result in a high risk, such as use of new technologies or novel application of existing technologies, data matching, invisible processing, and tracking of location or behaviour. When these involve things like evaluation or scoring, systematic monitoring, and large-scale processing, the requirement to do a DPIA is triggered.

In any case, if you have a major project that involves the use of personal data it is also good practice to do a DPIA. Read our [list of processing operations 'likely to result in high risk'](#) for examples of operations that require a DPIA, and further detail on which criteria are high risk in combination with others.

Further reading outside this guidance

See '[How do we decide whether to do a DPIA?](#)' in our detailed guidance on DPIAs.

What should we assess in our DPIA?

Your DPIA needs to describe the nature, scope, context and purposes of any processing of personal data. It needs to make clear how and why you are going to use AI to process the data. You need to detail:

- how you will collect, store and use data;

- the volume, variety and sensitivity of the data;
- the nature of your relationship with individuals; and
- the intended outcomes for individuals or wider society, as well as for you.

Whether a system using AI is generally more or less risky than a system not using AI depends on the specific circumstances. You therefore need to evaluate this based on your own context. Your DPIA should show evidence of your consideration of less risky alternatives, if any, that achieve the same purpose of the processing, and why you didn't choose them. This consideration is particularly relevant where you are using public task or legitimate interests as a lawful basis. See ['How do we identify our purposes and lawful basis'](#).

When considering the impact your processing has on individuals, it is important to consider both allocative harms and representational harms:

- Allocative harms are the result of a decision to allocate goods and opportunities among a group. The impact of allocative decisions may be loss of financial opportunity, loss of livelihood, loss of freedom, or in extreme circumstances, loss of life.
- Representational harms occur when systems reinforce the subordination of groups along identity lines. For example, through stereotyping, under-representation, or denigration, meaning belittling or undermining their human dignity.

Example of allocative harm

An organisation may use an AI system in recruitment that disproportionately classifies applications from male candidates as suitable compared to women. The use of this system has implications for the allocation of job opportunities to female candidates and the relevant economic results.

Example of representational harm

An individual belonging to an ethnic minority group uploads pictures of their holiday photos on an internet platform. The image recognition system operated by the platform assigns labels to their 'selfie' photos that are denigrating reflecting racist tropes.

In the context of the AI lifecycle, a DPIA will best serve its purpose if you undertake it at the earliest stages of project development. It should feature, at a minimum, the following key components.

How do we describe the processing?

Your DPIA should include:

- a systematic description of the processing activity, including data flows and the stages when AI processes and automated decisions may produce effects on individuals;
- an explanation of any relevant variation or margins of error in the performance of the system which may affect the fairness of the personal data processing (see ['What do we need to know about accuracy and statistical accuracy'](#)); and
- a description of the scope and context of the processing, including:

- what data you will process;
- the number of data subjects involved;
- the source of the data; and
- to what extent individuals are likely to expect the processing.

Your DPIA should identify and record the degree of any human involvement in the decision-making process and at what stage this takes place. Where automated decisions are subject to human intervention or review, you should implement processes to ensure this is meaningful and also detail the fact that decisions can be overturned.

It can be difficult to describe the processing activity of AI systems, particularly when they involve complex models and data sources. However, such a description is necessary as part of a DPIA. In some cases, although it is not a legal requirement, it may be good practice for you to maintain two versions of an assessment, with:

- the first presenting a thorough technical description for specialist audiences; and
- the second containing a more high-level description of the processing and explaining the logic of how the personal data inputs relate to the outputs affecting individuals (this may also support you in fulfilling your obligation to [explain AI decisions to individuals](#)).

Your DPIA should set out your roles and obligations as a controller and include any processors involved. Where AI systems are partly or wholly outsourced to external providers, both you and any other organisations involved should also assess whether joint controllership exists under Article 26 of the UK GDPR; and if so, collaborate in the DPIA process as appropriate.

If you use a processor, you can illustrate some of the more technical elements of the processing activity in a DPIA by reproducing information from that processor. For example, a flow diagram from a processor's manual. However, you should generally avoid copying large sections of a processor's literature into your own assessment.

Further Reading



[Relevant provisions in the legislation - see Article 35\(7\)\(a\) and Recitals 84, 90 and 94 of the UK GDPR](#)

External link

Do we need to consult anyone?

You must, where appropriate:

- seek and document the views of individuals whose data you will be processing during the AI lifecycle, or their representatives, unless there is a good reason not to;
- consult all relevant internal stakeholders;
- consult with your processor, if you use one; and
- consider seeking legal advice or other expertise.

Unless there is a good reason not to do so, you should seek and document the views of individuals whose personal data you process, or their representatives, on the intended processing operation during a DPIA. It

is therefore important that you can describe the processing in a way that those you consult can understand. However, if you can demonstrate that consultation would compromise commercial confidentiality, undermine security, or be disproportionate or impracticable, these can be reasons not to consult.

You can help to identify the potential risks of your systems by engaging with:

- independent domain experts who have a deep understanding of the context in which your system will be deployed; and
- people with lived experience within that context that could also be impacted by the system.

Further Reading

 [Relevant provisions in the legislation - see Article 28\(3\)\(f\) and Article 35 \(9\) of the UK GDPR](#) 

External link

How do we assess necessity and proportionality?

The deployment of an AI system to process personal data needs to be driven by evidence that there is a problem, and a reasoned argument that AI is a sensible solution to that problem, not by the mere availability of the technology. By assessing necessity in a DPIA, you can evidence that you couldn't accomplish these purposes in a less intrusive way.

A DPIA also allows you to demonstrate that your processing of personal data by an AI system is a proportionate activity. When assessing proportionality, you need to weigh up your interests in using AI against the risks it may pose to the rights and freedoms of individuals. For AI systems, you need to think about any detriment to individuals that could follow from bias or inaccuracy in the algorithms and data sets being used.

Within the proportionality element of a DPIA, you need to assess whether individuals would reasonably expect an AI system to conduct the processing. If AI systems complement or replace human decision-making, you should document in the DPIA how the project might compare human and algorithmic accuracy side-by-side to better justify their use.

You should also describe any trade-offs that are made, for example between statistical accuracy and data minimisation, and document the methodology and rationale for these.

How do we identify and assess risks to individuals?

The DPIA process will help you to objectively identify the relevant risks to individuals' interests. You should assign a score or level to each risk, measured against the likelihood and the severity of the impact on individuals.

The use of personal data in the development and deployment of AI systems may not just pose risks to individuals' information rights. When considering sources of risk, your DPIA should consider the potential impact of other material and non-material damage or harm on individuals.

For example, machine learning systems may reproduce discrimination from historic patterns in data, which could fall foul of equalities legislation. Similarly, AI systems that stop content being published based on the analysis of the creator's personal data could impact their freedom of expression. In these contexts, you

should consider the relevant legal frameworks beyond data protection.

Further Reading

 [Relevant provisions in the legislation - see Articles 35\(7\)\(c\) and Recitals 76 and 90 of the UK GDPR](#) 

External link

How do we identify mitigating measures?

Against each identified risk to individuals' interests, you should consider options to reduce the level of assessed risk further. Examples of this could be data minimisation techniques or providing opportunities for individuals to opt out of the processing.

You should ask your DPO (if you have one) for advice when considering ways to reduce or avoid these risks, and you should record in your DPIA whether your chosen measure reduces or eliminates the risk in question.

It is important that DPOs or other information governance professionals or both are involved in AI projects from the earliest stages. There must be clear and open channels of communication between them and the project teams. This will ensure that they can identify and address these risks early in the AI lifecycle.

Data protection should not be an afterthought, and a DPO's professional opinion should not come as a surprise at the eleventh hour.

You can use a DPIA to document the safeguards you put in place to ensure the individuals responsible for the development, testing, validation, deployment, and monitoring of AI systems are adequately trained and have an understanding of the data protection implications of the processing.

Your DPIA can also evidence the organisational measures you have put in place, such as appropriate training, to mitigate risks associated with human error. You should also document any technical measures designed to reduce risks to the security and accuracy of personal data processed in your AI system.

Once you have introduced measures to mitigate the risks you have identified, the DPIA should document the residual levels of risk posed by the processing.

You are not required to eliminate every risk identified. However, if your assessment indicates a high risk to the data protection rights of individuals that you are unable to sufficiently reduce, you are required to consult the ICO before you can go ahead with the processing.

How do we conclude our DPIA?

You should record:

- what additional measures you plan to take;
- whether each risk has been eliminated, reduced or accepted;
- the overall level of 'residual risk' after taking additional measures;
- the opinion of your DPO, if you have one; and
- whether you need to consult the ICO.

What happens next?

Although you must carry out your DPIA before the processing of personal data begins, you should also consider it to be a 'live' document. This means reviewing the DPIA regularly and undertaking a reassessment where appropriate (eg if the nature, scope, context or purpose of the processing, and the risks posed to individuals, alter for any reason).

For example, depending on the deployment, it could be that the demographics of the target population may shift, or that people adjust their behaviour over time in response to the processing itself. This is a phenomenon in AI known as 'concept drift' (for more, see '[How should we define and prioritise different statistical accuracy measures?](#)').

Further Reading



[Relevant provisions in the legislation - see Articles 35\(11\), 36\(1\) and 39\(1\)\(c\) and Recital 84 of the UK GDPR](#)

External link

Further reading outside this guidance

Read [our guidance on DPIAs](#) in the Guide to the UK GDPR, including [the list of processing operations likely to result in a high risk](#), for which DPIAs are legally required.

You should also read our detailed guidance on [how to do a DPIA](#), including each step described above.

You may also want to read the relevant sections of the Guide on:

- [lawfulness, fairness and transparency](#);
- [lawful basis for processing](#);
- [data minimisation](#); and
- [accuracy](#).

Further reading – European Data Protection Board

The European Data Protection Board (EDPB), which has replaced the Article 29 Working Party (WP29), includes representatives from the data protection authorities of each EU member state. It adopts guidelines for complying with the requirements of the EU version of the GDPR.

The EDPB has produced guidelines on:

- [Data protection impact assessments](#);
- [Data Protection Officers \('DPOs'\)](#); and
- [Automated individual decision-making and profiling](#)

EDPB guidelines are no longer directly relevant to the UK regime and are not binding under the UK regime. However, they may still provide helpful guidance on certain issues.

How should we understand controller / processor relationships in AI?

Why is controllership important for AI systems?

Often, several different organisations will be involved in developing and deploying AI systems which process personal data.

The UK GDPR recognises that not all organisations involved in the processing will have the same degree of control or responsibility. It is important to be able to identify who is acting as a controller, a joint controller or a processor so you understand which UK GDPR obligations apply to which organisation.

How do we determine whether we are a controller or a processor?

You should use our [existing guidance on controllers and processors](#) to help you with this. This is a complicated area, but some key points from that guidance are:

- You should take the time to assess, and document, the status of each organisation you work with in respect of all the personal data processing activities you carry out.
- If you exercise overall control of the purpose and means of the processing of personal data – you decide what data to process, why and how – you are a controller.
- If you don't have any purpose of your own for processing the data and you only act on a client's instructions, you are likely to be a processor – even if you make some technical decisions about how you process the data.
- Organisations that determine the purposes and means of processing will be controllers regardless of how they are described in any contract about processing services.

As AI usually involves processing personal data in several different phases or for several different purposes, it is possible that you may be a controller or joint controller for some phases or purposes, and a processor for others.

What type of decisions mean we are a controller?

Our guidance says that if you make any of the following overarching decisions, you will be a controller:

- to collect personal data in the first place;
- what types of personal data to collect;
- the purpose or purposes the data are to be used for;
- which individuals to collect the data about;
- how long to retain the data; and
- how to respond to requests made in line with individuals' rights.

For more information, see the [are we a controller?](#) checklist in our Guide to UK GDPR, and our more [detailed guidance on controllers and processors](#).

What type of decisions can we take as a processor?

Our guidance says that you are likely to be a processor if you don't have any purpose of your own for processing the data and you only act on a client's instructions. You may still be able to make some technical decisions as a processor about how the data is processed (the means of the processing). For example, where allowed in the contract, you may use your technical knowledge to decide:

- the IT systems and methods you use to process personal data;
- how you store the data;
- the security measures that will protect it; and
- how you retrieve, transfer, delete or dispose of that data.

How may these issues apply in AI?

When AI systems involve a number of organisations in the processing of personal data, assigning the roles of controller and processor can become complex. For example, when some of the processing happens in the cloud. This can raise broader questions outside the scope of this guidance.

For example, questions about the types of scenario that could result in an organisation becoming a controller, which may include when an organisation makes decisions about:

- the source and nature of the data used to train an AI model;
- the target output of the model (what is being predicted or classified);
- the broad kinds of ML algorithms that will be used to create models from the data (eg regression models, decision trees, random forests, neural networks);
- feature selection – the features that may be used in each model;
- key model parameters (eg how complex a decision tree can be, or how many models will be included in an ensemble);
- key evaluation metrics and loss functions, such as the trade-off between false positives and false negatives; and
- how any models will be continuously tested and updated: how often, using what kinds of data, and how ongoing performance will be assessed.

We will also consider questions about when an organisation is (depending on the terms of their contract) able to make decisions to support the provision of AI services, and still remain a processor. For example, in areas such as:

- the specific implementation of generic ML algorithms, such as the programming language and code libraries they are written in;
- how the data and models are stored, such as the formats they are serialised and stored in, and local caching;
- measures to optimise learning algorithms and models to minimise their consumption of computing resources (eg by implementing them as parallel processes); and
- architectural details of how models will be deployed, such as the choice of virtual machines, microservices, APIs.

We intend to address these issues in more detail in future guidance products, including additional

AI-specific material, as well as revisions to our cloud computing guidance. As we undertake this work, we will consult and work closely with key stakeholders, including government, to explore these issues and develop a range of scenarios when the organisation remains a data processor as it provides AI services.

In our work to date we have developed some indicative example scenarios:

Example

An organisation provides a cloud-based service consisting of a dedicated cloud computing environment with processing and storage, and a suite of common tools for ML. These services enable clients to build and run their own models, with data they have chosen, but using the tools and infrastructure the organisation provides in the cloud. The clients will be controllers, and the provider is likely to be a processor.

The clients are controllers as they take the overarching decisions about what data and models they want to use, the key model parameters, and the processes for evaluating, testing and updating those models.

The provider as a processor could still decide what programming languages and code libraries those tools are written in, the configuration of storage solutions, the graphical user interface, and the cloud architecture.

Example

An organisation provides live AI prediction and classification services to clients. It develops its own AI models, and allows clients to send queries via an API ('what objects are in this image?') to get responses (a classification of objects in the image).

First, the prediction service provider decides how to create and train the model that powers its services, and processes data for these purposes. It is likely to be a controller for this element of the processing.

Second, the provider processes data to make predictions and classifications about particular examples for each client. The client is more likely to be the controller for this element of the processing, and the provider is likely to be a processor.

Example

An AI service provider isolates different client-specific models. This enables each client to make overarching decisions about their model, including whether to further process personal data from their own context to improve their own model.

As long as the isolation between different controllers is complete and auditable, the client will be the sole controller and the provider will be a processor.

Further reading outside this guidance

This is a complicated area, and you should refer to our specific guidance for more information:

- [Controllers and processors: in brief](#)
- [Controllers and processors: in more detail](#)
- [Contracts and liabilities between controllers and processors](#)

We also intend to explore these issues in more detail when we review our cloud computing guidance in 2021.

The Court of Justice of the European Union (CJEU) has also considered the concepts of controller, joint controller and processor in the following judgments:

- [ULD v Wirtschaftsakademie \(Case C-210/16\)](#)
- [Fashion ID \(Case C-40/17\)](#)

How should we manage competing interests when assessing AI-related risks?

Your use of AI must comply with the requirements of data protection law. However, there can be a number of different values and interests to consider, and these may at times pull in different directions. These are commonly referred to as 'trade-offs', and the risk-based approach of data protection law can help you navigate them. There are several significant examples relating to AI, which we discuss in detail elsewhere:

- The interests in training a sufficiently accurate AI system and in reducing the quantity of personal data processed to train that system (see '[How should we balance data minimisation and statistical accuracy](#)').
- Producing an AI system which is sufficiently statistically accurate and which avoids discrimination (see '[What are the technical approaches to mitigate discrimination risk in ML models?](#)').
- Striking the appropriate balance between explainability and statistical accuracy, security, and commercial secrecy (see the [Explaining Decisions Made with AI guidance](#), and '[What about AI security risks exacerbated by explainable AI?](#)').

If you are using AI to process personal data you therefore need to identify and assess these interests, as part of your broader consideration of the risks to the rights and freedoms of individuals and how you will meet your obligations under the law.

The right balance depends on the specific sectoral and social context you operate in, and the impact the processing may have on individuals. However, there are methods you can use to assess and mitigate trade-offs that are relevant to many use cases.

How can we manage these trade-offs?

In most cases, striking the right balance between these multiple trade-offs is a matter of judgement, specific to the use case and the context an AI system is meant to be deployed in.

Whatever choices you make, you need to be accountable for them. Your efforts should be proportionate to the risks the AI system you are considering to deploy poses to individuals. You should:

- identify and assess any existing or potential trade-offs, when designing or procuring an AI system, and assess the impact it may have on individuals;
- consider available technical approaches to minimise the need for any trade-offs;
- consider any techniques which you can implement with a proportionate level of investment and effort;
- have clear criteria and lines of accountability about the final trade-off decisions. This should include a robust, risk-based and independent approval process;
- where appropriate, take steps to explain any trade-offs to individuals or any human tasked with reviewing AI outputs; and
- review trade-offs on a regular basis, taking into account, among other things, the views of individuals whose personal data is likely to be processed by the AI (or their representatives) and any emerging techniques or best practices to reduce them.

You should document these processes and their outcomes to an auditable standard. This will help you to demonstrate that your processing is fair, necessary, proportionate, adequate, relevant and limited. This is part of your responsibility as a controller under Article 24 and your compliance with the accountability principle under Article 5(2). You must also capture them with an appropriate level of detail where required as part of a DPIA or a legitimate interests assessment (LIA) undertaken in connection with a decision to rely on the "legitimate interests" lawful basis for processing personal data.

You should also document:

- how you have considered the risks to the individuals that are having their personal data processed;
- the methodology for identifying and assessing the trade-offs in scope; the reasons for adopting or rejecting particular technical approaches (if relevant);
- the prioritisation criteria and rationale for your final decision; and
- how the final decision fits within your overall risk appetite.

You should also be ready to halt the deployment of any AI systems, if it is not possible to achieve a balance that ensures compliance with data protection requirements.

Outsourcing and third-party AI systems

When you either buy an AI solution from a third party, or outsource it altogether, you need to conduct an independent evaluation of any trade-offs as part of your due diligence process. You are also required to specify your requirements at the procurement stage, rather than addressing trade-offs afterwards.

Recital 78 of the UK GDPR says producers of AI solutions should be encouraged to:

- take into account the right to data protection when developing and designing their systems; and
- make sure that controllers and processors are able to fulfil their data protection obligations.

You should ensure that any system you procure aligns with what you consider to be the appropriate trade-offs. If you are unable to assess whether the use of a third party solution would be data protection compliant, then you should, as a matter of good practice, opt for a different solution. Since new risks and compliance considerations may arise during the course of the deployment, you should regularly review any outsourced services and be able to modify them or switch to another provider if their use is no longer compliant in your circumstances.

For example, a vendor may offer a CV screening tool which effectively scores promising job candidates but may ostensibly require a lot of information about each candidate to assist with the assessment. If you are procuring such a system, you need to consider whether you can justify collecting so much personal data from candidates, and if not, request the provider modify their system or seek another provider.

Further reading inside this guidance

See our section on [‘what data minimisation and privacy-preserving techniques are available for AI systems?’](#)

Culture, diversity and engagement with stakeholders

You need to make significant judgement calls when determining the appropriate trade-offs. While effective risk management processes are essential, the culture of your organisation also plays a fundamental role.

Undertaking this kind of exercise will require collaboration between different teams within the organisation. Diversity, incentives to work collaboratively, as well as an environment in which staff feel encouraged to voice concerns and propose alternative approaches are all important.

The social acceptability of AI in different contexts, and the best practices in relation to trade-offs, are the subject of ongoing societal debates. Consultation with stakeholders outside your organisation, including those affected by the trade-off, can help you understand the value you should place on different criteria.

What about mathematical approaches to minimise trade-offs?

In some cases, you can precisely quantify elements of the trade-offs. A number of mathematical and computer science techniques known as ‘constrained optimisation’ aim to find the optimal solutions for minimising trade-offs.

For example, the theory of differential privacy provides a framework for quantifying and minimising trade-offs between the knowledge that can be gained from a dataset or statistical model, and the privacy of the people in it. Similarly, various methods exist to create ML models which optimise statistical accuracy while also minimising mathematically defined measures of discrimination.

While these approaches provide theoretical guarantees, it can be hard to meaningfully put them into practice. In many cases, values like privacy and fairness are difficult to meaningfully quantify. For example, differential privacy may be able to measure the likelihood of an individual being uniquely identified from a particular dataset, but not the sensitivity of that identification. Therefore, they may not always be appropriate. If you do decide to use mathematical and computer science techniques to minimise trade-offs, you should always supplement these methods with a more qualitative and holistic approach. But the inability to precisely quantify the values at stake does not mean you can avoid assessing and justifying the

trade-off altogether; you still need to justify your choices.

In many cases trade-offs are not precisely quantifiable, but this should not lead to arbitrary decisions. You should perform contextual assessments, documenting and justifying your assumptions about the relative value of different requirements for specific AI use cases.

How do we ensure transparency in AI?

■ [Latest updates](#)

15 March 2023 - This is a new chapter with new content new high-level content on the transparency principle as it applies to AI. The main guidance on transparency and explainability resides within our existing [Explaining Decisions Made with AI](#) product.

What are our transparency obligations towards people?

You need to be transparent about how you process personal data in an AI system, to comply with the principle of transparency.

Before you begin your processing, you must consider your transparency obligations towards individuals whose personal data you plan to process. The core issues about AI and the transparency principle are addressed in 'Explaining decisions made with AI' guidance, so are not discussed in detail here

At a high level, you need to include the following in the privacy information:

- your purposes for processing their personal data;
- your retention periods for that personal data; and
- who you will share it with.

If you collect data directly from individuals, you must provide that privacy information to them at the time you collect it, before you use it to train a model or apply that model on those individuals. If you collect it from other sources, you must provide this information within a reasonable period and no later than one month, or even earlier if you contact that person or disclose that data to someone else.

We have published the guidance on [Explaining Decisions Made with AI](#) that sets out how you can let people know how you use their data in your AI lifecycle. It offers good practice that can help you comply with the transparency principle.

Supplementary reading in ICO guidance

- [Explaining decisions made with AI](#)
- [What is Transparency?](#)
- [Right to be informed - What common issues might come up in practice?](#)

How do we ensure lawfulness in AI?

At a glance

This section sets out what to consider when you decide on your lawful basis for processing personal data in the context of AI development and deployment. It also includes up-to-date content on the status of AI-driven inferences.

Who is this section for?

This section is aimed at compliance-focused roles, including senior management, who are responsible for ensuring the processing using AI is lawful, fair, and transparent. It will also be of interest for AI developers to help them understand what kind of data they are likely to have access to, depending on the purpose they are using it for.

In detail

- [What should we consider when deciding lawful bases?](#)
- [How should we distinguish purposes between AI development and deployment?](#)
- [Can we rely on consent?](#)
- [Can we rely on performance of a contract?](#)
- [Can we rely on legal obligation, public task or vital interests?](#)
- [Can we rely on legitimate interests?](#)
- [What about special category data and data about criminal offences?](#)
- [Using AI systems to make inferences](#)
- [What about inferences and affinity groups?](#)

What should we consider when deciding lawful bases?

The development and deployment of AI systems involve processing personal data in different ways for different purposes. You must break down and separate each distinct processing operation, and identify the purpose and an appropriate lawful basis for each one, in order to comply with the principle of lawfulness.

Whenever you are processing personal data – whether to train a new AI system, or make predictions using an existing one – you must have an appropriate lawful basis to do so.

Different lawful bases may apply depending on your particular circumstances. However, some lawful bases may be more likely to be appropriate for the training and / or deployment of AI than others. In some cases, more than one lawful basis may be appropriate.

At the same time, you must remember that:

- it is **your responsibility** to decide which lawful basis applies to your processing;

- you must always choose the lawful basis that **most closely reflects the true nature of your relationship** with the individual and the purpose of the processing;
- you should make this determination **before** you start your processing;
- you should **document** your decision;
- you **cannot swap** lawful bases at a later date without good reason;
- you must **include your lawful basis** in your privacy notice (along with the purposes); and
- if you are processing **special categories of data** you need **both** a lawful basis **and** an additional condition for processing.

Further reading outside this guidance

Read our [guidance on lawful basis for processing](#).

How should we distinguish purposes between AI development and deployment?

In many cases, when determining your purpose(s) and lawful basis, it will make sense for you to separate the **research and development** phase (including conceptualisation, design, training and model selection) of AI systems from the **deployment** phase. This is because these are distinct and separate purposes, with different circumstances and risks.

Therefore, it may sometimes be more appropriate to choose different lawful bases for your AI development and deployment. For example, you need to do this when:

- the AI system was developed for a general-purpose task, and you then deploy it in different contexts for different purposes. For example, a facial recognition system could be trained to recognise faces, but that functionality could be used for multiple purposes, such as preventing crime, authentication, and tagging friends in a social network. Each of these further applications might require a different lawful basis;
- you implement an AI system from a third party, any processing of personal data undertaken by the developer will have been for a different purpose (eg to develop the system) to what you intend to use the system for, therefore you may need to identify a different lawful basis; and
- processing of personal data for the purposes of training a model may not directly affect the individuals, but once the model is deployed, it may automatically make decisions which have legal or significant effects. This means the provisions on automated decision-making apply; as a result, a different range of available lawful bases may apply at the development and deployment stages.

The following sections outline some AI-related considerations for each of the UK GDPR's lawful bases. They do not consider Part 3 of the DPA (law enforcement processing) at this stage.

Can we rely on consent?

Consent may be an appropriate lawful basis in cases where you have a direct relationship with the individuals whose data you want to process.

However, you must ensure that consent is freely given, specific, informed and unambiguous, and involves a clear affirmative act on the part of the individuals.

The advantage of consent is that it can lead to more trust and buy-in from individuals when they are using your service. Providing individuals with control can also be a factor in your DPIAs.

However, for consent to apply, individuals must have a genuine choice about whether you can use their data. This may have implications depending on what you intend to do with the data – it can be difficult to ensure you collect valid consent for more complicated processing operations, such as those involved in AI. For example, the more things you want to do with the data, the more difficult it is to ensure that consent is genuinely specific and informed.

The key is that individuals understand how you are using their personal data and have consented to this use. For example, if you want to collect a wide range of features to explore different models to predict a variety of outcomes, consent may be an appropriate lawful basis, provided that you inform individuals about these activities and obtain valid consent.

Consent may also be an appropriate lawful basis for the use of an individual's data during deployment of an AI system (eg for purposes such as personalising the service or making a prediction or recommendation).

However, you should be aware that for consent to be valid, individuals must also be able to withdraw consent as easily as they gave it. If you are relying on consent as the basis of processing data with an AI system during deployment (eg to drive personalised content), you should be ready to accommodate the withdrawal of consent for this processing.

Further Reading

 [Relevant provisions in the legislation - see UK GDPR Articles 4\(11\), 6\(1\)\(a\) 7, 8, 9\(2\)\(a\) and Recitals 32, 38, 40, 42, 43, 171](#) 

External link

Further reading outside this guidance

Read our [guidance on consent](#).

Further reading – European Data Protection Board

The European Data Protection Board (EDPB), which has replaced the Article 29 Working Party (WP29), includes representatives from the data protection authorities of each EU member state. It adopts guidelines for complying with the requirements of the EU version of the GDPR.

The EDPB has produced [guidelines on consent](#).

EDPB guidelines are no longer directly relevant to the UK regime and are not binding under the UK regime. However, they may still provide helpful guidance on certain issues.

Can we rely on performance of a contract?

This lawful basis applies where the processing using AI is objectively necessary to deliver a contractual service to the relevant individual, or to take steps prior to entering into a contract at the individual's request (eg to provide an AI-derived quote for a service).

If there is a less intrusive way of processing their data to provide the same service, or if the processing is not in practice objectively necessary for the performance of the contract, then you cannot rely on this lawful basis for the processing of data with AI.

Furthermore, even if it is an appropriate ground for the **use** of the system, this may not be an appropriate ground for processing personal data to **develop** an AI system. If an AI system can perform well enough **without** being trained on the individual's personal data, performance of the contract does not depend on such processing. Since machine learning models are typically built using very large datasets, whether or not a single individual's data is included in the training data should have a negligible effect on the system's performance.

Similarly, even if you can use performance of a contract as a lawful basis to provide a quote prior to a contract, this does not mean you can also use it to justify using that data to develop the AI system.

You should also note that you are unlikely to be able to rely on this basis for processing personal data for purposes such as 'service improvement' of your AI system. This is because in most cases, collection of personal data about the use of a service, details of how users engage with that service, or for the development of new functions within that service are not objectively necessary for the provision of a contract. This is because the service can be delivered without such processing.

Conversely, use of AI to process personal data for purposes of personalising content may be regarded as necessary for the performance of a contract – but only in some cases. Whether this processing can be regarded as 'intrinsic' to your service depends on:

- the nature of the service;
- the expectations of individuals; and
- whether you can provide your service without this processing (ie if the personalisation of content by means of an AI system is not integral to the service, you should consider an alternative lawful basis).

Further Reading

 [Relevant provisions in the legislation - see Article 6\(1\)\(b\) and Recital 44 of the UK GDPR](#) 
External link

Further reading outside this guidance

Read our [guidance on contracts](#).

Further reading – European Data Protection Board

[EDPB guidelines on processing under Article 6\(1\)\(b\) in the context of online services.](#)

Can we rely on legal obligation, public task or vital interests?

There are some examples in which the use of an AI system to process personal data may be a **legal obligation**. You may also be required to audit your AI systems to ensure they are compliant with various legislation (including but not limited to data protection), and this may involve processing of personal data. For example, to test how the system performs on different kinds of people. Such processing could rely on legal obligation as a basis, but this would only cover the auditing and testing of the system, not any other use of that data. You must be able to identify the obligation in question, either by reference to the specific legal provision or else by pointing to an appropriate source of advice or guidance that sets it out clearly.

Similarly, if you use AI as part of the exercise of your official authority, or to perform a task in the **public interest** set out by law, the necessary processing of personal data involved may be based on those grounds. This is likely to be relevant to public authorities using AI to deliver public services.

In a limited number of cases, the processing of personal data by an AI system might be based on protecting the **vital interests** of the individuals. For example, for emergency medical diagnosis of patients who are otherwise incapable of providing consent (eg processing an FMRI scan of an unconscious patient by an AI diagnostic system).

It is however very unlikely that vital interests could also provide a basis for **developing** an AI system, because this would rarely directly and immediately result in protecting the vital interests of those individuals, even if the models that are eventually built might later be used to save the lives of other individuals. For the development of potentially life-saving AI systems, it would be better to rely on other lawful bases.

Relevant provisions in the legislation

 [See Article 6\(1\)\(c\) and Recitals 41, 45 of the UK GDPR for provisions about legal obligation](#) 

External link

 [See Article 6\(1\)\(e\) and 6\(3\), and Recitals 41, 45 and 50 of the UK GDPR for provisions about public task](#) 

External link

 [See Article 6\(1\)\(d\), Article 9\(2\)\(c\) and Recital 46 of the UK GDPR for provisions about vital interests](#) 

External link

 [See Sections 7 and 8, and Schedule 1 paras 6 and 7 of the Data Protection Act 2018](#) 

External link

Further reading outside this guidance

Can we rely on legitimate interests?

Depending on your circumstances, you could base your processing of personal data for both development and ongoing use of AI on the legitimate interests lawful basis.

It is important to note that while legitimate interests is the most flexible lawful basis for processing, it is not always the most appropriate. For example, if the way you intend to use people's data would be unexpected or cause unnecessary harm. It also means you are taking on additional responsibility for considering and protecting people's rights and interests. You must also be able to demonstrate the necessity and proportionality of the processing.

Additionally, if you are a public authority you can only rely on legitimate interests if you are processing for a legitimate reason other than performing your tasks as a public authority.

There are three elements to the legitimate interests lawful basis, and it can help to think of these as the 'three-part test'. You need to:

- identify a legitimate interest (the 'purpose test');
- show that the processing is necessary to achieve it (the 'necessity test'); and
- balance it against the individual's interests, rights and freedoms (the 'balancing test').

There can be a wide range of interests that constitute 'legitimate interests' in data protection law. These can be your own or those of third parties, as well as commercial or societal interests. However, the key is understanding that while legitimate interests may be more flexible, it comes with additional responsibilities. It requires you to assess the impact of your processing on individuals and be able to demonstrate that there is a compelling benefit to the processing.

You should address and document these considerations as part of your legitimate interests assessment (LIA). As described above, in the initial research and development phase of your AI system, your purposes may be quite broad, but as more specific purposes are identified, you may need to review your LIA accordingly (or identify a different lawful basis).

Example

An organisation seeks to rely on legitimate interests for processing personal data for the purposes of training a machine learning model.

Legitimate interests may allow the organisation the most room to experiment with different variables for its model.

However, as part of its legitimate interests assessment, the organisation has to demonstrate that the range of variables and models it intends to use is a reasonable approach to achieving its outcome.

It can best achieve this by properly defining all of its purposes and justifying the use of each type of

data collected – this will allow the organisation to work through the necessity and balancing aspects of its LIA. Over time, as purposes are refined, the LIA is revisited.

For example, the mere possibility that some data might be useful for a prediction is not by itself sufficient for the organisation to demonstrate that processing this data is necessary for building the model.

Further Reading

 [Relevant provisions in the legislation - see UK GDPR Article 6\(1\)\(f\) and Recitals 47-49](#) 
External link

Further reading outside this guidance

Read our [guidance on legitimate interests](#).

We have also published a [lawful basis assessment tool](#) which you can use to help you decide what basis is appropriate for you, as well as a [legitimate interests template \(Word\)](#).

What about special category data and data about criminal offences?

If you intend to use AI to process special category data or data about criminal offences, then you will need to ensure you comply with the requirements of Articles 9 and 10 of the UK GDPR, as well as the DPA 2018.

Special category data is personal data that needs more protection because it is sensitive. In order to process it you need a lawful basis under Article 6, as well as a separate condition under Article 9, although these do not have to be linked. For more detail, see our detailed guidance on special category data and [‘How should we address risks of bias and discrimination’](#).

Further Reading

 [Relevant provisions in the legislation - see Articles 9 and 10 of the UK GDPR](#) 
External link

Further reading outside this guidance

Read our guidance on [special category data](#) and on [criminal offence data](#).

Using AI systems to make inferences

You may intend to use AI systems to:

- guess or predict details about someone, using information from various sources; or
- analyse and find correlations between datasets, and use these to categorise, profile or make predictions.

In other words, you may use AI systems to make inferences about individuals or groups. Whether an inference is personal data depends on whether it relates to an identified or identifiable individual.

It may also be possible for you to infer or guess details about someone which fall within what constitutes special categories of data. Whether or not this counts as special category data and triggers Article 9 depends on how certain that inference is, and whether you are deliberately drawing that inference.

That inference is likely to be special category data, if your use of AI means you:

- can (or intend to) infer relevant information about an individual; or
- intend to treat someone differently on the basis of the inference (even if it's not with a reasonable degree of certainty).

Further reading in ICO guidance

See our guidance on "[What is personal data](#)", including:

- [What are identifiers and related factors?](#)
- [What is the meaning of 'relates to'?](#)
- [What about inferences and educated guesses?](#)

What about inferences and affinity groups?

Whether inferences about groups are personal data depends on the circumstances. For example, how easy it is to identify an individual through group membership.

AI systems may aim to make predictions based on patterns within a population. In this sense they may appear to solely concern groups. However, if your AI system involves making inferences about a group – creating [affinity groups](#) – and linking these to a specific individual, then data protection law applies at multiple stages of the processing. More specifically:

- the development stage, involving processing of individuals' personal data to train the model; and
- the deployment stage, where you apply the results of the model to other individuals that were not part of the training dataset on the basis of its predictive features.

This means that even if an individual's personal data is not part of your training dataset, data protection law applies when you use that model on them. This is because it involves you processing their personal data to make a decision or prediction about them, using your model to do so.

It is also important that you don't just consider obvious and immediate tangible damage to people. But also more subtle intangible harms and how the system might affect people's rights and freedoms more generally. This includes any impact on society as a whole. For example, DPIAs require you to consider risks to rights and freedoms of all those that the system might affect.

Additionally, data protection by design requires you to take appropriate steps to:

- implement the data protection principles effectively; and
- integrate necessary safeguards into your processing at the design stage, and throughout the lifecycle.

In the context of AI, your data protection considerations therefore must include:

- the individuals whose personal data you process to train your system; and
- the impact your system has on the rights and freedoms of individuals and society once it is deployed.

If you use an affinity group to profile individuals, you need to comply with the data protection principles, including fairness.

What do we need to know about accuracy and statistical accuracy?

■ [Latest updates](#)

15 March 2023 - This is new chapter with old content. Following the restructuring under the data protection principles, the statistical accuracy content – that used to reside with the chapter ‘What do we need to do to ensure lawfulness, fairness, and transparency in AI systems?’ - has moved into a new chapter that will focus on the accuracy principle. Statistical accuracy continues to remain key for fairness but we felt it was more appropriate to host it under a chapter that focuses on the accuracy.

At a glance

Statistical accuracy refers to the proportion of answers that an AI system gets correct or incorrect.

This section explains the controls you can implement so that your AI systems are sufficiently statistically accurate to ensure that the processing of personal data complies with the fairness principle.

Who is this section for?

This section is aimed at technical specialists, who are best placed to assess the statistical accuracy of an AI system and what personal data is required to improve it. It will also be useful for those in compliance-focused roles to understand how statistical accuracy is linked to fairness.

In detail

- [What is the difference between accuracy in data protection law and ‘statistical accuracy’ in AI?](#)
- [How should we define and prioritise different statistical accuracy measures?](#)
- [What should we consider about statistical accuracy?](#)
- [What should we do about statistical accuracy post-deployment?](#)

What is the difference between accuracy in data protection law and ‘statistical accuracy’ in AI?

It is important to note that the word ‘accuracy’ has a different meaning in the contexts of data protection and AI. Accuracy in data protection is one of the fundamental principles, requiring you to ensure that personal data is accurate and, where necessary, kept up to date. It requires you to take all reasonable steps to make sure the personal data you process is not ‘incorrect or misleading as to any matter of fact’ and, where necessary, is corrected or deleted without undue delay.

Broadly, accuracy in AI (and, more generally, in statistical modelling) refers to how often an AI system guesses the correct answer, measured against correctly labelled test data. The test data is usually separated from the training data prior to training, or drawn from a different source (or both). In many

contexts, the answers the AI system provides will be personal data. For example, an AI system might infer someone's demographic information or their interests from their behaviour on a social network.

So, for clarity, in this guidance, we use the terms:

- 'accuracy' to refer to the accuracy principle of data protection law; and
- 'statistical accuracy' to refer to the accuracy of an AI system itself.

Fairness, in a data protection context, generally means that you should handle personal data in ways that people would reasonably expect and not use it in ways that have unjustified adverse effects on them. Improving the 'statistical accuracy' of your AI system's outputs is one of your considerations to ensure compliance with the fairness principle.

Data protection's **accuracy principle** applies to all personal data, whether it is information about an individual used as an input to an AI system, or an output of the system. However, this does not mean that an AI system needs to be 100% **statistically accurate** to comply with the accuracy principle.

In many cases, the outputs of an AI system are not intended to be treated as factual information about the individual. Instead, they are intended to represent a statistically informed guess as to something which may be true about the individual now or in the future. To avoid such personal data being misinterpreted as factual, you should ensure that your records indicate that they are statistically informed guesses rather than facts. Your records should also include information about the provenance of the data and the AI system used to generate the inference.

You should also record if it becomes clear that the inference was based on inaccurate data, or the AI system used to generate it is statistically flawed in a way which may have affected the quality of the inference.

Similarly, if the processing of the incorrect inference may have an impact on them, an individual may request the inclusion of additional information in their record countering the incorrect inference. This helps ensure that any decisions taken on the basis of the potentially incorrect inference are informed by any evidence that it may be wrong.

The UK GDPR mentions statistical accuracy in the context of profiling and automated decision-making at Recital 71. This states organisations should put in place 'appropriate mathematical and statistical procedures' for the profiling of individuals as part of their technical measures. You should ensure any factors that may result in inaccuracies in personal data are corrected and the risk of errors is minimised.

If you use an AI system to make inferences about people, you need to ensure that the system is sufficiently statistically accurate for your purposes. This does not mean that every inference has to be correct, but you do need to factor in the possibility of them being incorrect and the impact this may have on any decisions that you may take on the basis of them. Failure to do this could mean that your processing is not compliant with the fairness principle. It may also impact on your compliance with the data minimisation principle, as personal data, which includes inferences, must be adequate and relevant for your purpose.

Your AI system therefore needs to be sufficiently statistically accurate to ensure that any personal data generated by it is processed lawfully and fairly.

However, overall statistical accuracy is not a particularly useful measure, and usually needs to be broken down into different measures. It is important to measure and prioritise the right ones. If you are in a

compliance role and are unsure what these terms mean, you should consult colleagues in the relevant technical roles.

Further Reading

 [Relevant provisions in the legislation - see UK GDPR Articles 5\(1\)\(d\), 22 and Recital 71](#) 

External link

Further reading in this guidance

[Model evaluation – How should we evaluate our model’s statistical accuracy?](#)

Further reading outside this guidance

Read our [guidance on accuracy](#) as well as our guidance on the rights to [rectification](#) and [erasure](#).

[European guidelines on automated decision-making and profiling](#).

How should we define and prioritise different statistical accuracy measures?

Statistical accuracy, as a general measure, is about how closely an AI system’s predictions match the correct labels as defined in the test data.

For example, if an AI system is used to classify emails as spam or not spam, a simple measure of statistical accuracy is the number of emails that were correctly classified as spam or not spam, as a proportion of all the emails that were analysed.

However, such a measure could be misleading. For example, if 90% of all emails received to an inbox are spam, then you could create a 90% accurate classifier by simply labelling everything as spam. But this would defeat the purpose of the classifier, as no genuine email would get through.

For this reason, you should use alternative measures of statistical accuracy to assess how good a system is. If you are in a compliance role, you should work with colleagues in technical roles to ensure that you have in place appropriate measures of statistical accuracy given your context and the purposes of processing.

These measures should reflect the balance between two different kinds of errors:

- a **false positive** or ‘type I’ error: these are cases that the AI system incorrectly labels as positive (eg emails classified as spam, when they are genuine); or
- a **false negative** or ‘type II’ error: these are cases that the AI system incorrectly labels as negative when they are actually positive (eg emails classified as genuine, when they are actually spam).

It is important to strike the right balance between these two types of errors. There are more useful measures which reflect these two types of errors, including:

- **precision:** the percentage of cases identified as positive that are in fact positive (also called 'positive predictive value'). For example, if nine out of 10 emails that are classified as spam are actually spam, the **precision** of the AI system is 90%; or
- **recall (or sensitivity):** the percentage of all cases that are in fact positive that are identified as such. For example, if 10 out of 100 emails are actually spam, but the AI system only identifies seven of them, then its **recall** is 70%.

There are trade-offs between precision and recall, which can be assessed using statistical measures. If you place more importance on finding as many of the positive cases as possible (maximising recall), this may come at the cost of some false positives (lowering precision).

In addition, there may be important differences between the consequences of false positives and false negatives on individuals, which could affect the fairness of the processing.

Example

If a CV filtering system being used to assist with selecting qualified candidates for an interview produces a false positive, then an unqualified candidate may be invited to interview, wasting the employer's and the applicant's time unnecessarily.

If it produces a false negative, a qualified candidate will miss an employment opportunity and the organisation will miss a good candidate.

You should prioritise avoiding certain kinds of error based on the severity and nature of the risks.

In general, statistical accuracy as a measure depends on how possible it is to compare the performance of a system's outputs to some 'ground truth' (ie checking the results of the AI system against the real world). For example, a medical diagnostic tool designed to detect malignant tumours could be evaluated against high quality test data, containing known patient outcomes.

In some other areas, a ground truth may be unattainable. This could be because no high-quality test data exists or because what you are trying to predict or classify is subjective (eg whether a social media post is offensive). There is a risk that statistical accuracy is misconstrued in these situations, so that AI systems are seen as being highly statistically accurate even though they are reflecting the average of what a set of human labellers thought, rather than objective truth.

To avoid this, your records should indicate where AI outputs are not intended to reflect objective facts, and any decisions taken on the basis of such personal data should reflect these limitations. This is also an example of where you must take into account the **accuracy principle** – for more information, see our [guidance on the accuracy principle, which refers to accuracy of opinions](#).

Finally, statistical accuracy is not a static measure. While it is usually measured on static test data (held back from the training data), in real life situations AI systems are applied to new and changing populations. Just because a system is statistically accurate about an existing population's data (eg customers in the last year), it may not continue to perform well if there is a change in the characteristics of that population or any other population who the system is applied to in future. Behaviours may change, either of their own

accord, or because they are adapting in response to the system, and the AI system may become less statistically accurate with time.

This phenomenon is referred to in machine learning as 'concept / model drift', and various methods exist for detecting it. For example, you can measure the distance between classification errors over time; increasingly frequent errors may suggest drift.

You should regularly assess drift and retrain the model on new data where necessary. As part of your accountability responsibilities, you should decide and document appropriate thresholds for determining whether your model needs to be retrained, based on the nature, scope, context and purposes of the processing and the risks it poses. For example, if your model is scoring CVs as part of a recruitment exercise, and the kinds of skills candidates need in a particular job are likely to change every two years, you should anticipate assessing the need to re-train your fresh data at least that often.

In other application domains where the main features don't change so often (eg recognising handwritten digits), you can anticipate less drift. You will need to assess this based on your own circumstances.

Further reading inside this guidance

See '[What should we know about accuracy and statistical accuracy](#)'.

Further reading outside this guidance

See our [guidance on the accuracy principle](#).

See '[Learning under concept drift: an overview](#)' for a further explanation of concept drift.

What should we consider about statistical accuracy?

You should always think carefully from the start whether it is appropriate to automate any prediction or decision-making process. This should include assessing the effectiveness of the AI system in making statistically accurate predictions about the individuals whose personal data it processes.

You should assess the merits of using a particular AI system in light of consideration of its effectiveness in making statistically accurate, and therefore valuable, predictions. Not all AI systems demonstrate a sufficient level of statistical accuracy to justify their use.

If you decide to adopt an AI system, then to comply with the data protection principles, you should:

- ensure that all functions and individuals responsible for its development, testing, validation, deployment, and monitoring are adequately trained to understand the associated statistical accuracy requirements and measures;
- make sure data is clearly labelled as inferences and predictions, and is not claimed to be factual;
- ensure you have managed trade-offs and reasonable expectations; and
- adopt a common terminology that staff can use to discuss statistical accuracy performance measures,

including their limitations and any adverse impact on individuals.

What should we do about statistical accuracy post-deployment?

As part of your obligation to implement data protection by design and by default, you should consider statistical accuracy and the appropriate measures to evaluate it from the design phase and test these measures throughout the AI lifecycle. After deployment, you should implement monitoring, the frequency of which should be proportional to the impact an incorrect output may have on individuals. The higher the impact the more frequently you should monitor and report on it. You should also review your statistical accuracy measures regularly to mitigate the risk of concept drift. Your change policy procedures should take this into account from the outset.

Statistical accuracy is also an important consideration if you outsource the development of an AI system to a third party (either fully or partially) or purchase an AI solution from an external vendor. In these cases, you should examine and test any claims made by third parties as part of the procurement process.

Similarly, you should agree regular updates and reviews of statistical accuracy to guard against changing population data and concept/ model drift. If you are a provider of AI services, you should ensure that they are designed in such a way as to allow organisations to fulfil their data protection obligations.

Finally, the vast quantity of personal data you may hold and process as part of your AI systems is likely to put pressure on any pre-existing non-AI processes you use to identify and, if necessary, rectify/ delete inaccurate personal data, whether it is used as input or training/ test data. Therefore, you need to review your data governance practices and systems to ensure they remain fit for purpose.

How do we ensure fairness in AI?

■ [Latest updates](#)

15 March 2023 - This is a new chapter with new and old content.

The old content was extracted from the former chapter titled 'What do we need to do to ensure lawfulness, fairness, and transparency in AI systems?'. The new content includes information on:

- Data protection's approach to fairness, how it applies to AI and a non-exhaustive list of legal provisions to consider.
- The difference between fairness, algorithmic fairness, bias and discrimination.
- High level considerations when thinking about evaluating fairness and inherent trade-offs.
- Processing personal data for bias mitigation.
- Technical approaches to mitigate algorithmic bias.
- How are solely automated decision-making and relevant safeguards linked to fairness, and key questions to ask when considering Article 22 of the UK GDPR.

At a glance

This section explains how you should interpret data protection's fairness principle as it applies to AI. It highlights some of the provisions that can serve as a roadmap towards compliance. It sets out why fairness in data protection is not just about discrimination alongside how important the safeguards on solely automated decision-making that Article 22 provides are for fairness.

Fairness and discrimination are also concepts that exist in other legislation and regulatory frameworks that are relevant when using AI systems, in particular, the Equality Act 2010. This guidance only covers these two concepts in relation to data protection law. You will have other obligations in relation to fairness and discrimination that you will need to consider in addition to this guidance. You should read this in conjunction with Annex A that contains organisational and technical good practice measures to mitigate unfairness.

Who is this section for?

This section is aimed at senior management and those in compliance-focused roles, including DPOs, who are accountable for the governance and data protection risk management of an AI system. It will also be useful for technical specialists who are using, modifying, deploying or generally engaging with AI systems that process personal data. As well as members of the public who want to understand how the fairness principle interacts with AI.

In detail

- How does data protection approach fairness?
- What does data protection fairness mean at a high level?
- How does data protection fairness apply in an AI context?
- How should we approach AI governance and risk management?
- What about fairness, bias and discrimination?
- Is AI the best solution to begin with?
- What is the impact of Article 22 of the UK GDPR on fairness?

How does data protection approach fairness?

Fairness is a key principle of data protection and an overarching obligation when you process personal data. You must use personal data fairly to comply with various sections of the legislation, including Article 5(1)(a) of the UK GDPR, Section 2(1)(a) of the Data Protection Act (2018), as well as Part 3 and Part 4 of the legislation.

In simple terms, fairness means you should only process personal data in ways that people would **reasonably expect** and not use it in any way that could have **unjustified adverse effects** on them. You should not process personal data in ways that are unduly detrimental, unexpected or misleading to the individuals concerned.

If you use an AI system to infer data about people, you need to ensure that the system is sufficiently statistically accurate and avoids discrimination. This is in addition to considering the impact of individuals' reasonable expectations for this processing to be fair.

Any processing of personal data using AI that leads to unjust discrimination between people, will violate the fairness principle. This is because data protection aims to protect individuals' rights and freedoms with regard to the processing of their personal data, not just their information rights. This includes the right to privacy but also the right to non-discrimination. The principle of fairness appears across data protection law, both explicitly and implicitly. More specifically, fairness relates to:

- how you go about the processing; and
- the outcome of the processing (ie the impact it has on individuals).

Depending on your context, you may also have other sector-specific obligations about fairness, statistical accuracy or discrimination that you need to consider alongside your data protection obligations. You will also need to consider your Equality Act 2010 obligations: for more information you should refer to the Equality and Human Rights Commission (EHRC). If you need to process data in a certain way to meet those obligations, data protection does not prevent you from doing so.

We are likely to work closely with other regulators and experts to assess the fairness of outcomes and whether any adverse effects on individuals are justified. For example, the Financial Conduct Authority expects organisations to treat customers fairly. Also, the Consumer Rights Act 2015 that the Competition and Markets Authority oversees, requires businesses entering contracts with consumers to ensure contract terms or notices are fair.

Further reading outside this guidance

What does data protection fairness mean at a high level?

Article 5(1)(a) of the UK GDPR requires that you process personal data “lawfully, fairly and in a transparent manner”. In brief:

- “lawfully” means your processing must satisfy an Article 6 lawful basis (and, if required, an Article 9 condition), as well being lawful in more general terms;
- “fairly” means your processing should not lead to unjustified adverse outcomes for individuals, and should be within their reasonable expectations; and
- “in a transparent manner” means you must properly inform individuals about how and why you intend to process their personal data.

These three elements overlap and support each other. Together, they provide an overall framework that ensures fair processing of personal data and fair outcomes for individuals.

This means that fairness is not just about making sure people know and understand how and why you use their data. It is also not just about ensuring people have control over the processing, where appropriate.

It is essentially about taking into account the overall impact your processing has on people, and how you demonstrate that this impact is proportionate and justified.

How does data protection fairness apply in an AI context?

It is important to note, particularly in AI, that data protection’s specific requirements and considerations work together to ensure your AI systems process personal data fairly and lead to fair outcomes. These are listed below:

1. Data protection by design and by default

Data protection by design and by default requires you to consider data protection issues at the design stage of your processing activities, and throughout the AI lifecycle. This is set out in Article 25 of the UK GDPR.

It is about putting in place technical and organisational measures to:

- implement the data protection principles effectively; and
- integrate necessary safeguards into your processing.

AI increases the importance of embedding a data protection by design approach into your organisation’s culture and processes.

Even though AI also brings additional complexities compared to conventional processing, you should demonstrate how you have addressed these, for reasons of accountability and fairness. The AI lifecycle may surface the ‘problem of many hands’. This means that even though a large number of individuals are involved, the role of individuals in isolation may appear small. This makes it complicated to attribute responsibility for potential harms. The more significant the impact your AI system has on individuals whose personal data it processes, the more attention you should give to these complexities, considering what

safeguards are appropriate. You need to have proportionate and adequate safeguards for the processing to be fair.

You should take into account a variety of possible effects when considering impact, from material to non-material harm. This includes emotional consequences such as distress, and any significant economic or social disadvantage.

Further reading

- [Data protection by design and by default](#)
- [How should we approach AI governance and risk management?](#)

How should we approach AI governance and risk management?

2. Data protection impact assessments

DPIAs require you to consider the risks to the rights and freedoms of individuals, including the potential for any significant social or economic disadvantage. The focus is on the potential for harm to individuals or society at large, whether it is physical, material or non-material. This means that DPIAs do not just account for data rights but **rights and freedoms** more broadly and are an integral part of data protection by design and by default. They provide you with an opportunity to assess whether your processing will lead to fair outcomes.

DPIAs also require you to assess and demonstrate how and why:

- your processing is necessary to achieve your purpose; and
- the way you go about your processing is a proportionate way of achieving it.

Documenting the choices you make in the design of your AI system, alongside your assessment of the risks it poses, also helps you demonstrate that your processing is fair. DPIAs help you address AI risks if you see them as not just a box-ticking compliance exercise, but as an ongoing process, subject to regular review.

Further reading

- [When is processing 'necessary'?](#)
- [Data protection impact assessments](#)
- [What do we need to consider when undertaking data protection impact assessments for AI?](#)
- [How should we manage competing interests when assessing AI-related risks?](#)

3. Lawfulness

As noted above, many of the lawful bases for processing require you to assess:

- why the processing is necessary; and
- whether the way you go about it is proportionate to the outcome you seek to achieve.

Merely working through the requirements of a particular lawful basis does not automatically make your processing fair. However, it can go some way to demonstrating how your AI system:


- processes personal data fairly; and
- ensures its impacts on individuals are also fair.

Identifying and meeting the requirements of a lawful basis is likely to reduce the potential of unfair outcomes arising from your processing. This is because the lawful bases themselves provide a level of protection for individuals.

For example, a number of them require you to demonstrate the **necessity** and **proportionality** of your processing. This means that it meets a specific and limited purpose, and there is no reasonable and less intrusive way of achieving the same purpose. This “necessity test” acts to limit the processing, and in turn the potential of unjustified adverse effects on individuals.

This is most obviously the case with the legitimate interests lawful basis. This includes a “three-part test” that requires you to show:

- what the legitimate interest is;
- why the processing is necessary to achieve it; and
- how you balance the rights, freedoms and interests of individuals with your own (or those of third parties).

If you decide to use legitimate interest as your lawful basis, you should record how you assessed this test in your Legitimate interest assessment (LIA). The ICO has produced [a sample LIA template](#)  you can adjust to your context and use to decide whether or not the legitimate interests basis is likely to apply to your processing.

Further reading in this guidance:

- [Lawful basis for processing](#)
- [When is processing ‘necessary’?](#)
- [How do we identify our purposes and lawful basis when using AI?](#)

4. Transparency

Transparency is about being clear, open and honest with individuals from the start about who you are, as well as why and how you want to process their data. This enables them to make an informed choice about whether they wish to enter into a relationship with you, or try and negotiate its terms.

Transparency and fairness are closely linked. Recitals 39 and 60 of the UK GDPR highlight the importance of providing information to individuals to ensure “fair and transparent processing”. You are well-placed to demonstrate transparency, if you are open with people about how your AI system makes decisions about them, and how you use their personal data to train and test the system.

There are different ways you can explain AI decisions. These generally include two main categories:

- process-based explanations. These demonstrate how you have followed good governance processes and

best practices throughout your system’s design and use; and

- outcome-based explanations. These clarify the results of a particular decision. For example, explaining the reasoning in plain, clearly understandable and everyday language.

Outcome-based explanations also cover things such as whether:

- there was meaningful human involvement in the decision; and
- the actual outcome of your AI system’s decision-making meets the criteria you established in your design process.

One type of explanation is the “fairness explanation”. This is about helping people understand the steps you take to ensure your AI decisions are generally unbiased and equitable. Fairness explanations in AI can take four main approaches:

- dataset fairness;
- design fairness;
- outcome fairness; and
- implementation fairness.

Our guidance on explaining decisions made with AI provides more detail on the different ways of explaining AI decisions.

Further reading

[What goes into an explanation?](#)

5. Purpose limitation

Purpose limitation requires you to be clear and open about why you are collecting personal data and that what you intend to do with it is in line with individuals’ reasonable expectations.

Being clear about why you want to process personal data helps you ensure your processing is fair, as well as enabling you to demonstrate your accountability for it. For example, specifying your purpose helps avoid function creep.

If you plan to re-use personal data for a new purpose you usually need to ensure it is compatible with the original. This depends on a number of factors, including your relationship with the individuals. You can also use the data for a new purpose, if you get specific consent for the new purpose or if you have clear obligation set out in law.

Purpose limitation therefore clearly links to fairness, lawfulness and transparency.

AI systems may process personal data for different purposes at different points in the AI lifecycle. However, defining these purposes in advance and determining an appropriate lawful basis for each is a vital part of ensuring you undertake the processing fairly, and that it results in fair outcomes for individuals.

It may make sense to separate particular stages in the AI lifecycle and define the purposes for processing personal data in each one. For example, the data exploration and development phases of an AI system,

including stages such as design and model selection, are distinct from the deployment and ongoing monitoring phases.

Further reading in this guidance

[How do we identify our purposes and lawful basis when using AI?](#)

6. Data minimisation and storage limitation

The data minimisation principle is about only collecting the personal data you need to achieve your purpose. The storage limitation principle is about keeping that data only for as long as you need it.

Both principles relate to fairness as they involve necessity and proportionality considerations. For example:

- you must not process more data than you need just because that data may become useful at some point in the future. This processing is not necessary for your purpose. This raises additional questions about whether it is fair and lawful in the first place, beyond what the impact of processing more data than you need has on individuals; and
- if your AI system keeps data for longer than you need to achieve your purpose, this processing is also unnecessary and therefore unfair. You must take a proportionate approach to retention periods, balancing your needs with the impact of the retention on individuals' privacy.

Clearly establishing what data you need, and how long you need it for, allows you to demonstrate your compliance with these two principles. In turn, this can also demonstrate how your processing is fair overall.

As this guidance notes, data minimisation can appear challenging to achieve in AI. However, the data minimisation principle does not mean your AI system cannot process personal data at all. Instead, it requires you to be clear about what personal data is adequate, relevant and limited, based on your AI system's use case.

Further reading

- [How should we assess security and data minimisation in AI?](#)
- [What data minimisation and privacy-preserving techniques are available for AI systems?](#)

7. Security

The security principle requires you to protect the data you hold from unauthorised or unlawful processing, accidental loss, destruction or damage.

Processing personal data securely is also part of ensuring your processing is fair. Recital 71 of the UK GDPR sets out that the technical and organisational security measures you put in place have to be appropriate to the risks your processing poses to the rights and freedoms of individuals.

Using AI to process any personal data has important implications for your security risk profile, and you need to assess and manage these carefully. For example, the UK GDPR's security requirements apply to both the data you process and also the systems and services you use for that processing.

Further reading

[What security risks does AI introduce?](#)

8. Accountability

Failure to allocate accountability appropriately between processors and controllers can lead to non-compliance with the fairness principle. Accountability gaps can lead to unjustifiably adverse effects for individuals or undermine their ability to exercise rights. For example, the right to contest a solely automated decision with significant effects, which would go against individuals' reasonable expectations. See the section '[What are the accountability and governance implications of AI?](#)' for more detailed guidance on accountability.

9. Accuracy

The personal data you process to train your models or their outputs should be up-to-date and remain as such. Inaccurate input data can lead to inaccurate inferences about individuals that would go against their reasonable expectations or lead to adverse outcomes. For more in-depth analysis of how the accuracy principle applies to AI please read the chapter on [accuracy and statistical accuracy](#).

10. Profiling and automated decision-making (ADM)

Article 22 of the UK GDPR restricts the processing of people's personal data to make solely automated decisions that have legal or similarly significant effects on them, with certain exceptions. Used correctly, profiling and automated decision-making can be useful for many organisations. For example, they help you make decisions fairly and consistently.

The UK GDPR does not prevent you from carrying out profiling or using an AI system to make automated decisions about people, unless those decisions are solely automated and have legal or similarly significant effects on them. In those cases, it places certain safeguards on that processing.

Recital 71 provides more clarity about Article 22. It directly references fairness, saying that you should take into account the "specific circumstances and context" of your processing and implement technical and organisational measures to ensure it is "fair and transparent". These measures should:

- ensure personal data is processed in a manner that takes account of the risks to the rights and interests of individuals; and
- prevent discriminatory effects on the basis of special category data.

From a fairness perspective, profiling and automated decision-making in AI systems may give rise to discrimination. Annex A in this guidance addresses discrimination-specific risks of AI, or what has been referred to as "[algorithmic fairness](#)".

As part of assessing whether your processing is fair, you also need to think about whether any profiling is fundamentally linked with the reason for using your service. As well as what people would reasonably expect you to do with their data.

Further reading

- [What is the impact of Article 22 of the UK GDPR on fairness?](#)
- [What do we need to do about statistical accuracy?](#)
- [What does the UK GDPR say about automated decision-making and profiling?](#)

11. Individual rights

Individuals have a number of rights relating to their personal data. These enable them to exercise control over that data by scrutinising your processing, and in some cases challenging it. For example, by exercising rights such as erasure, restriction and objection).

In AI, these rights apply wherever you use personal data at any point in the AI lifecycle and are important for fairness.

Ensuring you can support these rights requires you to think about the impact your processing has on individuals. For example, if you use solely automated systems to make a decision in an employment context that has a legal or similarly significant effect on individuals, you may need to put in place a process that would enable individuals to challenge a decision, depending on your context.

This is the case even when you are using a partly automated decision-making process that involves AI. Thinking about how you will tell individuals about how your system will use their data to train or contribute through its recommendations to decisions, will also make you think more carefully about its impact and consequences on them. In turn, this can help you ensure that what you tell them is clear and understandable. While this forms part of your obligations under the right to be informed, it can also shape individuals' expectations about what your system does with their data.

Further reading

- [Individual rights](#)
- [How do we ensure individual rights in our AI systems?](#)
- [Explaining decisions made with AI – Legal framework](#)

What about fairness, bias and discrimination?

In detail

- [How do bias and discrimination relate to fairness?](#)
- [How should we address risks of bias and discrimination?](#)
- [Why might an AI system lead to discrimination?](#)
- [Should we just remove all sensitive data?](#)
- [What is the difference between fairness in data protection law and “algorithmic fairness”?](#)
- [What are the technical approaches to mitigate discrimination risk in ML models?](#)
- [Can we process special category data to assess and address discrimination in AI systems?](#)
- [What about special category data, discrimination and automated decision-making?](#)
- [What if we accidentally infer special category data through our use of AI?](#)
- [What can we do to mitigate these risks?](#)
- [Is AI using personal data the best solution to your problem?](#)

How do bias and discrimination relate to fairness?

Fairness in data protection law includes fair treatment and non-discrimination. It is not just about the distribution of benefits and opportunities between members of a group. It is also about how you balance different, competing interests. For example, your own interests and the interests of individuals who are members of that group.

In this guidance we differentiate between bias and discrimination. Bias is an aspect of decision-making. It is a trait often detected not just in AI systems but also humans or institutions. We refer to discrimination as the adverse effects that result from bias. For example, a prejudicial approach in favour of one solution over another.


How should we address risks of bias and discrimination?

As AI systems learn from data which may be unbalanced and/or reflect discrimination, they may produce outputs which have discriminatory effects on people based on their gender, race, age, health, religion, disability, sexual orientation or other characteristics.

The fact that AI systems learn from data does not guarantee that their outputs will not lead to discriminatory effects. The data used to train and test AI systems, as well as the way they are designed, and used, might lead to AI systems which treat certain groups less favourably without objective justification.

The following sections and the [Annex A: Fairness in the AI lifecycle](#) give guidance on interpreting the discrimination-related requirements of data protection law in the context of AI, as well as making some suggestions about best practice.

The following sections **do not** aim to provide guidance on legal compliance with the UK’s anti-discrimination

legal framework, notably the [UK Equality Act 2010](#) . This sits alongside data protection law and applies to a wide range of organisations, both as employers and service providers. It gives individuals protection from direct and indirect discrimination, whether generated by a human or an automated decision-making system (or some combination of the two).

Demonstrating that an AI system is not unlawfully discriminatory under the EA2010 is a complex task, but it is separate and additional to your obligations relating to discrimination under data protection law. Compliance with one will not guarantee compliance with the other.

Data protection law addresses concerns about unjust discrimination in several ways.

First, processing of personal data must be 'fair'. Fairness means you should handle personal data in ways people reasonably expect and not use it in ways that have unjustified adverse effects on them. Any processing of personal data using AI that leads to unjust discrimination between people, will violate the fairness principle.

Second, data protection aims to protect individuals' [rights and freedoms](#)– with regard to the processing of their personal data. This includes the right to privacy but also the right to non-discrimination. Specifically, the requirements of data protection by design and by default mean you have to implement appropriate technical and organisational measures to take into account the risks to the rights and freedoms of data subjects and implement the data protection principles effectively. Similarly, a data protection impact assessment should contain measures to address and mitigate those risks, which include the risk of discrimination.

Third, the UK GDPR specifically notes that processing personal data for profiling and automated decision-making may give rise to discrimination, and that you should use appropriate technical and organisational measures to prevent this.

Further reading outside this guidance

[Artificial Intelligence in public services – Equality and Human Rights Commission](#)

Why might an AI system lead to discrimination?

Before addressing what data protection law requires you to do about the risk of AI and discrimination, and suggesting best practices for compliance, it is helpful to understand how these risks might arise. The following content contains some technical details, so understanding how it may apply to your organisation may require attention of staff in both compliance and technical roles.

Example

A bank develops an AI system to calculate the credit risk of potential customers. The bank will use the AI system to approve or reject loan applications.

The system is trained on a large dataset containing a range of information about previous borrowers,

such as their occupation, income, age, and whether or not they repaid their loan.

During testing, the bank wants to check against any possible gender bias and finds the AI system tends to give women lower credit scores.

In this case, the AI system puts members of a certain group (women) at a disadvantage, and so would appear to be discriminatory. Note that this may not constitute unlawful discrimination under equalities law, if the deployment of the AI system can be shown to be a proportionate means of achieving a legitimate aim.

There are many different reasons why the system may be giving women lower credit scores.

One is **imbalanced training data**. The proportion of different genders in the training data may not be balanced. For example, the training data may include a greater proportion of male borrowers because in the past fewer women applied for loans and therefore the bank doesn't have enough data about women.

Machine learning algorithms used to create an AI system are designed to be the best fit for the data it is trained and tested on. If the men are over-represented in the training data, the model will pay more attention to the statistical relationships that predict repayment rates for men, and less to statistical patterns that predict repayment rates for women, which might be different.

Put another way, because they are **statistically** 'less important', the model may systematically predict lower loan repayment rates for women, even if women in the training dataset were on average more likely to repay their loans than men.

These issues will apply to any population under-represented in the training data. For example, if a facial recognition model is trained on a disproportionate number of faces belonging to a particular ethnicity and gender (eg white men), it will perform better when recognising individuals in that group and worse on others.

Another reason is that **the training data may reflect past discrimination**. For example, if in the past, loan applications from women were rejected more frequently than those from men due to prejudice, then any model based on such training data is likely to reproduce the same pattern of discrimination.

Certain domains where discrimination has historically been a significant problem are more likely to experience this problem more acutely, such as police stop-and-search of young black men, or recruitment for traditionally male roles.

Should we just remove all sensitive data?

Data protection provides additional protections for special category data, while UK equality law is concerned with protected characteristics. Here we use 'sensitive data' as an umbrella term for both groups.

It's important to note that discrimination issues can occur even if the training data does not contain any protected characteristics like gender or race.

This is because a variety of features in the training data are often closely correlated with protected characteristics in non-obvious ways (eg occupation). These "proxy variables" enable the model to reproduce patterns of discrimination associated with those characteristics, even if the designers did not intend this.

Therefore, removing particular attributes to mitigate the risk of discrimination will not necessarily achieve the intended outcome. This approach is sometimes known as “fairness through unawareness”. However, simply removing special category data (or protected characteristics) does not guarantee that other proxy variables cannot essentially reproduce previous patterns.

For example, even if you remove an attribute about gender from a dataset, it may still be possible to infer it from other data that you retain. For example, if more women traditionally work part-time in a sector, a model using working hours to make a recommendation in the context of redundancies may end up discriminating on the basis of gender.

These problems can occur in any statistical model, so the following considerations may apply to you even if you don’t consider your statistical models to be ‘AI’. However, they are more likely to occur in AI systems because they can include a greater number of features and may identify complex combinations of features which are proxies for protected characteristics. Many modern ML methods are more powerful than traditional statistical approaches because they are better at uncovering non-linear patterns in high dimensional data. However, these may also include patterns that reflect discrimination. For example, ML models can pick up [redundant encodings](#) in large datasets and replicate any biases associated with them.

Other causes of potentially discriminatory AI systems include:

- prejudices or bias in the way variables are measured, labelled or aggregated;
- biased cultural assumptions of developers;
- inappropriately defined objectives (eg where the ‘best candidate’ for a job embeds assumptions about gender, race or other characteristics); or
- the way the model is deployed (eg via a user interface which doesn’t meet accessibility requirements).

What is the difference between fairness in data protection law and “algorithmic fairness”?

Computer scientists have been developing mathematical techniques to measure if AI models treat individuals from different groups in potentially discriminatory ways. This field is referred to as “algorithmic fairness”. It reflects a statistical approach to fairness concerned with the distribution of classifications or predictions leading to the real-world allocation of resources, opportunities or capabilities. This is not the same as fairness in data protection which is broader than that, and considers imbalances between affected groups and the stakeholders processing their data.

When deciding what algorithmic fairness metrics you use you must consider legal frameworks relevant to your context, including equality law.

Statistical approaches can be useful in identifying discriminatory impacts. But they are not likely to guarantee your system complies with fairness or explaining why and how any unfairness takes place, even less what mitigation measures are efficient. This is because they cannot fully capture the social, historical and political nuances of each use case that relate to how, where, why personal data was processed.

Example

An organisation uses algorithmic fairness metrics to evaluate whether a system has shortlisted a

disproportionate number of women to men for a specific job. The metrics do not address more substantive elements such as the terms of employment or the suitability of the candidates.

As a result, you should view algorithmic fairness metrics as part of a broader non-technical framework that you need to put in place.

Further reading outside this guidance

[Fairness definitions explained](#)

[Algorithmic fairness](#) metrics and relevant toolkits may assist you in identifying and mitigating risks of unfair outcomes. However, fairness is not a goal that algorithms can achieve alone. Therefore, you should take a holistic approach, thinking about fairness across different dimensions and not just within the bounds of your model or statistical distributions.

You should think about:

- the power and information imbalance between you and individuals whose personal data you process;
- the underlying structures and dynamics of the environment your AI will be deployed in;
- the implications of creating self-reinforcing feedback loops;
- the nature and scale of any potential harms to individuals resulting from the processing of their data; and
- how you will make well-informed decisions based on rationality and causality rather than mere correlation.

In general, you should bear in mind the following:

Statistical approaches are just one piece of the puzzle: You need to take a broader approach to fairness. This is because vital elements are not captured by algorithmic fairness metrics, such as governance structures or legal requirements. Additionally, it may be difficult (and in some cases, misguided) to mathematically measure and remove bias that may be encoded in various features of your model.

Context is key: The conditions under which decision-making takes place is equally important as the decision-making process itself.

Fairness in terms of data protection in the context of AI is not static: AI-driven or supported decisions can be consequential, changing the world they are applied in, and potentially creating risks for cumulative discrimination.

The root causes are important: AI should not distract your decision-makers from addressing the root causes of unfairness that AI systems may detect and replicate.

Patterns are not destiny: AI models do not just memorise but seek to replicate patterns. The decisions they give rise to will influence the status quo, which in turn will impact the input data that inform future

predictions. Without thoughtful adoption, AI can lead to a vicious cycle where past patterns of unfairness are replicated and entrenched

What are the technical approaches to mitigate discrimination risk in ML models?

While discrimination is a broader problem that cannot realistically be 'fixed' through technology, various approaches exist which aim to mitigate AI-driven discrimination.

As explained above, some of these involve algorithmic fairness. This is a field of different mathematical techniques to measure how AI models treat individuals from different groups in potentially discriminatory ways and reduce them.

The techniques it proposes do not necessarily align with relevant non-discrimination law in the UK, and in some cases may contradict it, so should not be relied upon as a means of complying with such obligations. However, depending on your context, some of these approaches may be appropriate technical measures to ensure personal data processing is fair and to minimise the risks of discrimination arising from it.

In cases of **imbalanced training data**, it may be possible to balance it out by adding or removing data about under/ overrepresented subsets of the population (eg adding more data points on loan applications from women). This is part of pre-processing techniques.

In cases where the **training data reflects past discrimination**, you could either modify the data, change the learning process, or modify the model after training. These are part of in-processing and post-processing techniques.

In order to measure whether these techniques are effective, there are various mathematical 'fairness' measures against which you can measure the results.

Simply removing any protected characteristics from the inputs the model uses to make a prediction is unlikely to be enough, as there are often variables which are proxies for the protected characteristics. Other measures involve comparing how the AI system distributes positive or negative outcomes (or errors) between protected groups. Some of these measures conflict with each other, meaning you cannot satisfy all of them at the same time. Which of these measures are most appropriate, and in what combinations, if any, will depend on your context, as well as any applicable relevant laws (eg equality law).

You should also consider the impact of these techniques on the statistical accuracy of the AI system's performance. For example, to reduce the potential for discrimination, you might modify a credit risk model so that the proportion of positive predictions between people with different protected characteristics (eg men and women) are equalised. This may help prevent discriminatory outcomes, but it could also result in a higher number of statistical errors overall which you will also need to manage as well.

In practice, there may not always be a tension between statistical accuracy and avoiding discrimination. For example, if discriminatory outcomes in the model are driven by a relative lack of data about a statistically small minority of the population, then statistical accuracy of the model could be increased by collecting more data about them, whilst also equalising the proportions of correct predictions.

However, in that case, you would face a different choice between:

- collecting more data on the minority population in the interests of reducing the disproportionate number of statistical errors they face; or

- not collecting such data due to the risks doing so may pose to the other rights and freedoms of those individuals.

Unfairness and discrimination is not limited to impacts on groups for which there is Equality Act protection, but consideration should be given to whether the use of AI may result in unfair outcomes for other groups as well. Therefore, you must think about how you can protect minorities or vulnerable populations, while addressing risks of exacerbating pre-existing power imbalances. You should balance your bias mitigation goals with your data minimisation obligations. For example, if you can show that additional data is genuinely useful to protect minorities, then it is likely to be appropriate to process that additional data.

Can we process special category data to assess and address discrimination in AI systems?

In order to assess and address the risks of discrimination in your AI system, you may need a dataset containing data about individuals that includes:

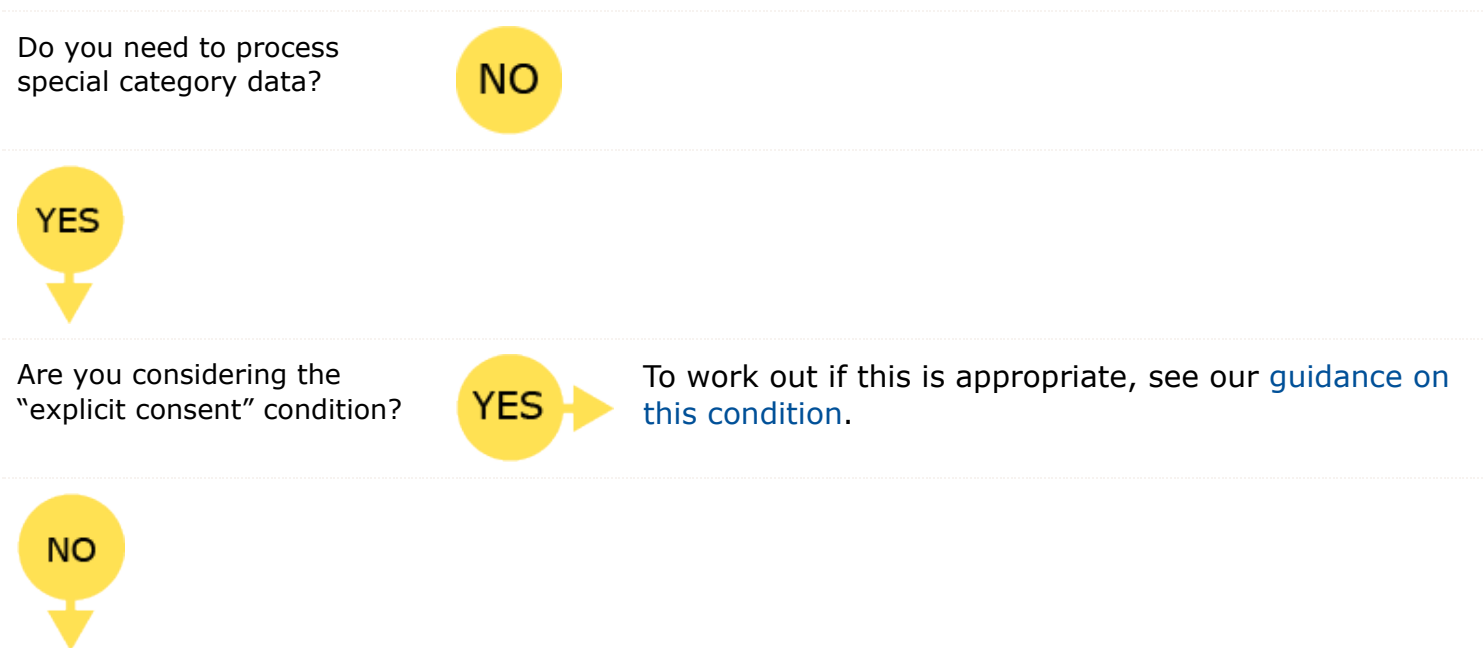
- special category data under data protection law; and/or
- protected characteristics such as those outlined in the Equality Act 2010.

For example, you could use the dataset to test how your system performs with different groups, and also potentially to re-train your model to avoid discriminatory effects.

If your processing for this purpose involves special category data, then in addition to having a lawful basis under Article 6 of the UK GDPR you must meet one of the conditions in Article 9. Some of these also require authorisation by law or a basis in law, which can be found in Schedule 1 of the DPA 2018.

There is no single condition in Article 9 that is specifically about the purpose of assessing and addressing discrimination in AI systems. This means that which, if any, of these conditions are appropriate depends on your individual circumstances.

You can use the following diagram to understand what each condition requires. It has links to more information about each one.



Does the Article 9 condition say that it requires [authorisation by law or a basis in law](#)?



This means you are considering:

- [vital interests](#)
- [not-for-profit bodies](#)
- [made public by the data subject](#)
- [legal claims and judicial acts](#)

To work out if any of these are appropriate, see our guidance. You don't need a DPA Schedule 1 condition, or an [appropriate policy document](#).



This means you are considering:

- employment, social security and social protection
- health and social care
- public health
- archiving, research and statistics
- substantial public interest

Are you considering the "employment, social security and social protection" condition?



To work out if this is appropriate, see our [guidance on this condition](#).

The relevant legal authorisation is set out in the DPA, at Schedule 1 condition 1. You need an [appropriate policy document](#).



Are you considering the "health and social care" condition?



To work out if this is appropriate, see our [guidance on this condition](#).

The relevant basis in UK law is set out in the DPA, at Schedule 1 condition 2. You don't need an [appropriate policy document](#).



Are you considering the “public health” condition?



To work out if this is appropriate, see our [guidance on this condition](#).

The relevant basis in UK law is set out in the DPA, at Schedule 1 condition 3. You don’t need an [appropriate policy document](#).



Are you considering the “archiving, research and statistics” condition?



To work out if this is appropriate, see our [guidance on this condition](#).

The relevant basis in UK law is set out in the DPA, at Schedule 1 condition 4. You don’t need an [appropriate policy document](#).



Are you considering the “substantial public interest” condition?



To work out if this is appropriate, see our [guidance on this condition](#).

The relevant basis in UK law is set out in the DPA, at Section 10(3). You need to meet one of the 23 specific [substantial public interest conditions](#) in Schedule 1 (at paragraphs 6 to 28). In almost all cases, you must also have an [appropriate policy document](#).

[An accessible, written description of this diagram \(suitable for screen readers\) is available here.](#)

Example: using special category data to assess discrimination in AI, to identify and promote or maintain equality of opportunity

An organisation using a CV scoring AI system to assist with recruitment decisions needs to test whether its system might be discriminating by religious or philosophical beliefs. While the system does not

directly use information about the applicants' religion, there might be features in the system which are indirect proxies for religion, such as previous occupation or qualifications. In a labour market where certain religious groups have been historically excluded from particular professions, a CV scoring system may unfairly under-rate candidates on the basis of those proxies.

The organisation collects the religious beliefs of a sample of job applicants in order to assess whether the system is indeed producing disproportionately negative outcomes or erroneous predictions for applicants with particular religious beliefs.

The organisation relies on the substantial public interest condition in Article 9(2)(g), and the equality of opportunity or treatment condition in Schedule 1 (8) of the DPA 2018. This provision can be used to identify or keep under review the existence or absence of equality of opportunity or treatment between certain protected groups, with a view to enabling such equality to be promoted or maintained.

Example: using special category data to assess discrimination in AI, for research purposes

A university researcher is investigating whether facial recognition systems perform differently on the faces of people of different racial or ethnic origin, as part of a research project.

In order to do this, the researcher assigns racial labels to an existing dataset of faces that the system will be tested on, thereby processing special category data. They rely on the archiving, research and statistics condition in Article 9(2)(j), read with Schedule 1 paragraph 4 of the DPA 2018.

Is special category data in data protection law the same as protected characteristics under the Equality Act?

Not in all cases. Some of the protected characteristics outlined in the Equality Act **are** classified as special category data. For example, race, religion or belief, and sexual orientation.

Other protected characteristics aren't. For example, testing for discriminatory impact by age does not involve special category data, even though age is a protected characteristic. In contrast, testing for discriminatory impact by ethnic origin does involve special category data.

You also need to be aware that some protected characteristics may constitute special category data even when the link is not obvious. For example, disability, pregnancy, and gender reassignment may be special category data in so far as they concern information about a person's health. Similarly, because civil partnerships were until recently only available to same-sex couples, data that indicates someone is in a civil partnership may indirectly reveal their sexual orientation.

You should take this into account, as there are different data protection considerations depending on the kinds of discrimination you are testing for.

You can see the overlap of special category data and protected characteristics in Table 1.

Table 1.

Protected characteristics in the Equality Act 2010	Special category data in UK data protection
<ul style="list-style-type: none">• race• religion or belief• sexual orientation	<ul style="list-style-type: none">• racial or ethnic origin• religious or philosophical beliefs• sexual orientation
<ul style="list-style-type: none">• age• disability• gender reassignment• marriage and civil partnership• pregnancy and maternity• sex	<ul style="list-style-type: none">• political opinions• trade union membership• genetic data• biometric data (where used for identification purposes)• health• sex life

What else do we need to consider?

You should also note that when you are processing personal data that results from specific technical processing about the physical, physiological or behavioural characteristics of an individual, and allows or confirms that individual’s unique identification, that data is biometric data.

Where you use biometric data for the **purpose** of uniquely identifying an individual, it is also special category data.

So, if you use biometric data for testing and mitigating discrimination in your AI system, but not for the purpose of confirming the identity of the individuals within the dataset or making any kind of decision in relation to them, the biometric data may not come under Article 9. The data is still regarded as biometric data under the UK GDPR, but may not be special category data.

Similarly, if the personal data does not allow or confirm an individual’s unique identification, then it is not biometric data (or special category data).

Additionally, even when you are not processing data classified as special category data in data protection law, you still need to consider:

- the broader questions of lawfulness, fairness and the risks the processing poses as a whole; and
- the possibility for the data to either be special category data anyway, or becoming so during the processing (ie if the processing involves analysing or inferring any data to do with health or genetic status).

Finally, if the personal data you are using to assess and improve potentially discriminatory AI were originally processed for a different purpose, you should consider:

- whether your new purpose is compatible with the original purpose;

- how you will obtain fresh consent, if required. For example, if the data was initially collected on the basis of consent, even if the new purpose is compatible you still need to collect a fresh consent for the new purpose; and
- if the new purpose is incompatible, how you will ask for consent.

Further Reading

 [See Article 9 and Recitals 51 to 56 of the UK GDPR](#) 

External link

 [See Schedule 1 of the DPA 2018](#) 

External link

Further reading outside this guidance

Read our guidance on [purpose limitation](#) and [special category data](#).

What about special category data, discrimination and automated decision-making?

Using special category data to assess the potential discriminatory impacts of AI systems does not usually constitute automated decision-making under data protection law. This is because it does not involve directly making any decisions about individuals.

Similarly, re-training a discriminatory model with data from a more diverse population to reduce its discriminatory effects does not involve directly making decisions about individuals and is therefore not classed as a decision with legal or similarly significant effect.

However, in some cases, simply re-training the AI model with a more diverse training set may not be enough to sufficiently mitigate its discriminatory impact. Rather than trying to make a model fair by **ignoring** protected characteristics when making a prediction, some approaches directly **include** such characteristics when making a classification, to ensure members of potentially disadvantaged groups are protected. Including protected characteristics could one of the measures you take to comply with the requirement to make 'reasonable adjustments' under the Equality Act 2010.

For example, if you were using an AI system to assist with sorting job applicants, rather than attempting to create a model which ignores a person's disability, it may be more effective to include their disability status in order to ensure the system does not indirectly discriminate against them. Not including disability status as an input to the automated decision could mean the system is more likely to indirectly discriminate against people with a disability because it will not factor in the effect of their condition on other features used to make a prediction.

However, if you process disability status using an AI system to make decisions about individuals, which produce legal or similarly significant effects on them, you must have explicit consent from the individual, or be able to meet one of the substantial public interest conditions laid out in Schedule 1 of the DPA.

You need to carefully assess which conditions in Schedule 1 may apply. For example, the equality of

opportunity monitoring provision mentioned above cannot be relied on in such contexts, because the processing is carried out for the purposes of decisions about a particular individual. Therefore, such approaches will only be lawful if based on a different substantial public interest condition in Schedule 1.

What if we accidentally infer special category data through our use of AI?

There are many contexts in which non-protected characteristics, such as the postcode you live in, are proxies for a protected characteristic, like race. Recent advances in machine learning, such as 'deep' learning, have made it even easier for AI systems to detect patterns in the world that are reflected in seemingly unrelated data. Unfortunately, this also includes detecting patterns of discrimination using complex combinations of features which might be correlated with protected characteristics in non-obvious ways.

For example, an AI system used to score job applications to assist a human decision-maker with recruitment decisions might be trained on examples of previously successful candidates. The information contained in the application itself may not include protected characteristics like race, disability, or mental health.

However, if the examples of employees used to train the model were discriminated against on those grounds (eg by being systematically under-rated in performance reviews), the algorithm may learn to reproduce that discrimination by inferring those characteristics from proxy data contained in the job application, despite the designer never intending it to.

So, even if you don't use protected characteristics in your model, it is very possible that you may inadvertently use a model which has detected patterns of discrimination based on those protected characteristics and is reproducing them in its outputs. As described above, some of those protected characteristics are also special category data.

Special category data is defined as personal data that 'reveals or concerns' the special categories. If the model learns to use particular combinations of features that are sufficiently revealing of a special category, then the model may be processing special category data.

As stated in our guidance on special category data, if you use profiling with the **intention** of inferring special category data, then this is special category data irrespective of whether the inferences are incorrect.

Furthermore, for the reasons stated above, there may also be situations where your model infers special category as an intermediate step to another (non-special-category data) inference. You may not be able to tell if your model is doing this just by looking at the data that went into the model and the outputs that it produces. It may do so with high statistical accuracy, even though you did not intend for it to do so.

If you are using machine learning with personal data you should proactively assess the chances that your model might be inferring protected characteristics or special category data or both in order to make predictions, and actively monitor this possibility throughout the lifecycle of the system. If your system is indeed inferring special category or criminal conviction data (whether unintentional or not), you must have an appropriate Article 9 or 10 condition for processing. If it is unclear whether or not your system may be inferring such data, you may want to identify a condition to cover that possibility and reduce your compliance risk, although this is not a legal requirement.

As noted above, if you are using such a model to make legal or similarly significant decisions in a solely

automated way, this is only lawful if you have the person's consent or you meet the substantial public interest condition (and an appropriate provision in Schedule 1).

Further reading outside this guidance

Read our guidance on [special category data](#).

What can we do to mitigate these risks?

The most appropriate approach to managing the risk of discriminatory outcomes in ML systems will depend on the particular domain and context you are operating in.

You should determine and document your approach to bias and discrimination mitigation from the very beginning of any AI application lifecycle, so that you can take into account and put in place the appropriate safeguards and technical measures during the design and build phase. Annex A has more information about good practice steps for mitigating bias and discrimination.

Establishing clear policies and good practices for the procurement and lawful processing of high-quality training and test data is important, especially if you do not have enough data internally. Whether procured internally or externally, you should satisfy yourself that the data is representative of the population you apply the ML system to (although for reasons stated above, this will not be sufficient to ensure fairness). For example, for a high street bank operating in the UK, the training data could be compared against the most recent Census.

Your senior management should be responsible for signing-off the chosen approach to manage discrimination risk and be accountable for its compliance with data protection law. While they are able to leverage expertise from technology leads and other internal or external subject matter experts, to be accountable your senior leaders still need to have a sufficient understanding of the limitations and advantages of the different approaches. This is also true for DPOs and senior staff in oversight functions, as they will be expected to provide ongoing advice and guidance on the appropriateness of any measures and safeguards put in place to mitigate discrimination risk.

In many cases, choosing between different risk management approaches will require trade-offs. This includes choosing between safeguards for different protected characteristics and groups. You need to document and justify the approach you choose.

Trade-offs driven by technical approaches are not always obvious to non-technical staff so data scientists should highlight and explain these proactively to business owners, as well as to staff with responsibility for risk management and data protection compliance. Your technical leads should also be proactive in seeking domain-specific knowledge, including known proxies for protected characteristics, to inform algorithmic 'fairness' approaches.

You should undertake robust testing of any anti-discrimination measures and should monitor your ML system's performance on an ongoing basis. Your risk management policies should clearly set out both the process, and the person responsible, for the final validation of an ML system both before deployment and, where appropriate, after an update.

For discrimination monitoring purposes, your organisational policies should set out any variance tolerances

against the selected Key Performance Metrics, as well as escalation and variance investigation procedures. You should also clearly set variance limits above which the ML system should stop being used.

If you are replacing traditional decision-making systems with AI, you should consider running both concurrently for a period of time. You should investigate any significant difference in the type of decisions (eg loan acceptance or rejection) for different protected groups between the two systems, and any differences in how the AI system was predicted to perform and how it does in practice.

Beyond the requirements of data protection law, a diverse workforce is a powerful tool in identifying and managing bias and discrimination in AI systems, and in the organisation more generally.

Finally, this is an area where best practice and technical approaches continue to develop. As a result, we will keep this guidance (and Annex A) under review as these approaches mature and evolve.

You should invest the time and resources to ensure you continue to follow best practice and your staff remain appropriately trained on an ongoing basis. In some cases, AI may actually provide an opportunity to uncover and address existing discrimination in traditional decision-making processes and allow you to address any underlying discriminatory practices.

Further reading inside this guidance

See our guidance on [‘How should we manage competing interests when assessing AI-related risks?’](#)

Further reading outside this guidance

- [UK Equality Act 2010](#)
- [European Charter of Fundamental Rights](#)

Is AI using personal data the best solution to your problem?

It is useful to reflect on whether an AI system is the most appropriate solution to the problem you are trying to solve in the first place. Some problems are solved by being open to unpredictability, rather than identifying patterns of the past and assuming continuity. At least that may be the individual’s reasonable expectation based on your context.

AI systems follow predetermined rules and are often unable to adapt to novel or edge cases that don’t neatly map onto input data they have been trained on. In these cases, you should carefully evaluate whether it is appropriate to integrate an AI system into a decision-making process. Certain problems are not ones that algorithms can interpret and solve, especially if they require individualised human discretion or are particularly contested. In such cases, there may be more merit in AI playing a decision-support role. For example, providing context for a human decision-maker rather than making the decision itself or influencing it in a meaningful way.

As we explain in the Fairness in the AI lifecycle section, the choices you make about formulating a problem can have profound consequences in terms of the impacts your system has on individuals and wider society.

You need to think about the underlying social dynamics and nuances, such as structural discrimination. In general, you need to think whether your AI system is doing the right thing, not just if it is correctly doing what you asked it to do.

What is the impact of Article 22 of the UK GDPR on fairness?

In detail

- [What is the purpose of Article 22?](#)
- [What do we need to ask ourselves to understand whether our processing falls under Article 22?](#)
- [Will impacted individuals be able to contest an automated decision?](#)

What is the purpose of Article 22?

Data protection law applies to all automated individual decision-making and profiling. Article 22 of the UK GDPR has additional rules to protect individuals if you are carrying out solely automated decision-making that has legal or similarly significant effects on them.

This may apply in the AI context, eg where you are using an AI system to make these kinds of decisions. You may decide to use automated decision-making in order to achieve scale at speed, or reduce costs by employing fewer humans in a decision-making process.

However, you can only carry out this type of decision-making where the decision is:

- necessary for the entry into or performance of a contract;
- authorised by law that applies to you; or
- based on the individual's explicit consent.

You therefore have to identify if your processing falls under Article 22 and, where it does, make sure that you:

- give individuals information about the processing;
- introduce simple ways for them to request human intervention or challenge a decision; and
- carry out regular checks to make sure your systems are working as intended.

In summary, Article 22 requirements essentially ensure that processing is fair by seeking to protect data subjects from solely automated decision making, save in circumstances where domestic law authorises it and provides suitable safeguards, or where it would otherwise be fair as the data subject has explicitly consented to it or it is necessary for the entry into or performance of a contract between the data subject and the controller.

What do we need to ask ourselves to understand whether our processing falls under Article 22?

AI systems can play a wide variety of roles, from decision-support, to triaging, to classifying or retrieving information. This means they can be involved at different stages of your decision-making process and to different degrees.

When an AI system is involved in a decision that impacts individuals in a legal or similarly significant way,

you must ask:

- what kind of decision is it (ie is it solely automated)?;
- when does the decision take place?;
- what is the context in which the system makes the decision?; and
- what steps are involved in reaching it?

This will help you comply with data protection and Article 22 in particular.

A legal effect is something that affects someone's legal rights. For example, someone's entitlement to child or housing benefit. A similarly significant effect is more difficult to define but has the same sort of impact on someone's circumstances or choices. For example, a computer decision to offer someone a job, or a decision to agree or decline a person's mortgage application. These effects can be positive or negative.

Understanding the full context in which a decision takes place will help you identify:

- whether Article 22 applies; and
- if it does, what you need to do to comply with it in your own context.

To do this, you should ask yourself:

1. What is the actual decision?

It is crucial that you identify which step of the decision-making process produces or overwhelmingly determines the direct legal or similarly significant effects. This ensures you have clarity on the level of human agency over the final outcome.

2. When does the human-determined decision take place?

In general, mere human involvement in the AI lifecycle does not necessarily make the decision 'AI-assisted', nor does it qualify as meaningful human review. The sequencing of the human and AI factors is crucial.

In some cases for example, a human may provide input data into an AI system, that will then process it to make predictions or classifications.

If those outputs have significant or legal effects, Article 22 will apply because the decision itself is solely automated. The human's involvement **in the decision** is not meaningful, as they are merely supplying the data that the system uses to make that decision.

In most cases, for human review to be meaningful, human involvement should come after the automated decision has taken place and it must relate to the actual outcome.

3. What is the context of the decision?

Identifying the structures, assumptions and conditions in which the decision takes place will help you have a clearer idea of the impact your system has. You should include the following considerations about:

- how your AI system interacts with human reviewers;
- the decision-making options that your system's design or introduction creates or prevents;
- the links between the AI outputs and the actual impact on groups and individuals; and

- what, if any, human-led processes your AI system intends to replace.

Will impacted individuals be able to contest an automated decision?

For the processing to be fair and compliant with Article 22, individuals must be able to contest a decision in a timely manner. Meaningful transparency is fundamental to support this and you must put in place the appropriate measures to ensure individuals can exercise their rights. See the section on '[What steps should we take to fulfil rights related to automated decision-making?](#)' for more information.

Supplementary reading in this guidance

- [How do we ensure individual rights in our AI systems?](#)
- [How do we ensure individual rights relating to solely automated decisions with legal or similar effect?](#)
- [What is the role of human oversight?](#)

Further reading in other ICO guidance

[Automated decision-making and profiling](#)

How should we assess security and data minimisation in AI?

At a glance

This section explains 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. A number of techniques are presented to help both data minimisation and effective AI development and deployment

Who is this section for?

This section is aimed at technical specialists, who are best placed to assess the security of an AI system and what personal data is required. It will also be useful for those in compliance-focused roles to understand the risks associated with security and data minimisation in AI.

In detail

- [What security risks does AI introduce?](#)
- [What types of privacy attacks apply to AI models?](#)
- [What steps should we take to manage the risks of privacy attacks on AI models?](#)
- [What data minimisation and privacy-preserving techniques are available for AI systems?](#)

What security risks does AI introduce?

You must process personal data in a manner that ensures appropriate levels of security against its unauthorised or unlawful processing, accidental loss, destruction or damage. In this section we focus on the way AI can adversely affect security by making known risks worse and more challenging to control.

What are our security requirements?

There is no 'one-size-fits-all' approach to security. The appropriate security measures you should adopt depend on the level and type of risks that arise from specific processing activities.

Using AI to process any personal data has important implications for your security risk profile, and you need to assess and manage these carefully.

Some implications may be triggered by the introduction of new types of risks, eg adversarial attacks on machine learning models (see section [‘What types of privacy attacks apply to AI models?’](#)).

Further reading outside this guidance

Read our [guidance on security](#) in the Guide to the UK GDPR, and the [ICO/NCSC Security Outcomes](#), for general information about security under data protection law.

Information security is a key component of our AI auditing framework but is also central to our work as the information rights regulator. The ICO is planning to expand its general security guidance to take into account the additional requirements set out in the new UK GDPR.

While this guidance will not be AI-specific, it will cover a range of topics that are relevant for organisations using AI, including software supply chain security and increasing use of open-source software.

What is different about security in AI compared to 'traditional' technologies?

Some of the unique characteristics of AI mean compliance with data protection law's security requirements can be more challenging than with other, more established technologies, both from a technological and human perspective.

From a technological perspective, AI systems introduce new kinds of complexity not found in more traditional IT systems that you may be used to using. Depending on the circumstances, your use of AI systems is also likely to rely heavily on third party code relationships with suppliers, or both. Also, your existing systems need to be integrated with several other new and existing IT components, which are also intricately connected. Since AI systems operate as part of a larger chain of software components, data flows, organisational workflows and business processes, you should take a holistic approach to security. This complexity may make it more difficult to identify and manage some security risks, and may increase others, such as the risk of outages.

From a human perspective, the people involved in building and deploying AI systems are likely to have a wider range of backgrounds than usual, including traditional software engineering, systems administration, data scientists, statisticians, as well as domain experts.

Security practices and expectations may vary significantly, and for some there may be less understanding of broader security compliance requirements, as well as those of data protection law more specifically. Security of personal data may not always have been a key priority, especially if someone was previously building AI applications with non-personal data or in a research capacity.

Further complications arise because common practices about how to process personal data securely in data science and AI engineering are still under development. As part of your compliance with the security principle, you should ensure that you actively monitor and take into account the state-of-the-art security practices when using personal data in an AI context.

It is not possible to list all known security risks that might be exacerbated when you use AI to process personal data. The impact of AI on security depends on:

- the way the technology is built and deployed;
- the complexity of the organisation deploying it;
- the strength and maturity of the existing risk management capabilities; and
- the nature, scope, context and purposes of the processing of personal data by the AI system, and the risks posed to individuals as a result.

The following hypothetical scenarios are intended to raise awareness of some of the known security risks and challenges that AI can exacerbate. The following content contains some technical details, so

understanding how it may apply to your organisation may require attention of staff in both compliance and technical roles.

Our key message is that you should review your risk management practices ensuring personal data is secure in an AI context.

How should we ensure training data is secure?

ML systems require large sets of training and testing data to be copied and imported from their original context of processing, shared and stored in a variety of formats and places, including with third parties. This can make them more difficult to keep track of and manage.

Your technical teams should record and document all movements and storing of personal data from one location to another. This will help you apply appropriate security risk controls and monitor their effectiveness. Clear audit trails are also necessary to satisfy accountability and documentation requirements.

In addition, you should delete any intermediate files containing personal data as soon as they are no longer required, eg compressed versions of files created to transfer data between systems.

Depending on the likelihood and severity of the risk to individuals, you may also need to apply de-identification techniques to training data before it is extracted from its source and shared internally or externally.

For example, you may need to remove certain features from the data, or apply privacy enhancing technologies (PETs), before sharing it with another organisation.

How should we ensure security of externally maintained software used to build AI systems?

Very few organisations build AI systems entirely in-house. In most cases, the design, building, and running of AI systems will be provided, at least in part, by third parties that you may not always have a contractual relationship with.

Even if you hire your own ML engineers, you may still rely significantly on third-party frameworks and code libraries. Many of the most popular ML development frameworks are [open source](#).

Using third-party and open source code is a valid option. Developing all software components of an AI system from scratch requires a large investment of time and resources that many organisations cannot afford, and especially compared to open source tools, would not benefit from the rich ecosystem of contributors and services built up around existing frameworks.

However, one important drawback is that these standard ML frameworks often depend on other pieces of software being already installed on an IT system. To give a sense of the risks involved, a recent [study](#) found the most popular ML development frameworks include up to 887,000 lines of code and rely on 137 external dependencies. Therefore, implementing AI will require changes to an organisation's software stack (and possibly hardware) that may introduce additional security risks.

Example

The recruiter hires an ML engineer to build the automated CV filtering system using a Python-based ML framework. The ML framework depends on a number of specialist open-source programming libraries, which needed to be downloaded on the recruiter's IT system.

One of these libraries contains a software function to convert the raw training data into the format required to train the ML model. It is later discovered the function has a security vulnerability. Due to an unsafe default configuration, an attacker introduced and executed malicious code remotely on the system by disguising it as training data.

This is not a far-fetched example, in January of 2019, such a [vulnerability](#) was discovered in 'NumPy', a popular library for the Python programming language used by many machine learning developers.

What should we do in this situation?

Whether AI systems are built in-house, externally, or a combination of both, you will need to assess them for security risks. As well as ensuring the security of any code developed in-house, you need to assess the security of any externally maintained code and frameworks.

In many respects, the standard requirements for maintaining code and managing security risks will apply to AI applications. For example:

- your external code security measures should include subscribing to security advisories to be notified of vulnerabilities; or
- your internal code security measures should include adhering to coding standards and instituting source code review processes.

Whatever your approach, you should ensure that your staff have appropriate skills and knowledge to address these security risks.

Having a secure pipeline from development to deployment will further mitigate security risks associated with third party code by separating the ML development environment from the rest of your IT infrastructure where possible. Using '[virtual machines](#)' or '[containers](#)' - emulations of a computer system that run inside, but isolated from the rest of the IT system may help here; these can be pre-configured specifically for ML tasks. In addition, it is possible to train an ML model using a programming language and framework suitable for exploratory development, but then convert the model into another more secure format for deployment.

Further reading outside this guidance

Read our report on [Protecting personal data in online services: learning from the mistakes of others](#) (PDF) for more information. Although written in 2014, the report's content in this area may still assist you.

The ICO is developing further security guidance, which will include additional recommendations for the oversight and review of externally maintained source code from a data protection perspective, as well as its implications for security and data protection by design.

What types of privacy attacks apply to AI models?

The personal data of the people who an AI system was trained on might be inadvertently revealed by the outputs of the system itself.

It is normally assumed that the personal data of the individuals whose data was used to train an AI system cannot be inferred by simply observing the predictions the system returns in response to new inputs. However, new types of privacy attacks on ML models suggest that this is sometimes possible.

In this section, we focus on two kinds of these privacy attacks – ‘model inversion’ and ‘membership inference’.

What are model inversion attacks?

In a model inversion attack, if attackers already have access to some personal data belonging to specific individuals included in the training data, they can infer further personal information about those same individuals by observing the inputs and outputs of the ML model. The information attackers can learn about goes beyond generic inferences about individuals with similar characteristics.

Example one – model inversion attack

An early [demonstration](#) of this kind of attack concerned a medical model designed to predict the correct dosage of an anticoagulant, using patient data including genetic biomarkers. It proved that an attacker with access to some demographic information about the individuals included in the training data could infer their genetic biomarkers from the model, despite not having access to the underlying training data.

Further reading outside this guidance

For further details of a model inversion attack, see [‘Algorithms that remember: model inversion attacks and data protection law’](#)

Example two – model inversion attack

Another recent [example](#) demonstrates that attackers could reconstruct images of faces that a Facial Recognition Technology (FRT) system has been trained to recognise. FRT systems are often designed to allow third parties to query the model. When the model is given the image of a person whose face it

recognises, the model returns its best guess as to the name of the person, and the associated confidence rate.

Attackers could probe the model by submitting many different, randomly generated face images. By observing the names and the confidence scores returned by the model, they could reconstruct the face images associated with the individuals included in the training data. While the reconstructed face images were imperfect, researchers found that they could be matched (by human reviewers) to the individuals in the training data with 95% accuracy (see Figure 2.)



Figure 2. A face image recovered using model inversion attack (left) and corresponding training set image (right), from Fredriksen et al., [‘Model Inversion Attacks that Exploit Confidence Information’](#).

What are membership inference attacks?

Membership inference attacks allow malicious actors to deduce whether a given individual was present in the training data of a ML model. However, unlike in model inversion, they don’t necessarily learn any additional personal data about the individual.

For example, if hospital records are used to train a model which predicts when a patient will be discharged, attackers could use that model in combination with other data about a particular individual (that they already have) to work out if they were part of the training data. This would not reveal any individual’s data from the training data set itself, but in practice it would reveal that they had visited one of the hospitals that generated the training data during the period the data was collected.

Similar to the earlier FRT example, membership inference attacks can exploit confidence scores provided alongside a model’s prediction. If an individual was in the training data, then the model will be disproportionately confident in a prediction about that person because it has seen them before. This allows the attacker to infer that the person was in the training data.

The gravity of the consequences of models’ vulnerability to membership inference will depend on how sensitive or revealing membership might be. If a model is trained on a large number of people drawn from the general population, then membership inference attacks pose less risk. But if the model is trained on a vulnerable or sensitive population (eg patients with dementia, or HIV), then merely revealing that someone is part of that population may be a serious privacy risk.

What are black box and white box attacks?

There is an important distinction between 'black box' and 'white box' attacks on models. These two approaches correspond to different operational models.

In white box attacks, the attacker has complete access to the model itself, and can inspect its underlying code and properties (although not the training data). For example, some AI providers give third parties an entire pre-trained model and allow them to run it locally. White box attacks enable additional information to be gathered, such as the type of model and parameters used, which could help an attacker in inferring personal data from the model.

In black box attacks, the attacker only has the ability to query the model and observe the relationships between inputs and outputs. For example, many AI providers enable third parties to access the functionality of an ML model online to send queries containing input data and receive the model's response. The examples we have highlighted above are both black box attacks.

White and black box attacks can be performed by providers' customers or anyone else with either authorised or unauthorised access to either the model itself, or its query or response functionality.

What about models that include training data by design?

Model inversion and membership inferences show that AI models can inadvertently contain personal data. You should also note that there are certain kinds of ML models which actually contain parts of the training data in its raw form within them **by design**. For example, '[support vector machines](#)' (SVMs) and '[k-nearest neighbours](#)' (KNN) models contain some of the training data in the model itself.

In these cases, if the training data is personal data, access to the model by itself means that the organisation purchasing the model will already have access to a subset of the personal data contained in the training data, without having to exert any further efforts. Providers of such ML models, and any third parties procuring them, should be aware that they may contain personal data in this way.

Unlike model inversion and membership inference, personal data contained in models like this is not an attack vector. Any personal data contained in these models would be there by design and easily retrievable by the third party. Storing and using these models therefore constitutes processing of personal data and as such, the standard data protection provisions apply.

Further reading outside this guidance

See scikit learn's [module on 'Support Vector Machines'](#).

See scikit learn's [module on 'K-nearest Neighbours'](#).

What steps should we take to manage the risks of privacy attacks on AI models?

If you train models and provide them to others, you should assess whether those models may contain personal data or are at risk of revealing it if attacked, and take appropriate steps to mitigate these risks.

You should assess whether the training data contains identified or identifiable personal data of individuals,

either directly or by those who may have access to the model. You should assess the means that may be reasonably likely to be used, in light of the vulnerabilities described above. As this is a rapidly developing area, you should stay up-to-date with the state of the art in both methods of attack and mitigation.

Security and ML researchers are still working to understand what factors make ML models more or less vulnerable to these kinds of attacks, and how to design effective protections and mitigation strategies.

One possible cause of ML models being vulnerable to privacy attacks is known as 'overfitting'. This is where the model pays too much attention to the details of the training data, effectively almost remembering particular examples from the training data rather than just the general patterns. Overfitting can happen where there are too many features included or where there are too few examples in the training data (or both). Model inversion and membership inference attacks can exploit this.

Avoiding overfitting will help, both in mitigating the risk of privacy attacks and also in ensuring that the model is able to make good inferences on new examples it hasn't seen before. However, avoiding overfitting will not completely eliminate the risks. Even models which are not overfitted to the training data can still be vulnerable to privacy attacks.

In cases where confidence information provided by a ML system can be exploited, as in the FRT example above, the risk could be mitigated by not providing it to the end user. This would need to be balanced against the need for genuine end users to know whether or not to rely on its output and will depend on the particular use case and context.

If you are going to provide a whole model to others via an Application Programming Interface (API), you will not be subject to white box attacks in this way, because the API's users will not have direct access to the model itself. However, you might still be subjected to black box attacks.

To mitigate this risk, you could monitor queries from the API's users, in order to detect whether it is being used suspiciously. This may indicate a privacy attack and would require prompt investigation, and potential suspension or blocking of a particular user account. Such measures may become part of common real-time monitoring techniques used to protect against other security threats, such as 'rate-limiting' (reducing the number of queries that can be performed by a particular user in a given time limit).

If your model is going to be provided in whole to a third party, rather than being merely accessible to them via an API, then you will need to consider the risk of 'white box' attacks. As the model provider, you will be less easily able to monitor the model during deployment and thereby assess and mitigate the risk of privacy attacks on it.

However, you remain responsible for assessing and mitigating the risk that personal data used to train your models may be exposed as a result of the way your clients have deployed the model. You may not be able to fully assess this risk without collaborating with your clients to understand the particular deployment contexts and associated threat models.

As part of your procurement policy there should be sufficient information sharing between each party to perform your respective assessments as necessary. In some cases, ML model providers and clients will be joint controllers and therefore need to perform a joint risk assessment.

In cases where the model actually contains examples from the training data by default (as in SVMs and KNNs), this is a transfer of personal data, and you should treat it as such.

What about AI security risks raised by explainable AI?

Recent [research](#) has demonstrated how some proposed methods to make ML models explainable can unintentionally make it easier to conduct privacy attacks on models. For example, when providing an explanation to individuals, there may be a risk that doing so reveals proprietary information about how the AI model works. However, you must take care not to conflate commercial interests with data protection requirements (eg commercial security and data protection security), and instead you should consider the extent to which such a trade-off genuinely exists.

Given that the kind of explanations you may need to provide to data subjects about AI need to be 'in a concise, transparent, intelligible and easily accessible form, using clear and plain language', they will not normally risk commercially sensitive information. However, there may be cases where you need to consider the right of individuals to receive an explanation, and (for example) the interests of businesses to maintain trade secrets, noting that data protection compliance cannot be 'traded away'.

Both of these risks are active areas of research, and their likelihood and severity are the subject of debate and investigation. We will continue to monitor and review these risks and may update this guidance accordingly.

Further reading outside this guidance

ICO and The Alan Turing Institute guidance on '[Explaining decisions made with artificial intelligence](#)'.

What about adversarial examples?

While the main data protection concerns about AI involve accidentally revealing personal data, there are other potential novel AI security risks, such as 'adversarial examples'.

These are examples fed to an ML model, which have been deliberately modified so that they are reliably misclassified. These can be images which have been manipulated, or even real-world modifications such as stickers placed on the surface of the item. Examples include pictures of turtles which are classified as guns, or road signs with stickers on them, which a human would instantly recognise as a 'STOP', but an image recognition model does not.

While such adversarial examples are concerning from a security perspective, they might not raise data protection concerns if they don't involve personal data. The security principle refers to security of the personal data – protecting it against unauthorised processing. However, adversarial attacks don't necessarily involve unauthorised processing of personal data, only a compromise to the system.

However, there may be cases in which adversarial examples can be a risk to the rights and freedoms of individuals. For example, some attacks have been demonstrated on facial recognition systems. By slightly distorting the face image of one individual, an adversary can trick the [facial recognition system](#) into misclassifying them as another (even though a human would still recognise the distorted image as the correct individual). This would raise concerns about the system's statistical accuracy, especially if the system is used to make legal or similarly significant decisions about individuals.

You may also need to consider the risk of adversarial examples as part of your obligations under the Network and Information Systems Regulations 2018 (NIS). The ICO is the competent authority for 'relevant

digital service providers' under NIS. These include online search engines, online marketplaces and cloud computing services. A 'NIS incident' includes incidents which compromise the data stored by network and information systems and the related services they provide. This is likely to include AI cloud computing services. So, even if an adversarial attack does not involve personal data, it may still be a NIS incident and therefore within the ICO's remit.

Further reading outside this guidance

Read our [Guide to NIS](#).

For further information on adversarial attacks on facial recognition systems, see '[Efficient decision-based black-box adversarial attacks on face recognition](#)'.

What data minimisation and privacy-preserving techniques are available for AI systems?

What considerations about the data minimisation principle do we need to make?

The data minimisation principle requires you to identify the minimum amount of personal data you need to fulfil your purpose, and to only process that information, and no more. For example, Article 5(1)(c) of the UK GDPR says



'1. Personal data shall be

adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (data minimisation).'

However, AI systems generally require large amounts of data. At first glance it may therefore be difficult to see how AI systems can comply with the data minimisation principle, yet if you are using AI as part of your processing, you are still required to do so.

Whilst it may appear challenging, in practice this may not be the case. The data minimisation principle does not mean either 'process no personal data' or 'if we process more, we're going to break the law'. The key is that you only process the personal data you need for your purpose.

How you go about determining what is 'adequate, relevant and limited' is therefore going to be specific to your circumstances, and our existing guidance on data minimisation details the steps you should take.

In the context of AI systems, what is 'adequate, relevant and limited' is therefore also case specific. However, there are a number of techniques that you can adopt in order to develop AI systems that process only the data you need, while still remaining functional.

In this section, we explore some of the most relevant techniques for supervised Machine Learning (ML) systems, which are currently the most common type of AI in use.

Within your organisations, the individuals accountable for the risk management and compliance of AI systems need to be aware that such techniques exist and be able to discuss and assess different approaches with your technical staff. For example, the default approach of data scientists in designing and building AI systems might involve collecting and using as much data as possible, without thinking about ways they could achieve the same purposes with less data.

You must therefore implement risk management practices designed to ensure that data minimisation, and all relevant minimisation techniques, are fully considered from the design phase. Similarly, if you buy in AI systems or implement systems operated by third parties (or both), these considerations should form part of the procurement process due diligence.

You should also be aware that, while they may help you comply with the principle of data minimisation, the techniques described here do not eliminate other kinds of risk.

Also, while some techniques will not require any compromise to comply with data minimisation requirements, others may need you to balance data minimisation with other compliance or utility objectives. For example, making more statistically accurate and non-discriminatory ML models.

The first step you should take towards compliance with data minimisation is to understand and map out all the ML processes in which personal data might be used.

Further Reading

 [See Article 5\(1\)\(c\) and Recital 39, and Article 16 \(right to rectification\) and Article 17 \(right to erasure\) of the UK GDPR](#) 

External link

Further reading outside this guidance

[Read our guidance on the data minimisation principle](#)

How should we process personal data in supervised ML models?

Supervised ML algorithms can be trained to identify patterns and create models from datasets ('training data') which include past examples of the type of instances the model will be asked to classify or predict. Specifically, the training data contains both the 'target' variable (ie the thing that the model is aiming to predict or classify), and several 'predictor' variables (ie the input used to make the prediction).

For example, in the training data for a bank's credit risk ML model, the predictor variables might include the age, income, occupation, and location of previous customers, while the target variable will be whether or not the customers repaid their loan.

Once trained, ML systems can then classify and make predictions based on new data containing examples that the system has never seen before. A query is sent to the ML model, containing the predictor variables for a new instance (eg a new customer's age, income, occupation). The model responds with its best guess

as to the target variable for this new instance (eg whether or not the new customer will default on a loan).

Supervised ML approaches therefore use data in two main phases:

1. the **training phase**, when training data is used to develop models based on past examples; and
2. the **inference phase**, when the model is used to make a prediction or classification about new instances.

If the model is used to make predictions or classifications about individual people, then it is very likely that personal data will be used at both the training and inference phases.

What techniques should we use to minimise personal data when designing ML applications?

When designing and building ML applications, data scientists will generally assume that all data used in training, testing and operating the system will be aggregated in a centralised way, and held in its full and original form by a single entity in multiple places throughout the AI system's lifecycle.

However, where this is personal data, you need to consider whether it is necessary to process it for your purpose(s). If you can achieve the same outcome by processing less personal data then by definition, the data minimisation principle requires you to do so.

A number of techniques exist which can help you to minimise the amount of personal data you need to process.

How should we minimise personal data in the training stage?

As we have explained, the training phase involves applying a learning algorithm to a dataset containing a set of features for each individual which are used to generate the prediction or classification.

However, not all features included in a dataset will necessarily be relevant to your purpose. For example, not all financial and demographic features will be useful to predict credit risk. Therefore, you need to assess which features – and therefore what data – are relevant for your purpose, and only process that data.

There are a variety of standard feature selection methods used by data scientists to select features which will be useful for inclusion in a model. These methods are good practice in data science, but they also go some way towards meeting the data minimisation principle.

Also, as discussed in the ICO's previous report on AI and Big Data, the fact that some data might later in the process be found to be useful for making predictions is not enough to establish why you need to keep it for this purpose, nor does it retroactively justify its collection, use, or retention. You must not collect personal data on the off-chance that it might be useful in the future, although you may be able to hold information for a foreseeable event that may not occur, but only if you are able to justify it.

How should we balance data minimisation and statistical accuracy?

In general, when an AI system learns from data (as is the case with ML models), the more data it is trained on, the more statistically accurate it will be. That is, the more likely it will capture any underlying, statistically useful relationships between the features in the datasets. As explained in the section on 'What do we need to do about statistical accuracy?', the fairness principle means that your AI system needs to be

sufficiently statistically accurate for your purposes.

For example, a model for predicting future purchases based on customers' purchase history would tend to be more statistically accurate the more customers are included in the training data. And any new features added to an existing dataset may be relevant to what the model is trying to predict. For example, purchase histories augmented with additional demographic data might further improve the statistical accuracy of the model.

However, generally speaking, the more data points collected about each person, and the more people whose data is included in the data set, the greater the risks to those individuals, even if the data is collected for a specific purpose. The principle of data minimisation requires you not to use more data than is necessary for your purposes. So if you can achieve sufficient accuracy with fewer data points or fewer individuals being included (or both), you should do so.

Further reading outside this guidance

Read our report on [Big data, artificial intelligence, machine learning and data protection](#)

What privacy-enhancing methods should we consider?

There are also a range of techniques for enhancing privacy which you can use to minimise the personal data being processed at the training phase, including:

- perturbation or adding 'noise';
- synthetic data; and
- federated learning

Some of these techniques involve modifying the training data to reduce the extent to which it can be traced back to specific individuals, while retaining its use for the purposes of training well-performing models.

You can apply these types of privacy-enhancing techniques to the training data after you have already collected it. Where possible, however, you should apply them before collecting any personal data, as a part of mitigating the risks to individuals that large datasets can pose.

You can mathematically measure the effectiveness of these privacy-enhancing techniques in balancing the privacy of individuals and the utility of a ML system, using methods such as differential privacy.

Differential privacy is a way to measure whether a model created by an ML algorithm significantly depends on the data of any particular individual used to train it. While mathematically rigorous in theory, meaningfully implementing differential privacy in practice is still challenging.

You should monitor developments in these methods and assess whether they can provide meaningful data minimisation before attempting to implement them. They may not be appropriate or sufficiently mature to deploy in your particular context.

• Perturbation

Modification could involve changing the values of data points belonging to individuals at random (known as 'perturbing' or adding 'noise' to the data) in a way that preserves some of the statistical properties of those

features.

Generally speaking, you can choose how much noise to inject, with obvious consequences for how much you can still learn from the 'noisy data'.

For example, smartphone predictive text systems are based on the words that users have previously typed. Rather than always collecting a user's actual keystrokes, the system could be designed to create 'noisy' (ie false) words at random. This means it makes it substantially less certain which words were 'noise' and which words were actually typed by a specific user.

Although data would be less accurate at individual level, provided the system has enough users, you could still observe patterns, and use these to train your ML model at an aggregate level. The more noise you inject, the less you can learn from the data, but in some cases you may be able to inject sufficient noise to render the data pseudonymous in a way which provides a meaningful level of protection.

• **Synthetic data**

In some cases, you may be able to develop models using 'synthetic' data. This is data which does not relate to real people, but has been generated artificially. To the extent that synthetic data cannot be related to identified or identifiable living individuals, it is not personal data and therefore data protection obligations do not apply when you process it.

However, you will generally need to process some real data in order to determine realistic parameters for the synthetic data. Where that real data can be related to identified or identifiable individuals, then the processing of such data must comply with data protection laws.

Furthermore, in some cases, it may be possible to infer information about the real data which was used to estimate those realistic parameters, by analysing the synthetic data. For example, if the real data contains a single individual who is unusually tall, rich, and old, and your synthetic data contains a similar individual (in order to make the overall dataset statistically realistic), it may be possible to infer that the individual was in the real dataset by analysing the synthetic dataset. Avoiding such re-identification may require you to change your synthetic data to the extent that it would be too unrealistic to be useful for machine learning purposes.

• **Federated learning**

A related privacy-preserving technique is federated learning. This allows multiple different parties to train models on their own data ('local' models). They then combine some of the patterns that those models have identified (known as 'gradients') into a single, more accurate 'global' model, without having to share any training data with each other.

Federated learning is relatively new but has several large-scale applications. These include auto-correction and predictive text models across smartphones, but also for medical research involving analysis across multiple patient databases.

While sharing the gradient derived from a locally trained model presents a lower privacy risk than sharing the training data itself, a gradient can still reveal some personal information about the individuals it was derived from, especially if the model is complex with a lot of fine-grained variables. You therefore still need to assess the risk of re-identification. In the case of federated learning, participating organisations may be considered joint controllers even though they don't have access to each other's data.

Further reading inside this guidance

For more information on controllership in AI, read the section on [controller/processor relationships](#).

Further reading outside this guidance

See '[Rappor \(randomised aggregatable privacy preserving ordinal responses\)](#)' for an example of perturbation.

For an introduction to differential privacy, see '[Differential privacy: an introduction for statistical agencies](#)'.

How should we minimise personal data at the inference stage?

To make a prediction or classification about an individual, ML models usually require the full set of predictor variables for that person to be included in the query. As in the training phase, there are a number of techniques which you can use to minimise personal data, or mitigate risks posed to that data, at the inference stage, including:

- converting personal data into less 'human readable' formats;
- making inferences locally; and
- privacy-preserving query approaches.

We consider these approaches below.

• Converting personal data into less "human readable" formats

In many cases the process of converting data into a format that allows it to be classified by a model can go some way towards minimising it. Raw personal data will usually first have to be converted into a more abstract format for the purposes of prediction. For example, human-readable words are normally translated into a series of numbers (called a 'feature vector').

This means that if you deploy an AI model you may not need to process the human-interpretable version of the personal data contained in the query. For example, if the conversion happens on the user's device.

However, the fact that it is no longer easily human-interpretable does not imply that the converted data is no longer personal. Consider Facial Recognition Technology (FRT), for example. In order for a facial recognition model to work, digital images of the faces being classified have to be converted into 'faceprints'. These are mathematical representations of the geometric properties of the underlying faces (eg the distance between a person's nose and upper lip).

Rather than sending facial images themselves to your servers, photos could be converted to faceprints directly on the individuals' device which captures them before sending them to the model for querying. These faceprints would be less easily identifiable to any humans than face photos.

However, faceprints are still personal (indeed, biometric) data and therefore very much identifiable within

the context of the specific facial recognition models that they are created for. Also, when used for the purposes of uniquely identifying an individual, they would be special category data under data protection law.

- **Making inferences locally**

Another way to minimise the personal data involved in prediction is to host the ML model on the device from which the query is generated and which already collects and stores the individual's personal data. For example, an ML model could be installed on the user's own device and make inferences 'locally', rather than being hosted on a cloud server.

For example, models for predicting what news content a user might be interested in could be run locally on their smartphone. When the user opens the news app the day's news is sent to the phone and the local model would select the most relevant stories to show to the user, based on the user personal habits or profile information which are tracked and stored on the device itself and are not shared with the content provider or app store.

The constraint is that ML models need to be sufficiently small and computationally efficient to run on the user's own hardware. However, recent advances in purpose-built hardware for smartphones and embedded devices mean that this is an increasingly viable option.

It is important to note that local processing is not necessarily out of scope of data protection law. Even if the personal data involved in training is being processed on the user's device, the organisation which creates and distributes the model is still a controller in so far as it determines the means and purposes of processing.

Similarly, if personal data on the user's device is subsequently accessed by a third party, this activity would constitute 'processing' of that data.

- **Privacy-preserving query approaches**

If it is not feasible to deploy the model locally, other privacy-enhancing techniques exist to minimise the data that is revealed in a query sent to a ML model. These allow one party to retrieve a prediction or classification without revealing all of this information to the party running the model; in simple terms, they allow you to get an answer without having to fully reveal the question.

Further reading outside this guidance

See '[Privad: practical privacy in online advertising](#)' and '[Targeted advertising on the handset: privacy and security challenges](#)' for proof of concept examples for making inferences locally.

See '[TAPAS: trustworthy privacy-aware participatory sensing](#)' for an example of privacy-preserving query approaches.

Does anonymisation have a role?

There are conceptual and technical similarities between data minimisation and anonymisation. In some cases, applying privacy-preserving techniques means that certain data used in ML systems is rendered pseudonymous or anonymous.

However, you should note that pseudonymisation is essentially a security and risk reduction technique, and data protection law still applies to personal data that has undergone pseudonymisation. In contrast, 'anonymous information' means that the information in question is no longer personal data and data protection law does not apply to it.


Further reading outside this guidance

[The ICO is currently developing new guidance on anonymisation to take into account of new recent developments and techniques in this field.](#)

What should we do about storing and limiting training data?

Sometimes it may be necessary to retain training data in order to re-train the model, for example when new modelling approaches become available and for debugging. However, where a model is established and unlikely to be re-trained or modified, the training data may no longer be needed. If the model is designed to use only the last 12 months' worth of data, a data retention policy should specify that data older than 12 months be deleted.

Further reading outside this guidance

The European Union Agency for Cybersecurity (ENISA) has [a number of publications about PETs](#) , including research reports.

How do we ensure individual rights in our AI systems?

At a glance

This section explains the challenges to ensure individual rights in AI systems, including rights relating to solely automated decision-making with legal or similarly significant effect. It also covers the role of meaningful human oversight.

Who is this section for?

This section is aimed at those in compliance-focused roles who are responsible for responding to individual rights requests. The section makes reference to some technical terms and measures, which may require input from a technical specialist.

In detail

- [How do individual rights apply to different stages of the AI lifecycle?](#)
- [How do individual rights relate to data contained in the model itself?](#)
- [How do we ensure individual rights relating to solely automated decisions with legal or similar effect?](#)
- [What is the role of human oversight?](#)

How do individual rights apply to different stages of the AI lifecycle?

Under data protection law individuals have a number of rights relating to their personal data. Within AI, these rights apply wherever personal data is used at any of the various points in the development and deployment lifecycle of an AI system. This therefore covers personal data:

- contained in the training data;
- used to make a prediction during deployment, and the result of the prediction itself; or
- that might be contained in the model itself.

This section describes what you may need to consider when developing and deploying AI and complying with the individual rights of information, access, rectification, erasure, and to restriction of processing, data portability, and objection (rights referred to in Articles 13-21 of the UK GDPR). It does not cover each right in detail but discusses general challenges to complying with these rights in an AI context, and where appropriate, mentions challenges to specific rights.

Rights that individuals have about solely automated decisions that affect them in legal or similarly significant ways are discussed in more detail in [‘What is the role of human oversight?’](#), as these rights raise particular challenges when using AI.

How should we ensure individual rights requests for training data?

When creating or using ML models, you invariably need to obtain data to train those models.

For example, a retailer creating a model to predict consumer purchases based on past transactions needs a large dataset of customer transactions to train the model on.

Identifying the individuals that the training data is about is a potential challenge to ensuring their rights. Typically, training data only includes information relevant to predictions, such as past transactions, demographics, or location, but not contact details or unique customer identifiers. Training data is also typically subjected to various measures to make it more amenable to ML algorithms. For example, a detailed timeline of a customer's purchases might be transformed into a summary of peaks and troughs in their transaction history.

This process of transforming data prior to using it for training a statistical model, (for example, transforming numbers into values between 0 and 1) is often referred to as 'pre-processing'. This can create confusion about terminology in data protection, where 'processing' refers to any operation or set of operations which is performed on personal data. So 'pre-processing' (in machine learning terminology) is still 'processing' (in data protection terminology) and therefore data protection still applies.

Because these processes involve converting personal data from one form into another potentially less detailed form, they may make training data potentially much harder to link to a particular named individual. However, in data protection law this is not necessarily considered sufficient to take that data out of scope. You therefore still need to consider this data when you are responding to individuals' requests to exercise their rights.

Even if the data lacks associated identifiers or contact details, and has been transformed through pre-processing, training data may still be considered personal data. This is because it can be used to 'single out' the individual it relates to, on its own or in combination with other data you may process (even if it cannot be associated with a customer's name).

For example, the training data in a purchase prediction model might include a pattern of purchases unique to one customer.

In this example, if a customer provided a list of their recent purchases as part of their request, the organisation may be able to identify the portion of the training data that relates to them.

In these kinds of circumstances, you are obliged to respond to an individual's request, assuming you have taken reasonable measures to verify their identity and no other exceptions apply.

There may be times where you are not able to identify an individual in the training data, directly or indirectly. Provided you are able to demonstrate this, individual rights under Articles 15 to 20 do not apply. However, if the individual provides additional information that enables identification, this is no longer the case and you need to fulfil any request they make. You should consult our guidance on determining what is personal data for more information about identifiability.

We recognise that the use of personal data with AI may sometimes make it harder to fulfil individual rights to information, access, rectification, erasure, restriction of processing, and notification. If a request is manifestly unfounded or excessive, you may be able to charge a fee or refuse to act on the request. However, you should not regard requests about such data as manifestly unfounded or excessive just because they may be harder to fulfil in the context of AI or the motivation for requesting them may be unclear in comparison to other access requests you might typically receive.

If you outsource an AI service to another organisation, this could also make the process of responding to rights requests more complicated when the personal data involved is processed by them rather than you. When procuring an AI service, you must choose one which allows individual rights to be protected and enabled, in order to meet your obligations as a controller. If your chosen service is not designed to easily comply with these rights, this does not remove or change those obligations. If you are operating as a controller, your contract with the processor must stipulate that the processor assist you in responding to rights requests. If you are operating an AI service as a joint controller, you need to decide with your fellow controller(s) who will carry out which obligations. See the section [‘How should we understand controller/processor relationships in AI?’](#) for more details.

In addition to these considerations about training data and individual rights in general, below we outline some considerations about how particular individual rights (rectification, erasure, portability, and information) may relate to training data.

- **Right to rectification**

The right to rectification may apply to the use of personal data to train an AI system. The steps you should take for rectification depend on the data you process as well as the nature, scope, context and purpose of that processing. The more important it is that the personal data is accurate, the greater the effort you should put into checking its accuracy and, if necessary, taking steps to rectify it.

In the case of training data for an AI system, one purpose of the processing may be to find general patterns in large datasets. In this context, individual inaccuracies in training data may be less important, as they are not likely to affect the performance of the model, since they are just one data point among many, when compared to personal data that you might use to take action about an individual.

For example, you may think it more important to rectify an incorrectly recorded customer delivery address than to rectify the same incorrect address in training data. Your rationale is likely to be that the former could result in a failed delivery, but the latter would barely affect the overall statistical accuracy of the model.

However, in practice, the right of rectification does not allow you to disregard any requests because you think they are less important for your purposes.

- **Right to erasure**

You may also receive requests for the erasure of personal data contained within training data. You should note that whilst the right to erasure is not absolute, you still need to consider any erasure request you receive, unless you are processing the data on the basis of a legal obligation or public task (both of which are unlikely to be lawful bases for training AI systems – see the [section on lawful bases](#) for more information).

The erasure of one individual’s personal data from the training data is unlikely to affect your ability to fulfil the purposes of training an AI system (as you are likely to still have sufficient data from other individuals). You are therefore unlikely to have a justification for not fulfilling the request to erase their personal data from your training dataset.

Complying with a request to erase training data does not entail erasing all ML models based on this data, unless the models themselves contain that data or can be used to infer it (situations which we will cover in the section below).

- **Right to data portability**

Individuals have the right to data portability for data they have 'provided' to a controller, where the lawful basis of processing is consent or contract. 'Provided data' includes data the individual has consciously input into a form, but also behavioural or observational data gathered in the process of using a service.

In most cases, data used for training a model (eg demographic information or spending habits) counts as data 'provided' by the individual. The right to data portability therefore applies in cases where this processing is based on consent or contract.

However, as discussed above, pre-processing methods are usually applied which significantly change the data from its original form into something that can be more effectively analysed by machine learning algorithms. Where this transformation is significant, the resulting data may no longer count as 'provided'.

In this case the data is not subject to data portability, although it does still constitute personal data and as such other data protection rights still apply (eg the right of access). However, the original form of the data from which the pre-processed data was derived is still subject to the right to data portability (if provided by the individual under consent or contract and processed by automated means).

- **Right to be informed**

You must inform individuals if their personal data is going to be used to train an AI system, to ensure that processing is fair and transparent. You should provide this information at the point of collection. If the data was initially processed for a different purpose, and you later decide to use it for the separate purpose of training an AI system, you need to inform the individuals concerned (as well as ensuring the new purpose is compatible with the previous one). In some cases, you may not have obtained the training data from the individual, and therefore not have had the opportunity to inform them at the time you did so. In such cases, you should provide the individual with the information specified in Article 14 within a reasonable period, one month at the latest, unless a relevant exemption from Article 14(5) applies.

Since using an individual's data for the purposes of training an AI system does not normally constitute making a solely automated decision with legal or similarly significant effects, you only need to provide information about these decisions when you are taking them. However, you still need to comply with the main transparency requirements.

For the reasons stated above, it may be difficult to identify and communicate with the individuals whose personal data is contained in the training data. For example, training data may have been stripped of any personal identifiers and contact addresses (while still remaining personal data). In such cases, it may be impossible or involve a disproportionate effort to provide information directly to the individual.

Therefore, instead you should take appropriate measures to protect the individual's rights and freedoms and legitimate interests. For example, you could provide public information explaining where you obtained the data from that you use to train your AI system, and how to object.

How should we ensure individual rights requests for AI outputs?

Typically, once deployed, the outputs of an AI system are stored in a profile of an individual and used to take some action about them.

For example, the product offers a customer sees on a website might be driven by the output of the predictive model stored in their profile. Where this data constitutes personal data, it will generally be

subject to all of the rights mentioned above (unless exemptions or other limitations to those rights apply).

Whereas individual inaccuracies in training data may have a negligible effect, an inaccurate output of a model could directly affect the individual. Requests for rectification of model outputs (or the personal data inputs on which they are based) are therefore more likely to be made than requests for rectification of training data. However, as said above, predictions are not inaccurate if they are intended as prediction scores as opposed to statements of fact. If the personal data is not inaccurate then the right to rectification does not apply.

Personal data resulting from further analysis of provided data is not subject to the right to portability. This means that the outputs of AI models such as predictions and classifications about individuals are out of scope of the right to portability.

In some cases, some or all of the features used to train the model may themselves be the result of some previous analysis of personal data. For example, a credit score which is itself the result of statistical analysis based on an individual's financial data might then be used as a feature in an ML model. In these cases, the credit score is not included within scope of the right to data portability, even if other features are.

Further reading outside this guidance

Read our guidance on [individual rights](#), including:

- the [right to be informed](#);
- the [right of access](#);
- the [right to erasure](#);
- the [right to rectification](#); and
- the [right to data portability](#).

How do individual rights relate to data contained in the model itself?

In addition to being used in the inputs and outputs of a model, in some cases personal data might also be contained in a model itself. As explained in [‘what types of privacy attacks apply to AI models?’](#), this could happen for two reasons; by design or by accident.

How should we fulfil requests about models that contain data by design?

When personal data is included in models by design, it is because certain types of models, such as Support Vector Machines (SVMs), contain some key examples from the training data in order to help distinguish between new examples during deployment. In these cases, a small set of individual examples are contained somewhere in the internal logic of the model.

The training set typically contains hundreds of thousands of examples, and only a very small percentage of them end up being used directly in the model. Therefore, the chances that one of the relevant individuals makes a request are very small; but remains possible.

Depending on the particular programming library in which the ML model is implemented, there may be a

built-in function to easily retrieve these examples. In these cases, it is likely to be practically possible for you to respond to an individual's request. To enable this, where you are using models which contain personal data by design, you should implement them in a way that allows the easy retrieval of these examples.

If the request is for access to the data, you could fulfil this without altering the model. If the request is for rectification or erasure of the data, this may not be possible without re-training the model (either with the rectified data, or without the erased data), or deleting the model altogether.

While it is not a legal requirement, having a well-organised model management system and deployment pipeline will make it easier and cheaper to accommodate these requests, and re-training and redeploying your AI models accordingly will be less costly.

How should we fulfil requests about data contained in models by accident?

Aside from SVMs and other models that contain examples from the training data by design, some models might 'leak' personal data by accident. In these cases, unauthorised parties may be able to recover elements of the training data, or infer who was in it, by analysing the way the model behaves.

The rights of access, rectification, and erasure may be difficult or impossible to exercise and fulfil in these scenarios. Unless the individual presents evidence that their personal data could be inferred from the model, you may not be able to determine whether personal data can be inferred and therefore whether the request has any basis.

You should regularly and proactively evaluate the possibility of personal data being inferred from models in light of the state-of-the-art technology, so that you minimise the risk of accidental disclosure.

How do we ensure individual rights relating to solely automated decisions with legal or similar effect?

There are specific provisions in data protection law covering individuals' rights where processing involves solely automated individual decision-making, including profiling, with legal or similarly significant effects. These provisions cover both information you have to provide proactively about the processing and individuals' rights in relation to a decision made about them.

Under Articles 13 (2)(f) and 14 (2)(g), you must tell people whose data you are processing that you are doing so for automated decision-making and give them "meaningful information about the logic involved, as well as the significance and the envisaged consequences" of the processing for them. Under Article 15 (2)(h) you must also tell them about this if they submit a subject access request.

In addition, data protection requires you to implement suitable safeguards when processing personal data to make solely automated decisions that have a legal or similarly significant impact on individuals. These safeguards include the right for individuals to:

- obtain human intervention;
- express their point of view;
- contest the decision made about them; and
- obtain an explanation about the logic of the decision.

For processing involving solely automated decision-making that falls under Part 2 of the DPA 2018, these safeguards differ to those in the UK GDPR if the lawful basis for that processing is a requirement or authorisation by law.

For processing involving solely automated decision-making that falls under Part 3 of the DPA 2018, the applicable safeguards will depend on regulations provided in the particular law authorising the automated decision-making. Although the individual has the right to request that you reconsider the decision or take a new decision that is not based solely on automated processing.

These safeguards cannot be token gestures. Human intervention should involve a review of the decision, which must be carried out by someone with the appropriate authority and capability to change that decision. That person's review should also include an assessment of all relevant data, including any information an individual may provide.

The conditions under which human intervention qualifies as meaningful are similar to which render a decision non-solely automated (see ['What is the difference between solely automated and partly automated decision-making?'](#) below). However, a key difference is that in solely automated contexts, human intervention is only required on a case-by-case basis to safeguard the individual's rights, whereas for a system to qualify as **not** solely automated, meaningful human intervention is required in **every** decision.

Note that if you are using automated decision making, as well as implementing suitable safeguards, you must also have a suitable lawful basis. See ['How do we identify our purposes and lawful basis when using AI?'](#) and ['What is the impact of Article 22 of the UK GDPR?'](#) above.

Further reading outside this guidance

See the ICO and The Alan Turing [guidance on 'Explaining decisions made with Artificial Intelligence'](#)

See our [guidance on rights related to automated decision-making including profiling](#)

Also see our [in-depth guidance on rights related to automated decision-making including profiling](#)

Further reading – European Data Protection Board

The European Data Protection Board (EDPB), which has replaced the Article 29 Working Party (WP29), includes representatives from the data protection authorities of each EU member state. It adopts guidelines for complying with the requirements of the EU version of the GDPR.

[WP29 published guidelines on automated individual decision-making and profiling, which the EDPB endorsed in May 2018](#) [↗](#).

EDPB guidelines are no longer directly relevant to the UK regime and are not binding under the UK regime. However, they may still provide helpful guidance on certain issues.

Why could rights relating to automated decisions be a particular issue for AI systems?

The type and complexity of the systems involved in making solely automated decisions affect the nature and severity of the risk to people's data protection rights and raise different considerations, as well as compliance and risk management challenges.

Firstly, transparency issues arise because of complex algorithms, such as neural networks or complex software supply chains. Frictionless design such as ambient intelligence can exacerbate this challenge. A 'smart' home device that decides what job advert to serve depending on the content of an individual's speech may not enable that individual to understand a decision is being made about them. However, irrespective of the type of AI system you use, if it processes personal data you need to comply with data protection law.

Basic systems, which automate a relatively small number of explicitly written rules, are unlikely to be considered AI (eg a set of clearly expressed 'if-then' rules to determine a customer's eligibility for a product). However, the resulting decisions could still constitute automated decision-making within the meaning of data protection law.

It should also be relatively easy for a human reviewer to identify and rectify any mistake, if a decision is challenged by an individual because of a system's high interpretability.

However other systems, such as those based on ML, may be more complex and present more challenges for meaningful human review. ML systems make predictions or classifications about people based on data patterns. Even when they are highly [statistically accurate](#), they will occasionally reach the wrong decision in an individual case. Errors may not be easy for a human reviewer to identify, understand or fix.

While not every challenge from an individual will result in the decision being overturned, you should expect that many could be. There are two particular reasons why this may be the case in ML systems:

- **the individual is an 'outlier'**, ie their circumstances are substantially different from those considered in the training data used to build the AI system. Because the ML model has not been trained on enough data about similar individuals, it can make incorrect predictions or classifications; or
- **assumptions in the AI design can be challenged**, eg a continuous variable such as age, might have been broken up ('binned') into discrete age ranges, like 20-39, as part of the modelling process. Finer-grained 'bins' may result in a different model with substantially different predictions for people of different ages. The validity of this data pre-processing and other design choices may only come into question as a result of an individual's challenge.

What steps should we take to fulfil rights related to automated decision-making?

You should:

- consider the system requirements necessary to support a meaningful human review from the design phase. Particularly, the interpretability requirements and effective user-interface design to support human reviews and interventions;
- design and deliver appropriate training and support for human reviewers; and
- give staff the appropriate authority, incentives and support to address or escalate individuals' concerns and, if necessary, override the AI system's decision.

However, there are some additional requirements and considerations you should be aware of.

The ICO's and the Alan Turing Institute's 'Explaining decisions made with AI' guidance looks at how, and to

what extent, complex AI systems might affect your ability to provide meaningful explanations to individuals. However, complex AI systems can also impact the effectiveness of other mandatory safeguards. If a system is too complex to explain, it may also be too complex to meaningfully contest, intervene on, review, or put an alternative point of view against.

For example, if an AI system uses hundreds of features and a complex, non-linear model to make a prediction, then it may be difficult for an individual to determine which variables or correlations to object to. Therefore, safeguards around solely automated AI systems are mutually supportive, and should be designed holistically and with the individual in mind.

The information about the logic of a system and explanations of decisions should give individuals the necessary context to decide whether, and on what grounds, they would like to request human intervention. In some cases, insufficient explanations may prompt individuals to resort to other rights unnecessarily. Requests for intervention, expression of views, or contests are more likely to happen if individuals don't feel they have a sufficient understanding of how the decision was reached.

Transparency measures do not just relate to your model's internal logic but also to the constraints in which it operates. For example, it may be useful for an individual to know whether a gender classifier only uses binary variables, or what options it considers. This is because they may find being deprived of certain outcomes such as a non-binary label against their reasonable expectations and therefore the processing unfair.

It is good practice to communicate the inherent uncertainty of AI-driven decision-making and predictions. This helps individuals and groups evaluate the predictions your system makes and enables them to challenge any decisions that result.

For example, if loan applicants know the confidence interval that indicates the reliability of an inference, it could help those whose applications are rejected to more effectively challenge the decision.

It is also good practice for you to communicate different levels of uncertainty about the data sources involved in your decision-making process, as well as any potential biases in the data itself.

The process for individuals to exercise their rights should be simple and user friendly. For example, if you communicate the result of the solely automated decision through a website, the page should contain a link or clear information allowing the individual to contact a member of staff who can intervene, without any undue delays or complications.

You are also required to keep a record of all decisions made by an AI system as part of your accountability and documentation obligations. This should also include whether an individual requested human intervention, expressed any views, contested the decision, and whether you changed the decision as a result.

You should monitor and analyse this data. If decisions are regularly changed in response to individuals exercising their rights, you should then consider how you will amend your systems accordingly. Where your system is based on ML, this might involve including the corrected decisions into fresh training data, so that similar mistakes are less likely to happen in future.

More substantially, you may identify a need to collect more or better training data to fill in the gaps that led to the erroneous decision, or modify the model-building process (ie by changing the feature selection).

In addition to being a compliance requirement, this is also an opportunity for you to improve the performance of your AI systems and, in turn, build individuals' trust in them. However, if grave or frequent mistakes are identified, you need to take immediate steps to understand and rectify the underlying issues and, if necessary, suspend the use of the automated system.

There are also trade-offs that having a human-in-the-loop may entail. Either in terms of a further erosion of privacy, if human reviewers need to consider additional personal data in order to validate or reject an AI generated output, or the possible reintroduction of human biases at the end of an automated process.

Further reading outside this guidance

Read our guidance on [Documentation](#)

European [guidelines on automated decision-making and profiling](#).

ICO and The Alan Turing Institute guidance on '[Explaining decisions made with artificial intelligence](#)'.

What is the role of human oversight?

When AI is used to inform legal or similarly significant decisions about individuals, there is a risk that these decisions are made without appropriate human oversight. For example, whether they have access to financial products or job opportunities. This infringes Article 22 of the UK GDPR.

To mitigate this risk, you should ensure that people assigned to provide human oversight remain engaged, critical and able to challenge the system's outputs wherever appropriate.

What is the difference between solely automated and partly automated decision-making?

You can use AI systems in two ways:

- for **automated decision-making** (ADM), where the system makes a decision automatically; or
- as **decision-support**, where the system only **supports** a human decision-maker in their deliberation.

For example, you could use AI in a system which automatically approves or rejects a financial loan, or merely to provide additional information to support a loan officer when deciding whether to grant a loan application.

Whether solely automated decision-making is generally more or less risky than partly automated decision-making depends on the specific circumstances. You therefore need to evaluate this based on your own context.

Regardless of their relative merits, automated decisions are treated differently to human decisions in data protection law. Specifically, Article 22 of the UK GDPR restricts fully automated decisions which have legal or similarly significant effects on individuals to a more limited set of lawful bases and requires certain safeguards to be in place.

By contrast, the use of decision-support tools are not subject to these conditions. However, the human input needs to be **meaningful**. You should be aware that a decision does not fall outside the scope of

Article 22 just because a human has 'rubber-stamped' it. The degree and quality of human review and intervention before a final decision is made about an individual are key factors in determining whether an AI system is being used for automated decision-making or merely as decision-support.

Ensuring human input is meaningful in these situations is not just the responsibility of the human using the system. Senior leaders, data scientists, business owners, and those with oversight functions if you have them, among others, are expected to play an active role in ensuring that AI applications are designed, built, and used as intended.

If you are deploying AI systems which are designed as decision-support tools, and therefore are intended to be outside the scope of Article 22, you should be aware of existing guidance on these issues from both the ICO and the EDPB.

The key considerations are:

- human reviewers must be involved in checking the system's recommendation and should not just apply the automated recommendation to an individual in a routine fashion;
- reviewers' involvement must be active and not just a token gesture. They should have actual 'meaningful' influence on the decision, including the 'authority and competence' to go against the recommendation; and
- reviewers must 'weigh-up' and 'interpret' the recommendation, consider all available input data, and also take into account other additional factors.

Further Reading

 [See UK GDPR Article 22 and Recital 71](#) 

External link

 [See DPA 2018 Sections 14, 49 and 50](#) 

External link

Further reading outside this guidance

Read our [guidance on automated decision-making and profiling](#).

European [guidelines on automated decision-making and profiling](#).

What are the additional risk factors in AI systems?

You need to consider the meaningfulness of human input in any automated decision-making system you use, however basic it may be.

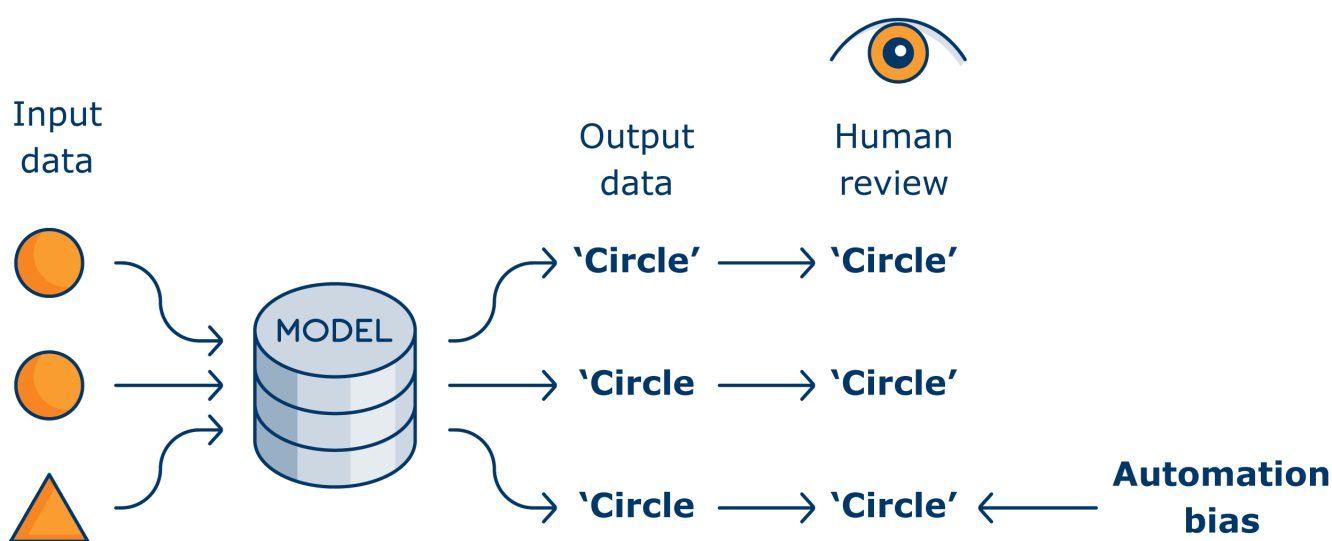
However, in more complex AI systems, there are two additional factors that could potentially cause a system intended as decision-support to inadvertently fail to ensure meaningful human input and therefore fall into the scope of Article 22. They are:

- automation bias; and
- lack of interpretability.

What does 'automation bias' mean?

AI models are based on mathematics and data. Because of this, people tend to think of them as objective and trust their output regardless of how statistically accurate it is.

The terms **automation bias** or **automation-induced complacency** describe how human users routinely rely on the output generated by a decision-support system and stop using their own judgement or stop questioning whether the output might be wrong.



What does 'lack of interpretability' mean?

Some types of AI systems may have outputs which are difficult for a human reviewer to interpret, for example those which rely on complex, high-dimensional 'deep learning' models.

If the outputs of AI systems are not easily interpretable, and other explanation tools are not available or reliable, there is a risk that a human is not able to meaningfully assess the output of an AI system and factor it into their own decision-making.

If meaningful reviews are not possible, the reviewer may start to just agree with the system's recommendations without judgement or challenge. This means the resulting decisions are effectively 'solely automated'.

Should we distinguish solely from partly automated AI systems?

Yes. You should take a clear view on the intended use of any AI system from the beginning. You should specify and document clearly whether you are using AI to support or enhance human decision-making, or to make solely automated decisions.

Your senior management should review and sign-off the intended use of any AI system, making sure that it is in line with your organisation's risk appetite. This means senior management needs to have a solid understanding of the key risk implications associated with each option and be ready and equipped to provide an appropriate degree of challenge.

You must also ensure clear lines of accountability and effective risk management policies are in place from the outset. If AI systems are only intended to support human decisions, then your policies should specifically address additional risk factors such as automation bias and lack of interpretability.

It is possible that you:

- may not know in advance whether a solely or partly automated AI application will meet your needs best; or
- believe that a solely automated AI system will more fully achieve the intended outcome of your processing, but that it may carry more risks to individuals than a partly automated system.

In these cases, your risk management policies and DPIAs should clearly reflect this and include the risk and controls for each option throughout the AI system's lifecycle.

How can we address risks of automation bias?

You may think you can address automation bias chiefly by improving the effectiveness of the training and monitoring of human reviewers. While training is a key component of effective AI risk management, you should have controls to mitigate automation bias in place from the start of the project, including the scoping and design phases as well as development and deployment.

During the design and build phase all relevant parts of your organisation should work together to develop design requirements that support a meaningful human review from the outset (eg business owners, data scientists and those with oversight functions if you have them).

You must think about what features you expect the AI system to consider and which additional factors the human reviewers should take into account before finalising their decision. For example, the AI system could consider quantitatively measurable properties like how many years of experience a job applicant has, while a human reviewer qualitatively assesses other aspects of an application (eg written communication).

If human reviewers can only access or use the same data used by the AI system, then arguably they are not taking into account other additional factors. This means that their review may not be sufficiently meaningful, and the decision may end up being considered as 'solely automated'.

Where necessary, you should consider how to capture additional factors for consideration by the human reviewers. For example, they might interact directly with the person the decision is about to gather such information.

Those in charge of designing the front-end interface of an AI system must understand the needs, thought processes, and behaviours of human reviewers and allow them to effectively intervene. It may therefore be helpful to consult and test options with human reviewers early on.

However, the features of the AI systems you use also depend on the data available, the type of model(s) selected, and other system building choices. You need to test and confirm any assumptions made in the design phase once the AI system has been trained and built.

How can we address risks of interpretability?

You should also consider interpretability from the design phase. However, interpretability is challenging to define in absolute terms and can be measured in different ways. For example, can the human reviewer:

- predict how the system's outputs will change if given different inputs;
- identify the most important inputs contributing to a particular output; and
- identify when the output might be wrong?

This is why it is important that you define and document what interpretability means, and how to measure it, in the specific context of each AI system you wish to use and the personal data that system will process.

Some AI systems are more interpretable than others. For example, models that use a small number of human-interpretable features (eg age and weight), are likely to be easier to interpret than models that use a large number of features.

The relationship between the input features and the model's output can also be either simple or complicated. Simple rules, which set conditions under which certain inferences can be made, as is the case with decision trees, are easier to interpret.

Similarly, linear relationships (where the value of the output increases proportional to the input) may be easier to interpret than relationships that are non-linear (where the output value is not proportional to the input) or non-monotonic (where the output value may increase or decrease as the input increases).

One approach to address low interpretability is the use of 'local' explanations, using methods like Local Interpretable Model-agnostic Explanation (LIME), which provides an explanation of a specific output rather than the model in general.

LIMEs use a simpler surrogate model to summarise the relationships between input and output pairs that are similar to those in the system you are trying to interpret. In addition to summaries of individual predictions, LIMEs can sometimes help detect errors (eg to see what specific part of an image has led a model to classify it incorrectly).

However, they do not represent the logic underlying the AI system and its outputs and can be misleading if misused, especially with certain kinds of models (eg high-dimensional models). You should therefore assess whether in your context, LIME and similar approaches will help the human decision-maker to meaningfully interpret the AI system and its output.

Many statistical models can also be designed to provide a confidence score alongside each output, which could help a human reviewer in their own decision-making. A lower confidence score indicates that the human reviewer needs to have more input into the final decision. (See [What do we need to do about statistical accuracy?](#))

Assessing the interpretability requirements should be part of the design phase, allowing you to develop explanation tools as part of the system, if required.

This is why your risk management policies should establish a robust, risk-based, and independent approval process for each processing operation that uses AI. They should also set out clearly who is responsible for the testing and final validation of the system before it is deployed. Those individuals should be accountable for any negative impact on interpretability and the effectiveness of human reviews and only provide sign-off if AI systems are in line with the adopted risk management policy.

How should we train our staff to address these risks?

Training your staff is pivotal to ensuring an AI system is considered partly automated. As a starting point, you should train (or retrain) your human reviewers to:

- understand how an AI system works and its limitations;
- anticipate when the system may be misleading or wrong and why;
- have a healthy level of scepticism in the AI system's output and given a sense of how often the system could be wrong;
- understand how their own expertise is meant to complement the system, and provide them with a list of factors to take into account; and
- provide meaningful explanations for either rejecting or accepting the AI system's output – a decision they should be responsible for. You should also have a clear escalation policy in place.

In order for the training to be effective, it is important that:

- human reviewers have the authority to override the output generated by the AI system and they are confident that they will not be penalised for so doing. This authority and confidence cannot be created by policies and training alone: a supportive organisational culture is also crucial; and
- any training programme is kept up to date in line with technological developments and changes in processes, with human reviewers being offered 'refresher' training at intervals, where appropriate.

We have focused here on the training of human reviewers; however, it is worth noting that you should also consider whether any other function requires additional training to provide effective oversight (eg risk or internal audit).

What monitoring should we undertake?

The analysis of why, and how many times, a human reviewer accepted or rejected the AI system's output is a key part in an effective risk monitoring system.

If risk monitoring reports flag that your human reviewers are routinely agreeing with the AI system's outputs, and cannot demonstrate they have genuinely assessed them, then their decisions may effectively be classed as solely automated under UK GDPR.

You need to have controls in place to keep risk within target levels. When outcomes go beyond target levels, you should have processes to swiftly assess compliance and take action if necessary. This might include temporarily increasing human scrutiny, or ensuring that you have an appropriate lawful basis and safeguards, in case the decision-making does effectively become fully automated.

Further reading outside this guidance

ICO and The Alan Turing Institute guidance on '[Explaining decisions made with artificial intelligence](#)'.

Annex A: Fairness in the AI lifecycle

■ [Latest update](#)

15 March 2023 - This is a new chapter with new content. This section is about data protection fairness considerations across the AI lifecycle, from problem formulation to decommissioning. It sets out why fundamental aspects of building AI such as underlying assumptions, abstractions used to model a problem, the selection of target variables or the tendency to over-rely on quantifiable proxies may have an impact on fairness. This chapter also explains the different sources of bias that can lead to unfairness and possible mitigation measures. Technical terms are also explained in the updated glossary.

At a glance

This section is about data protection fairness considerations across the AI lifecycle, from problem formulation to decommissioning. It sets out potential sources of bias. For example, your choice of training data, or the historical bias embedded in the environment you choose to collect your data from. It also explains how fairness can be impacted by fundamental aspects of AI development. For example, the choice around what aspects of the real-world problem to include or omit from your decision-making process, or how you assume what you are trying to predict is linked to the data you process.

However, fairness for data protection is not the only concept of fairness you need to consider. There may be sector-specific concepts as well as obligations in relation to discrimination under the Equality Act. This guidance only covers data protection fairness.

Technical terms, such as reward function or regularisation, are explained in the glossary of the main guidance on AI and data protection.

Who is this section for?

This section is predominantly for AI engineers and key decision-makers in the development and use of AI products and services. It also provides useful foundational knowledge for data protection officers, risk managers, as well as information for the broader public around the processes or decisions that may lead to unfair outcomes in the context of AI and any mitigations.

In detail

- [How to use this annex](#)
- [Project initiation and design](#)
- [Before you process personal data](#)
- [Data collection](#)
- [Data analysis and pre-processing](#)

- [Model development](#)
- [Model evaluation](#)
- [Deployment and monitoring](#)
- [Decommissioning or replacing an AI system](#)

How to use this annex

Examining the different data protection fairness considerations across the AI lifecycle is not a linear process. You are likely to need to take an iterative approach, moving back and forth between different steps.

The AI lifecycle does not have a standardised format, so you should see the one we've used here as indicative. We divide the lifecycle into the following stages:

- Project initiation and design
- Before you process personal data
- Data collection and procurement
- Data analysis and pre-processing
- Model development
- Model evaluation
- Model deployment and monitoring
- Retiring or decommissioning

If you process personal data, you must consider the data protection principles across the AI lifecycle, not just at specific stages. For example, when you use AI to make solely automated decisions, you must consider an individual's ability to understand and contest those decisions, throughout the lifecycle.

Organisational measures to help you comply with fairness include:

- adopting a data protection by design approach;
- conducting a DPIA; and
- ensuring appropriate governance measures around your deployment act as safeguards against unfairness.

The section on fairness in the main guidance on AI and data protection sets out some of these relevant approaches.

The technical approaches we set out in this annex are neither exhaustive nor a 'silver bullet'. Determining which are appropriate will be a process of trial and error, particularly as techniques evolve and mature, or demonstrate limitations as they do so. For any technique you use, you should assess:

- whether it works as intended;
- how it interacts or influences your entire decision-making process;
- whether it reinforces or replicates unfair discrimination; and
- any detrimental unintended consequences that may arise (eg creating additional risks to privacy, data protection and fundamental rights).

The specific technical and organisational measures you need to adopt depend on the nature, scope, context, and purpose of your processing and the particular risks it poses. In any case, you should test the effectiveness of these measures at the design stage and detail:

- how the measures mitigate risks; and
- whether they introduce other risks, and steps you will take to manage these.

It is important to acknowledge that most of the technical approaches we set out in this section seek to address the risk of discrimination. You should remember that data protection's fairness is broader than discrimination and can be too context-specific to be neatly addressed by technical approaches alone. Nevertheless, these technical approaches can help you.

If you are procuring AI as a service or off-the-shelf models, asking for documentation could assist you with your fairness compliance obligations as the controller for processing your customer data. This could include:

- information around the demographic groups a model was originally or continues to be trained on;
- what, if any, underlying bias has been detected or could emerge; or
- any algorithmic fairness testing that has already been conducted.

Further reading in ICO guidance

[Outcome fairness](#) (Explaining decisions made with AI)

1. Project initiation and design

Frame the problem, set out your goals, and map your objectives

At this stage in the lifecycle, you should:

- effectively frame the real-world problem you seek to solve via AI; and
- clearly articulate your objectives.

This is known as the problem formulation stage, where you translate a real-life problem into a mathematical or computable one that is amenable to data science. You should pay attention to fairness at this stage. This is because these first steps influence the decisions you take later, such as trade-offs or benchmarking during development and testing.

You should clearly set out the problem with detail and clarity.

This enables you to test the validity of how you frame the real-world problem and translate it into a computable one. For example, whether the prediction you are trying to make is:

- directly observable in the available data; or
- a construct you can only observe indirectly (eg are you directly measuring whether someone is a 'good employee' or are you just observing indirect indicators, such as performance reviews and sales?).

Depending on your context, you may optimise for more than one objective, tackling a multi-objective

problem. For example, an organisation may decide to use AI as part of its recruitment strategy to sift job applications. The organisation can design the system to take into account multiple objectives. This could include minimising commuting time, mapping salary expectations to resources, maximising flexibility around working hours and other variables. If you clearly define and articulate the different objectives at this early stage, you are better placed to ensure that the impacts your system has are fair.

Examine the decision space

Problem formulation also involves evaluating the “decision space”. This refers to the set of actions that you make available to your decision-makers. When you consider using AI to tackle a complex issue, it is important to evaluate the effects of limiting the decision space to binary choices. This may lead to unfair outcomes, such as increased risk of making unfair decisions about individuals or groups.

Example

An organisation uses an AI system to approve or decline loans. If it limits the decision space to this binary choice, without considering the terms under which the loan is granted to different individuals, it may lead to unfair outcomes. For example, giving loans to vulnerable people under punitive conditions would make them unsustainable.

Certain types of decisions require a nuanced, case-by-case approach because of their socio-political impact, scale or legal implications. This is the context that Article 22 of the UK GDPR tries to capture. Not all problems can be effectively expressed via (or meaningfully solved by) algorithms that only reflect statistics and probabilities.

Problem formulation: what you assume, measure and infer

Problem formulation involves a concept known as the “construct space”. This involves using proxies, or “constructs”, for qualities that you cannot observe in the real world. For example, creditworthiness is a construct. This stage generally involves four aspects. These may introduce their own risks. For example:

- **Abstraction** reduces the complexity of a problem by filtering out irrelevant properties while preserving those necessary to solve it. As this implies a loss of information, you should ensure that you do not discard information that is critical to the solution or that can help you avoid unjustified adverse effects on individuals. One way of mitigating this risk is to consult additional expertise to improve your contextual knowledge.
- **Assumptions** relate to a series of decisions about interdependencies and causality that are taken as a given. For example, when you use a measurable, observable construct to represent an unmeasurable, unobservable property you assume an interdependency between the two. However, certain properties that are fundamental to answering your problem may not just be unmeasurable, but also unable to be represented by an unobservable property (or construct). This means that your model may be built on invalid assumptions that will impact the effects it has on individuals when you deploy it. This is why it is important to test and validate your assumptions at all stages.
- **Target variables** are what your model seeks to measure. Your choice of target variables can also have implications for whether or not your model’s outcomes are likely to be unfair. This can happen if the

target variable is a proxy for protected characteristics (because it is correlated with them). Or, if it is measured less accurately for certain groups, reflecting measurement bias (see later section). For example, predictive policing models may rely on past arrest records which may be racially biased. You should therefore examine your target variable's validity.

- **Measurability bias** is the tendency to over-rely on quantifiable proxies rather than qualitative ones. This can lead to omitting useful features from your model's design, or embedding unwanted bias by using unreliable proxies just because they are easily available. Effective problem formulation is about ensuring what you plan to measure captures the nuance and context of the actual problem.

When you formulate the problem, you should also think about how a decision-support system will interact with human decision-makers and reviewers. This is because the system's predictions may not be sufficient for them to make the most informed decision. How you formulate a problem and what limitations your formulation has may have an impact on how effective human oversight is in the final system. Humans need to be aware of these limitations.

Further reading outside this guidance

For more on target variables and their effects on fairness in the US context see '[Big data's disparate impact](#)'

For more on the link between the problem formulation stage and fairness see '[Problem formulation and fairness](#)'

Who are the impacted groups and individuals?

From a fairness perspective it is important that you are able to explain why your AI system is applied to specific groups of individuals and not others. This guidance refers to these as 'impacted groups'.

You may intend to apply your model to particular groups. However, you must also consider whether your system may influence other groups indirectly. For example, an AI system managing childcare benefits does not only impact the claimants but the children under their custody as well.

Also, when thinking about impacted groups, you must consider the possibility that because of their different contexts not all individuals in the group will be impacted in the same way.

Engage with individuals your system affects

To ensure the development and deployment of your AI system is fair, you should engage with people with lived experience relevant to your use case or their representatives, such as trade unions. Article 35(9) of the UK GDPR states that organisations, depending on their context, must seek the views of individuals or their representatives on the intended processing as part of a DPIA. Lived experience testimonies may challenge the assumptions made during your problem formulation stage. But this can help you to manage costly downstream changes, as well as mitigating risk.

For example, this kind of input can help you identify and understand how marginalisation affects the groups your system may impact. In turn, this may lead to design changes during the development process which ensure your system has a better chance of delivering fair outcomes when you deploy it. Setting up a way to incorporate this feedback post-deployment, as part of your monitoring measures, aligns with data

protection by design.

Further reading

For more information, read the subsection on culture, diversity and engagement with stakeholders in [“How should we manage competing interests when assessing AI-related risks?”](#)

You could adopt a participatory design approach

Independent domain expertise and lived experience testimony will help you identify and address fairness risks. This includes relative disadvantage and real-world societal biases that may otherwise appear in your datasets and consequently your AI outputs over time.

This approach is known as “participatory design” and is increasingly important to AI systems. It can include citizens’ juries, community engagement, focus groups or other methods. It is particularly important if AI systems are deployed rapidly across different contexts, creating risks for a system that may be fairness compliant in its country of origin to be non-compliant in the UK for instance.

Participatory design can also help you identify individuals’ reasonable expectations, by discussing available options with their representatives.

For engagement with impacted groups to be meaningful, they or their representatives need to know about:

- the possibilities and limitations of your AI system; and
- the risks inherent in the dynamic nature of the environments you deploy it in.

If you decide to adopt a participatory design approach, you should seek to validate your design choices with these stakeholders and reconcile any value conflicts or competing interests. You could also view participatory design as something you embed into your AI project lifecycle. This will help you refine your models based on actual experience over time.

Consult independent domain experts

Your organisation as a whole may not have the necessary knowledge of what marginalisation means in all possible contexts in which you may use your AI system. It is important that key decision-makers at management level acknowledge this. Seeking and incorporating advice from independent domain experts is good practice. Article 35(9) of the GDPR supports this approach.

Determining the relevant domain expertise depends on the circumstances in which you develop and deploy your AI system, along with the impacted groups it relates to.

Your senior decision-makers should invest appropriate resources and embed domain experts in the AI pipeline. This will make it easier to avoid unfairness downstream in the AI lifecycle as it will ensure your developers can draw on that expertise at various points of the production.

Additionally, decision-making processes that can lead to unfair outcomes can be sector specific. Depending on the context, multidisciplinary expertise from social sciences, such as sociology or ethnography, could be useful to understand how humans interact with AI systems on the ground.

Further reading outside this guidance

For more information on challenges AI developers face in identifying appropriate performance metrics and affected demographic groups see [‘Assessing the fairness of AI systems: AI practitioners’ processes, challenges, and needs for support’](#)

2. Before you process personal data

You should determine and document your approach to bias and discrimination mitigation from the very beginning of any AI application lifecycle. You should take into account and put in place the appropriate safeguards, technical and organisational measures during the design and build phase.

Consider what algorithmic fairness approaches are appropriate for your use case

Your choice of algorithmic fairness metrics should relate to your context, objectives and what your distribution of outcomes should reflect. For example, in the case of granting a loan, if you chose equality of opportunity as the metric you test your system against. This would mean that someone’s chance of being predicted to repay their loan, given that they in fact will go on to repay it, is the same regardless of their group. Equalised odds, on the other hand, is even stricter. It also requires that someone’s chance of being incorrectly predicted to default on their loan, when they will in fact go on to repay it, is the same regardless of their group.

It is good practice to identify the algorithmic fairness approach you have chosen and why., Depending on the context, you may also be required to do so. Our guidance on outcome fairness can help you be transparent about it to individuals.

You can apply algorithmic fairness approaches at different stages of your lifecycle. You can undertake **pre-processing** bias mitigation, for example by removing examples from your training dataset that you suspect may lead to discrimination. You can seek to mitigate bias **in-processing** by changing the learning process of your model during training so it incorporates particular algorithmic fairness metrics. You can also modify the model after the initial training, at the **post-processing** stage.

In the real world algorithmic fairness metrics eventually relate to the distribution of resources or opportunities. However, ensuring fair outcomes is not always dependent on distributions. Instead, this may require you to consider more qualitative and contextual aspects, such as how human rights are affected or under what terms a job applicant is invited for an interview. At the same time, in certain circumstances unequal distribution may be justifiable. For example, it may be preferable to distribute aid ‘unequally’ in favour of those who need it most.

You should document and justify how you make decisions about distributing resources or opportunities. Algorithmic fairness metrics are not always mathematically compatible with each other. Therefore, you may not be able to fulfil each metric you choose at the same time. As a result, you should clearly analyse how and why the metrics you choose serve your specific objectives. As well as any trade-offs you had to make to safeguard the rights, freedoms and interests of individuals.

Being able to demonstrate why you chose one statistical notion of algorithmic fairness over another can also assist you with your wider accountability obligations.

It is good practice to clearly document your algorithmic fairness objective and the relative disadvantage it is seeking to address.

If you use algorithmic fairness metrics to test your system, you should consider how they account for or interact with broader unfairness risks inherent in the AI lifecycle. For example, the context of your system's deployment, the groups it impacts, or the input of human reviewers.

Whatever approach you choose, it is important to note that algorithmic fairness metrics do not comprehensively help you to demonstrate that your AI system is fair. This is because of the variety of other factors involved in the entire decision-making process, as well as the often dynamic environment your system operates in. For example, a Live Facial Recognition (LFR) technology system deployed by a private actor in a public space. It processes indiscriminately the personal data of thousands of individuals per day without the necessary safeguards or a DPIA. This is not going to be lawful or fair just because it does not display statistical accuracy variations across groups. Crucially, data protection asks not only how you process personal data but also if you should process it in the first place.

Finally, when you consider your algorithmic fairness approaches, it is also important that you take into account compliance with the Equality Act 2010. This guidance does not cover any of your obligations under equality law. To ensure compliance with equality law, you should refer to guidance produced by the Equality and Human Rights Commission.

Supplementary reading in this guidance

[How should we manage competing interests when assessing AI-related risks?](#)

Further reading outside this guidance

A number of toolkits, questionnaires and frameworks have been proposed in order to examine algorithmic fairness and how to negotiate its inherent trade-offs. For more on see '[Risk identification questionnaire for detecting unintended bias in the machine learning development lifecycle](#)' and '[The landscape and gaps in open source fairness toolkits](#)'

What are the limitations of algorithmic fairness approaches?

Algorithmic fairness approaches have several limitations that you should consider when evaluating their efficiency or appropriateness for a specific use case.

- Fairness questions are highly contextual. They often require human deliberation and cannot always be addressed by automated means.
- They are designed to compare between a single category (eg race). As a result, algorithmic fairness metrics may struggle to address issues of intersectional discrimination. This discrimination manifests at the intersection of two or more protected groups without necessarily being visible when comparing pairs of groups one-by-one.
- They may require you to put additional safeguards in place to protect individual rights in accordance with data protection. For example, if you decide to use special category data or protected characteristics

to test for algorithmic fairness.

- Some approaches rely on false positives or false negatives. These errors may have different implications for different groups, which influences outcome fairness.

3. Data collection

Take a thoughtful approach to data collection

When you use AI to process personal data, you take decisions about what data to include and why. This applies whether you collect the data from individuals directly, or whether you obtain it from elsewhere.

The purpose limitation and data minimisation principles mean you have to be clear about the personal data you need to achieve your purpose, and only collect that data. Your data should be adequate for your intended purpose.

This also helps to ensure your processing is fair. In the AI context, data is not necessarily an objective interpretation of the world. Depending on the sampling method, datasets tend to mirror a limited aspect of the world and they may reflect outcomes or aspects born out of historical prejudices, stereotypes or inequalities. Awareness of the historical context is crucial for understanding the origin of unfairness in datasets.

This is why you should examine what data you need to collect in relation to your purposes.

Depending on your purpose, you should examine which of the underlying dynamics of the observed environment you want to retain from within the data you have collected, and which you want to dispose of.

To ensure your processing and outcomes are fair, your datasets should be accurate, complete and representative of the purpose of the processing. We have provided [guidance on dataset fairness](#) which can assist with your transparency compliance requirements as well. This will also help you to comply with the data minimisation and storage limitation principles, and improve business efficiency.

You should not assume that collecting **more** data is an effective substitute for collecting **better** data. Collecting personal data because it is convenient to do so, or because you may find a use for it in the future, does not make that processing fair or lawful. Data protection law requires you to process only the data you need to achieve your stated purpose.

Further reading in ICO guidance

[Dataset fairness](#) (Explaining decisions made with AI)

Further reading outside this guidance

For a discussion on methodologies for data collection see '[Lessons from archives: Strategies for collecting sociocultural data in machine learning](#)'

Assess risks of bias in the data you collect

Bias can arise from different sources. One of these includes the data you originally collect. It is important to carefully assess your datasets instead of just obtaining them from various sources and using them uncritically. This is so you can avoid inadvertently incorporating any underlying unwanted bias they may have.

Therefore, if you obtain datasets from other organisations (eg as a result of a data sharing process), you need to assess the risks of bias in the same way as you would if you collect the data from individuals yourself.

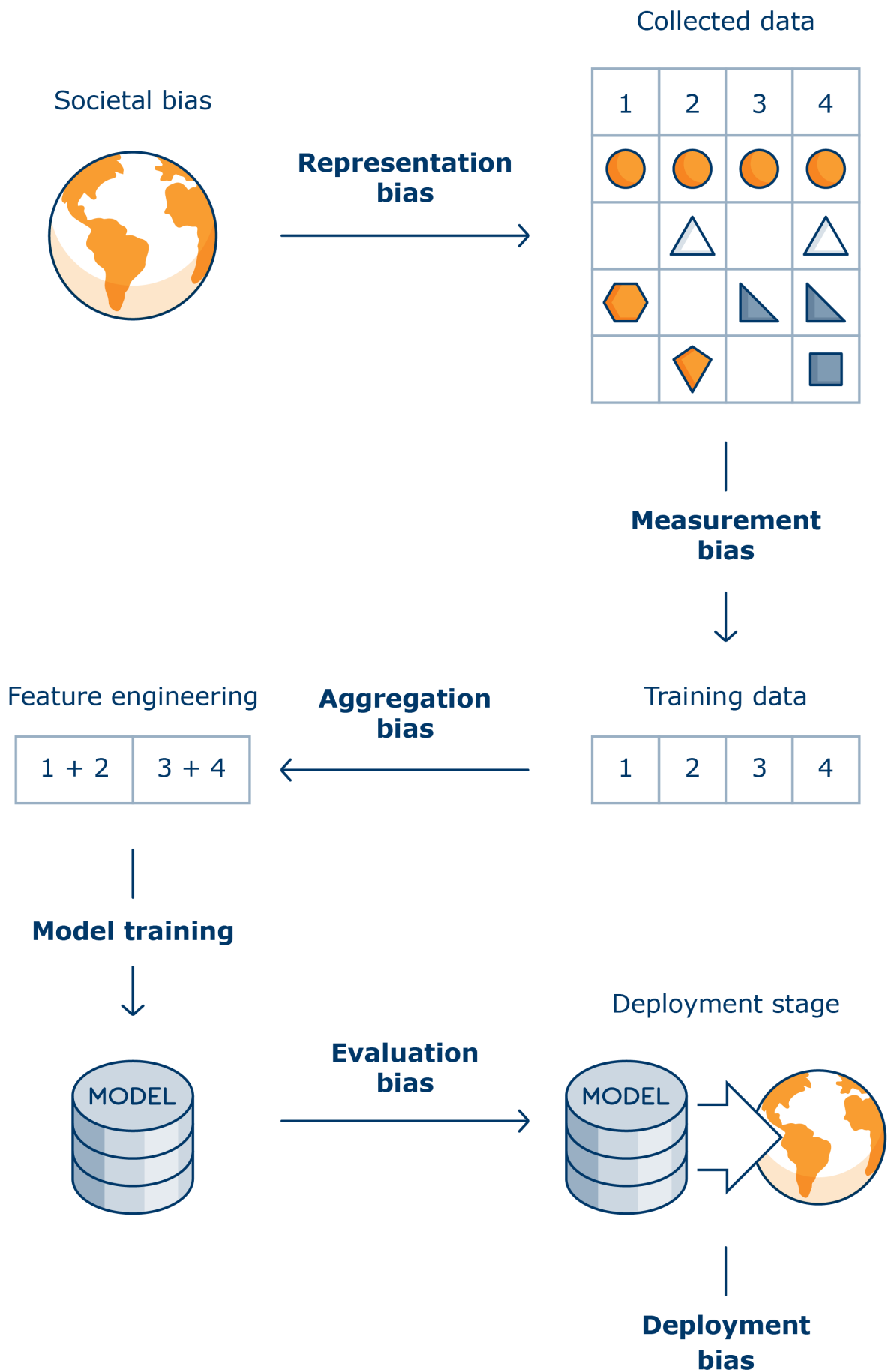
For example, the context in which the data was initially processed may reflect selection (sampling and measurement) biases that are not representative of the population you intend to apply your AI system to.

It is useful to ask your dataset suppliers for information on data collection methodologies and any biases their datasets might have encoded.

You should take into account two main categories of data-driven bias. These are **statistical** and **societal biases**. These categories inter-relate so you may find it difficult to consider them separately. Domain expertise can help you identify statistical or societal biases that your datasets may include, and plan your mitigation measures appropriately. In general, identifying the potential sources of bias at an early stage will help you design more effective mitigation approaches.

Later stages of the AI lifecycle may present additional types of bias, such as **evaluation bias** and **emergent bias**.

Potential sources of bias across the AI lifecycle:



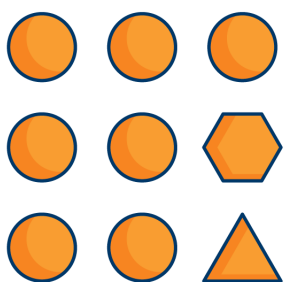
What is statistical bias?

Statistical bias is as an umbrella term for the following:

• **Representation or sampling bias.** This can occur when your dataset is imbalanced. This means your training dataset is not representative of the population that you will eventually apply your trained model to. Representation bias can result from sampling methods that are inappropriate for the stated purpose, changes in the sampling population or the fact important aspects of the problem are not observed in the training data. For example:

- you never get to observe what happens to people whose credit application was denied or fraud cases that were never detected;
- various image datasets used to train computer vision systems have been found to exhibit representation bias; and
- classification systems can perform poorly when encountering cases not present in the training data.

Collected data



Representation/sampling bias

-
- **Measurement bias.** This is about the impact of selected features and labels on model performance. It can be caused by differences in the accuracy of the process through which the selected features are measured between groups. For example, attempts to geolocate patterns of fraud could create unintended correlations with particular racial groups. Additionally, depending on the context, the selection of those features and labels may oversimplify the target variable you aim to capture and omit important nuances. For example, if a model predicts the risk of domestic abuse without factoring in an examination or interview of the alleged offenders, the estimated risk may be downplayed or exaggerated.

Data labelling



- **Aggregation bias.** Aggregation bias arises when a one-size-fits-all model is used for data in which there are underlying groups or types of examples that should be considered differently. Aggregation bias can also take place at the feature engineering stage.

Collating a 'triangle' dataset



Further reading outside this guidance

For more on the sources of bias see ['A framework for understanding sources of harm throughout the machine learning life cycle'](#)

For more on statistical and societal bias see ['Algorithmic fairness: choices, assumptions, and definitions'](#)

What is societal bias?

Societal bias is often referred to as **historical or structural bias** and relates to structural inequalities inherent in society that the data then reflects. In general, predictive AI models are trained to identify patterns in datasets and reproduce them. This means that there is an inherent risk that the models replicate any societal bias in those datasets. For example, if a model uses an engineering degree as a feature, it may replicate pre-existing societal bias arising from the historical level of male representation in that field.

Societal bias can also be institutional or interpersonal. For example, if human recruiters have systematically undermarked individuals from specific backgrounds, an AI system may reflect this bias in its labelling process. In general, it may be particularly risky to deploy AI-driven systems in environments that are known to have significant societal bias.

Consulting domain experts and people with lived experience will enable you to identify societal bias, explore its root causes and design appropriate mitigation measures. For example, consultation in this way may help you identify and mitigate risks resulting from the inappropriate labelling of training data. These can include:

- offering cognitive bias training for these staff members;
- updating your labelling protocol; or
- updating your monitoring process to ensure the protocol is being followed.



Societal bias (also referred to as historical or structural bias)

Data labelling

In order to establish a reliable [ground truth](#), it is important that you appropriately annotate your datasets. Having clear criteria for doing this helps you ensure fairness and accuracy during both the acquisition and preparation stages.

Labels can pose fairness risks. This is because they can reflect inaccurate representations of the real world or create an overly-simplistic view of something that is more complex. For example, attempting to label human emotions is unlikely to capture their full complexity, or the context in which they occur. Without

mitigation, this could lead to unexpected or unfair impacts in certain use cases (eg emotion recognition systems).

Additionally, misrepresentation or underrepresentation of certain data points can result in risks of harm, such as:

- underrepresentation in labelled data points, leading to harms related to the allocation of opportunities or resources. The impact of allocative decisions may be significant (eg loss of life or livelihood); and
- misrepresentation in labels leading to harms related to the reinforcing of biases or stereotypes. The impact of representative harms could include unfavourable bias for marginalised groups.

Training your staff about implicit bias and how it may impact their decisions is one way of mitigating these risks. Including community groups and impacted individuals in the labelling process is another, and is an example of participatory design in practice. You can then formulate your labelling criteria and protocol as a result.

In general, your training data labels should comply with data protection's accuracy principle. You should clearly tag output data as inferences and predictions, and not claim it to be factual. You should document clear criteria and lines of accountability for labelling data.

Further reading in this guidance

[What do we need to know about accuracy and statistical accuracy?](#)

4. Data analysis and pre-processing

Pre-processing in the AI lifecycle refers to the different processing techniques you apply to the data itself before you feed it into your algorithm. Depending on the circumstances, datasets are likely to be split between training, testing and validation data.

For data protection purposes, it is important to note that "processing" personal data includes the process of collecting that data. This means that processing in the data protection sense begins **before** the 'pre-processing' stage of the AI lifecycle.

If, depending on your context, you need to use bias mitigation techniques using algorithmic fairness metrics, you should track the provenance, development, and use of your training datasets. You should also detect any biases that may be encoded into the personal data you collect or procure. For example, if you sample uniformly from the training dataset, this may result in disparities for minorities that are under-represented within the data. Be mindful that common practices like imputing missing values can introduce biases that may result in unfairness.

To ensure fairness, you may decide not to use certain data sources or features to make decisions about people. For example, when using them may lead to direct or indirect discrimination. In these cases, you should assess, document and record the sources or features that you do not intend to use.

How should we approach bias mitigation at the pre-processing stage?

A number of pre-processing methods, including toolkits and algorithms exist that aim to remove bias from

training datasets by altering their statistical properties. That means you need to have access to the training data to use these methods. Also, if you are procuring your model from someone else, you may need to use post-processing techniques as the latter do not require that access. If such techniques are not adequate in your context, you may need to negotiate with your provider to define the terms of your contract. This will ensure measures are in place to adequately address unfairness risks.

You should bear in mind that the efficacy and the limitations of all these algorithmic fairness approaches are an ongoing area of research.

If your collected data reflects societal or statistical bias, you should consider modifying the training data, to remove examples of data that you suspect can lead to discrimination. If your dataset is imbalanced, you may need to collect more data on underrepresented groups. These datasets need to be able to also capture in-group variation to avoid aggregation bias.

Other pre-processing approaches include reweighing your data points or changing your labels. Depending on your selected algorithmic fairness metric, you can change the labels for the groups most likely to be vulnerable to discriminatory patterns.

You should also consider other techniques at the pre-processing stage. For example, data visualisations. These can observe changes in the distribution of key features in order for you to take appropriate actions. Creating histograms and correlations matrices to observe your training data's distribution patterns and how these change over time can provide insights about potential bias in that dataset.

5. Model development

ML models learn patterns encoded in the training data. They aim to generalise these patterns and apply those generalisations to unseen data. At the model development stage, you are likely to consider and test different algorithms, features and [hyperparameters](#).

We have provided [guidance about design fairness](#), that looks into how you can explain how your design decisions serve fairness goals to individuals.

Model selection

Your choice of model depends on the complexity of your problem and the quantity and quality of the data you have or need to process to address it. You need to consider the balance between your model's [inductive bias](#) and its variance, to ensure it is generalisable and therefore statistically accurate. Variance is the model's sensitivity to fluctuations in the data, meaning how often it registers noise or trivial features as important.

You should also consider your transparency obligations when selecting your model. For example, in the context of 'black-box' models, the lack of access to how a system arrived at a decision may obstruct individuals' ability to contest one they deem unfair. Additionally, in the case of solely automated decisions with legal or similarly significant effects, being deprived of access to the 'logic involved' would mean being deprived of their information rights. This would go against their reasonable expectations.

Further reading in this guidance

[How you can address risks of interpretability](#)

Further reading – ICO guidance

[Design fairness](#) (Explaining decisions made with AI)

What do you need to consider during feature engineering?

Feature engineering is a core stage in developing your model and transforming data from the original form into features for training your model. This forms part of the process of developing an AI model and involves concepts such as augmentation, aggregation and summarisation of the data. Feature engineering therefore can involve a set of processing operations such as the adaptation or alteration of personal data. This ensures that these new features are more amenable to modelling than the original data.

Feature engineering can mitigate risks of statistical bias, such as aggregation bias. For example, transformation techniques like re-weighting feature values in the training data to reduce the likelihood of discriminatory effects.

However, you should recognise that feature engineering may also **introduce** risks of statistical bias, specifically aggregation bias. For example, if you do not apply transformation techniques correctly, they could also lead to aggregation bias and therefore not solve the issue you are trying to address. It is therefore important that you carefully assess the impact of the technique to ensure it achieves your desired effect.

Example

Individuals who follow the Catholic, Protestant or Orthodox doctrines are grouped under one “Christian” data feature.

This may fail to account for their distinctions, leading to unexpected results, depending on the context. This is an example of how feature engineering can lead to unfair results.

Feature engineering can also give rise to measurement bias. This can happen when there is a difference in how inputs or target variables are measured between groups. For example, the fractions of the population that have computable records that will be used as features in a model may vary by race. This will affect how accurately those features are measured. For example, that may be the case when specific groups have historically less access to specific services. Deriving unbiased learnings from data collected in the real world is challenging because of unobvious correlations that exist within that data. This nevertheless can impact your model’s performance.

When transforming personal data to create the features, you must consider whether any features are proxies of protected characteristics or special category data. You can establish this through a “proxy analysis”. For example, if your model detects correlations between creditworthiness and correct capitalisation in loan applications, it may penalise people with dyslexia. Once your proxy analysis identifies this, you may be able to remove or adjust the feature to avoid these correlations.

Model optimisation

In ML, model optimisation typically intends to minimise prediction errors. When using algorithmic fairness approaches, you will typically include a fairness metric as another criteria alongside the objective of minimising prediction errors. This is called [‘multi-criteria optimisation’](#).

You can also intervene in the testing stage, where a held-out sample of the data (the test data) is used to test the generalisability of the model. At this point, it is possible to assign some cost or weight to the underrepresented group to ensure fairness in the prediction.

If you decide to test your model against one of the open source benchmark datasets, you should take into account any biases they contain.

Model design and testing has inherent trade-offs that you need to assess. For example, hyperparameter tuning - the process of identifying the optimal hyperparameters for your model. Different tunings may present distinct trade-offs between statistical accuracy and algorithmic fairness. If you have a clear idea of the context in which you will use your system, then you are better able to assess these trade-offs.

Model optimisation is not straightforward. However, you should consider individuals’ reasonable expectations when you address it. For example:

- there may be implicit assumptions in the decisions you make about the features that are relevant to your problem; and
- how you weigh these features may lead to one or more having a higher importance or prioritisation, and this may not be justifiable to the individuals concerned. For example, job applicants may expect an AI system used for sorting job applications to register recent work experience as of more relevance than a job they held one decade ago and weigh that variable accordingly.

In adaptive models, what you are optimizing for may have downstream consequences for fairness. A reinforcement learning system that explores the relationship of actions and reward functions may lead to unpredictable results when individuals or their personal data inform that learning process. For example, an ad delivery algorithm that is optimised to surface content to people that past data indicates are more likely to click on it. This may set in motion a self-reinforcing feedback loop that entrenches existing inequalities. For example, surfacing ads for computer engineering jobs predominantly to 90% men and 10 % women on the basis that historical data demonstrates that uneven distribution. This is likely to perpetuate and entrench that unequal access to certain professions.

You may need to properly evaluate how to rebalance the current distribution your historical data demonstrates with the optimal or ideal one you are aiming for. It may be useful for your training phase to engage with community groups and impacted individuals. Especially when labelled data is unavailable or when you are dealing with unsupervised ML systems. This could confirm whether the patterns and groupings that are emerging reflect the ground truth as per their knowledge and expectations.

Further reading outside this guidance

For a discussion around key questions AI developers need to ask during training and testing see '[Hard choices in artificial intelligence: addressing normative uncertainty through sociotechnical commitments](#)'

Overfitting and underfitting

Overfitting is where your model pays too much attention to the details of the training data. Essentially, the model remembers particular examples from the training data rather than just the underlying patterns. This can happen if it includes too many features. This potentially raises data minimisation questions. Or, if there are too few examples in the training data (or both).

Underfitting takes place when your model fails to capture a phenomenon in the training data.

Both underfitting and overfitting can result in statistical inaccuracies affecting individuals' reasonable expectations.

In order to address these risks, you should:

- examine the data's context (particularly whether there is appropriate representation of groups in your training data);
- evaluate the values you assign to your features before you feed them into the model;
- tweak your model by tuning its hyperparameters; and
- fit the most appropriate algorithm to the data. Your algorithm may not be the appropriate one, so you can test other algorithms and their performance.

Exploration v exploitation trade-off

If your AI system is adaptive and dynamic, you should assess any trade-off between exploration and exploitation.

AI systems generally operate by detecting historical correlations between inputs and outcomes. They then apply these to new data they receive. They assume the correlations still apply, and that the model is appropriately generalisable.

Exploitation is where a system makes optimal near-term decisions based on the existing information it has. You also need to be aware that over-reliance on historical data can impose constraints on what your model will predict that may be inappropriate for your specific context.

A system that prioritises exploitation risks entrenching current skewed distributions that the data may contain. This may risk embedding the discrimination that these correlations reflect. For example, following the job targeting example, displaying an advert only to people who fit the group profile most likely to click on it. This is based on the assumption that the current status quo will hold indefinitely and can have unfair outcomes. Demonstrating engineering jobs predominantly to men because the model exploits the fact it knows the existing skewed distribution of men to women engineers, may have fairness implications.

Exploration is where a system attempts to find new ways to fulfil its reward function. It can provide unknown benefits, including better rewards. For example, a recommendation model may add random

suggestions into the content it serves instead of solely predicting the contents an individual wants to read. This means the reader may engage with that content more if they want novel information or less if they prefer to access predictable or familiar content.

This means exploration's benefits may come at the expense of the model not taking an action that is known to have a specific payoff. For example, a job targeting algorithm that incorporates a level of randomness by displaying an advert to people who do not fit into the group profile it has specified as the most likely to click on it. This may result in less online engagement, and potentially less revenue for the company selling the advertising space.

Clearly articulating your system's objectives and thoughtfully approaching this exploration-exploitation trade-off will help you mitigate unfairness risks. This is because it is essentially about deciding the balance between the status quo, as unequal as that may be, and the need to not let the past predetermine the future. For example, your goal may be to make your workforce more diverse than it has been in the past. Therefore, incorporating more exploration in your e-recruiting models may make sense.

In-processing bias mitigation

[In-processing](#) techniques for bias mitigation take place during the training stage when your training data are fed into the model. Examples include making sure the algorithm:

- incorporates regularisation (see [Glossary](#)) terms;
- imposes algorithmic fairness constraints; or
- incorporates selected metrics into its objective function.

These types of technical approaches are not a "one size fits all" solution to in-processing bias mitigation. For example, some research indicates that in some cases overfitting can happen. For example, fairness performance during training not becoming generalisable on unseen test data.

Researchers have noted that some bias mitigation approaches can use protected characteristics or special category data only during training, while others require them during live deployment as well.

Further reading outside this guidance

For more about how different bias mitigation techniques use protected characteristics see '[How could equality and data protection law shape AI fairness for people with disabilities?](#)'

6. Model evaluation

Model evaluation is the stage when you record and assess the performance metrics of your model.

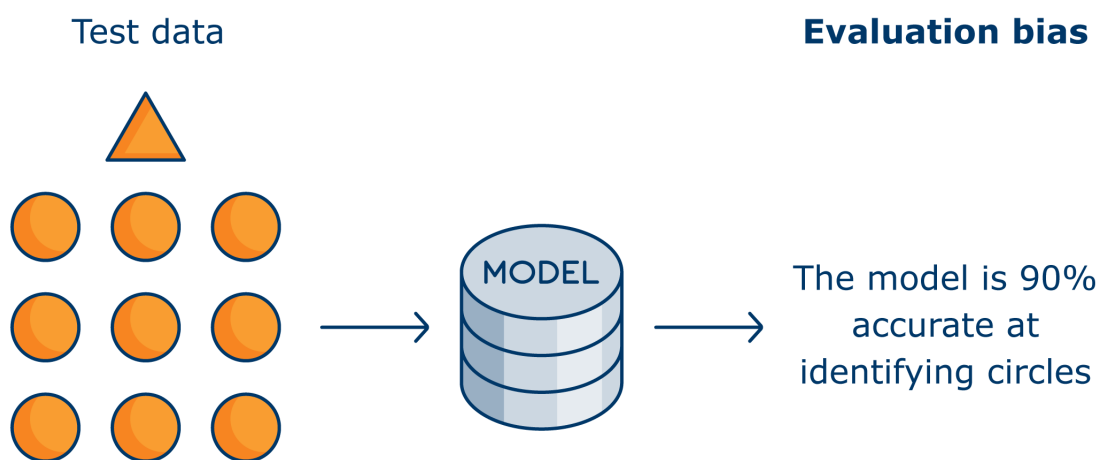
This stage can also suffer from evaluation bias if you use inappropriate performance metrics. For example, aggregate evaluation may not be appropriate to detect performance disparities between different demographic groups in the population. This may in turn give you the wrong idea about the statistical accuracy of your model. That is why you should conduct disaggregated evaluations of your AI systems.

Example

An LFR system has an aggregate error rate of 10%. This may obscure the fact that its model has a disproportionately poor performance in the context of minority groups.

Depending on the use case, collecting more data on the unrepresented group to reduce the disproportionate number of statistical errors they face is an option.

Time of collection is something to bear in mind for the validation data you are using. For example, data from four years ago may not represent the current context your system will be deployed in.



How should we evaluate our model's statistical accuracy?

You must take note of the trade-offs between precision (specificity) and recall (sensitivity). Your context will determine if you should prioritise one or the other. For example, if your system is used in a healthcare context for diagnosis, precision may be more appropriate to prioritise in order to avoid false negatives. Not all correct decisions are equally important and not all the wrong ones are equally detrimental.

Depending on your context, the false positive and false negative rates you choose to accept will reflect different levels of risk. For example, if approving a loan, the cost of a false positive (approving a loan for a lender likely to default) may exceed the profit of a true positive. This will influence where to set the decision boundary. To strike the balance that suits your use case you can:

- penalise certain types of errors more when specifying your [cost function](#) (see Glossary);
- tune your hyperparameters;
- increase your training dataset with new samples; or
- assign more weight to certain features over others.

How should we approach any trade-offs between statistical accuracy and algorithmic fairness?

Depending on your context, model evaluation may involve a trade-off between:

- the extent to which the model is statistically accurate; and
- whether it meets certain algorithmic fairness metrics.

It is important to note that from a data protection fairness perspective you must ensure that your overall processing is fair. You cannot simply trade off fairness in favour of statistical accuracy. You need to consider whether statistical accuracy in your context might still be unfair. For example, your system may accurately reproduce discriminatory patterns that in practice you want to remove.

The way you deploy your system may also lead to unjustifiably unfair outcomes. This may be the case even if it is statistically accurate if it causes individuals to be unexpectedly excluded from services and markets, or violates aspects of other legal frameworks, such as equality law. For example, risk prediction models for car insurance premiums may be more statistically accurate, leading to adjusted personalised prices. But this may make services less accessible to price sensitive individuals. These are people whose demand for a service or product is likely to change when their cost changes.

Further reading in this guidance

[What do we need to know about accuracy and statistical accuracy?](#)

What are the appropriate safeguards?

You should consider safeguards for how you apply your AI system to subgroups of the population.

If you embed algorithmic fairness constraints into your model or test it post-development, you should examine whether your safeguards protect individuals as well as groups. If your system processes personal data fairly about most people but unfairly for one individual, your processing will still infringe the fairness principle. Identifying what makes individuals vulnerable will also assist you in determining the appropriate safeguards.

You need to pay particular attention to “edge cases” or outliers. For example, cases where your ML model makes incorrect predictions or classifications because you have not trained it using sufficient data about similar individuals.

Generalisation and outliers

AI models are designed to generalise. This means they are intended to be sufficiently accurate when faced with unseen circumstances. Nevertheless, systems are less accurate for outliers, as by definition they represent a minority in the training data, making them more vulnerable to risks.

Your senior decision-makers should take into account that models may struggle to identify patterns for groups that are not sufficiently represented in the training data. This should help you determine whether the use of AI is the best solution in the first place or what safeguards are necessary. You should not treat individuals unfairly just because they constitute outliers.

What is the role of disaggregated evaluations?

Algorithmic fairness metrics can assess model performance at aggregate level. However, depending on the population you intend to deploy an AI system to, you may also need to assess its performance at both group and subgroup level. Disaggregated evaluations enable you to do this. This is when your system's performance is evaluated separately for different groups of people, such as those based on protected characteristics or special category data.

Additionally, algorithmic fairness approaches may not sufficiently capture issues of intersectional discrimination. This is when an individual is discriminated against on more than one ground in the same context. For example, discrimination experienced by black women on the basis of race and sex. Disaggregated evaluations for people that may be exposed to intersectoral discrimination may help you identify this risk.

Disaggregated evaluations can also help you explore the utility or the cost of different AI decisions for different groups. In turn, this can help you set the appropriate decision boundary for your system. For example, the most marginalised individuals within impacted groups may be at a higher risk of AI-driven harms, such as discrimination. However, you may also need to consider that vulnerable groups may not neatly fall under one of the protected characteristics or special category labels.

You could conduct disaggregated evaluations and keep a record of disparities in outcomes detected. This is not just to mitigate them, but to assist your senior management in understanding and addressing these issues appropriately and promptly.

If you procure AI models, you should seek assurances from the AI vendors about any bias testing they conducted on them or test the models yourself. It is unlikely for the model you procure to have been trained on the exact population you intend to deploy it on. Therefore, applying algorithmic fairness testing to detect discrepancy in statistical accuracy or distribution effects could be useful.

Further reading outside this guidance

Read the "[Fair ML book](#)" for more information about research demonstrating particularly high error rates in for minorities in ML systems. (PDF, external link)

For more on disaggregated evaluations see '[Designing disaggregated evaluations of AI systems: choices, considerations, and tradeoffs](#)'

Researchers have indicated AI systems may give rise to new forms of discrimination. For more see '[Protected grounds and the system of non-discrimination law in the context of algorithmic decision-making and artificial intelligence](#)'

Post-processing bias mitigation

A number of post-processing techniques are available for after the model has been built. For example, if you discover bias in your model, you can test other models on your testing data and combine the outputs of all models. You can do this with a view of harmonising the different bias tendencies within them depending on your use case and its constraints. You may also need to go back into your training stage and redefine your target variable or change the labels in your input data. As mentioned before, the AI lifecycle

is not a linear process but iterative.

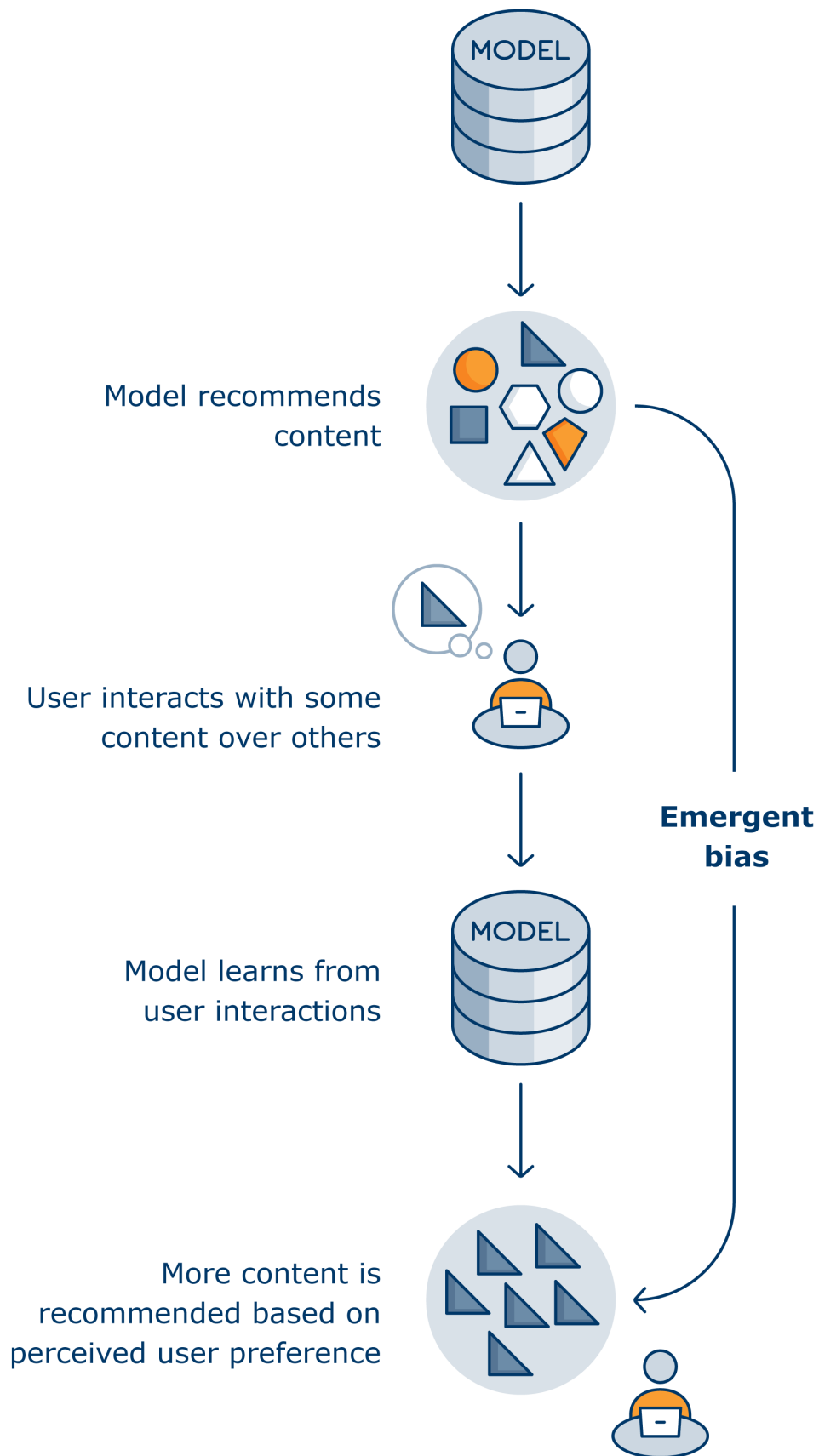
7. Deployment and monitoring

Ongoing monitoring

After deployment you should monitor your AI system's performance across groups. The frequency of the monitoring should be proportional to the impact of incorrect outputs on individuals and vulnerable groups. Depending on your context, you may want to monitor your model's performance on the basis of features that may be proxies for protected characteristics or marginalised and vulnerable groups. For example, postcodes or features with encoded historical biases.

You also need to keep your DPIA **under regular review**.

AI models are not applied in a static, unchanging world. Changes in the environment can also change its impact, including the individuals on which your system is deployed on or their behaviour. This can lead to **emergent bias** which you must take into account.



For example, in the case of reinforcement learning systems, feedback loops can give rise to biases that you may not have detected in the development and model validation stages. This can result from cultural biases on the side of the users (eg social media users favouring sensationalist content leading an AI system to surface it more), or the humans whose decisions the AI system supports. In the latter case, these biases can then be fed back into the model as inputs. That is why you need to monitor your AI system on an ongoing basis and track its performance throughout its deployment. For example, by using algorithmic fairness to account for distributions and statistical accuracy metrics.

Depending on your context and how people (human reviewers or data subjects) interact with your AI system, conducting user testing prior and during deployment could be useful.

You must also have a transparent and simple-to-use mechanism that allows impacted individuals and groups to contest decisions they deem unfair. This also enables you to consider whether a model needs retraining. It's good practice to record and monitor the number of individuals' challenges to your AI system's decision on fairness grounds. This will help you to evaluate your data protection compliance risk.

AI systems often enable you to process personal data on a large scale and at speed. Therefore, your processes to address the risk of unfair outcomes should be designed so that you can deal with these issues in a timely manner.

For example, an unfair outcome, and any second-order adverse effect arising from it, may spread quickly if you do not react in an appropriate timescale. The role of human review in mitigate this risk is also crucial.

Supplementary reading in this guidance

- [AI and data protection risk toolkit](#)
- [Outsourcing and third-party AI systems](#)

Further reading outside this guidance

For more on emergent bias see '[A broader view on bias in automated decision-making: reflecting on epistemology and dynamics](#)'

Human review as a fairness safeguard

Individuals relying on AI outputs to make decisions may not always be able to evaluate the statistical accuracy of those AI-driven predictions. This means that even when AI outputs inform rather than determine decisions, human reviewers' judgement of them may be inaccurate.

Conversely, [automation bias](#) (see Glossary) may lead your human reviewers to overestimate the credibility of an AI system, even when it is statistically inaccurate. See the section on [the role of human oversight for more information](#).

You should monitor the statistical accuracy of "hybrid" AI systems that involve human and AI cooperation. You should use the outcomes of this monitoring to improve your system or decide to withdraw it, if

unmitigable high risks arise after deployment.

You should also monitor the impact human reviewers have on the performance of your AI model. For example, to mitigate the risk of them incorporating unwanted bias into the system through their choices, particularly if these result in harmful decisions. If you do not do this appropriately, you may end up with a false sense of security that your model is working as intended and all its outcomes are desirable.

Additionally, keep a record of when your human reviewers “override” the decision your AI system recommends or makes. This will enable you to make an informed evaluation of both your reviewers and your systems.

Supplementary reading in this guidance

[What is the role of human oversight?](#)

Further reading in ICO guidance

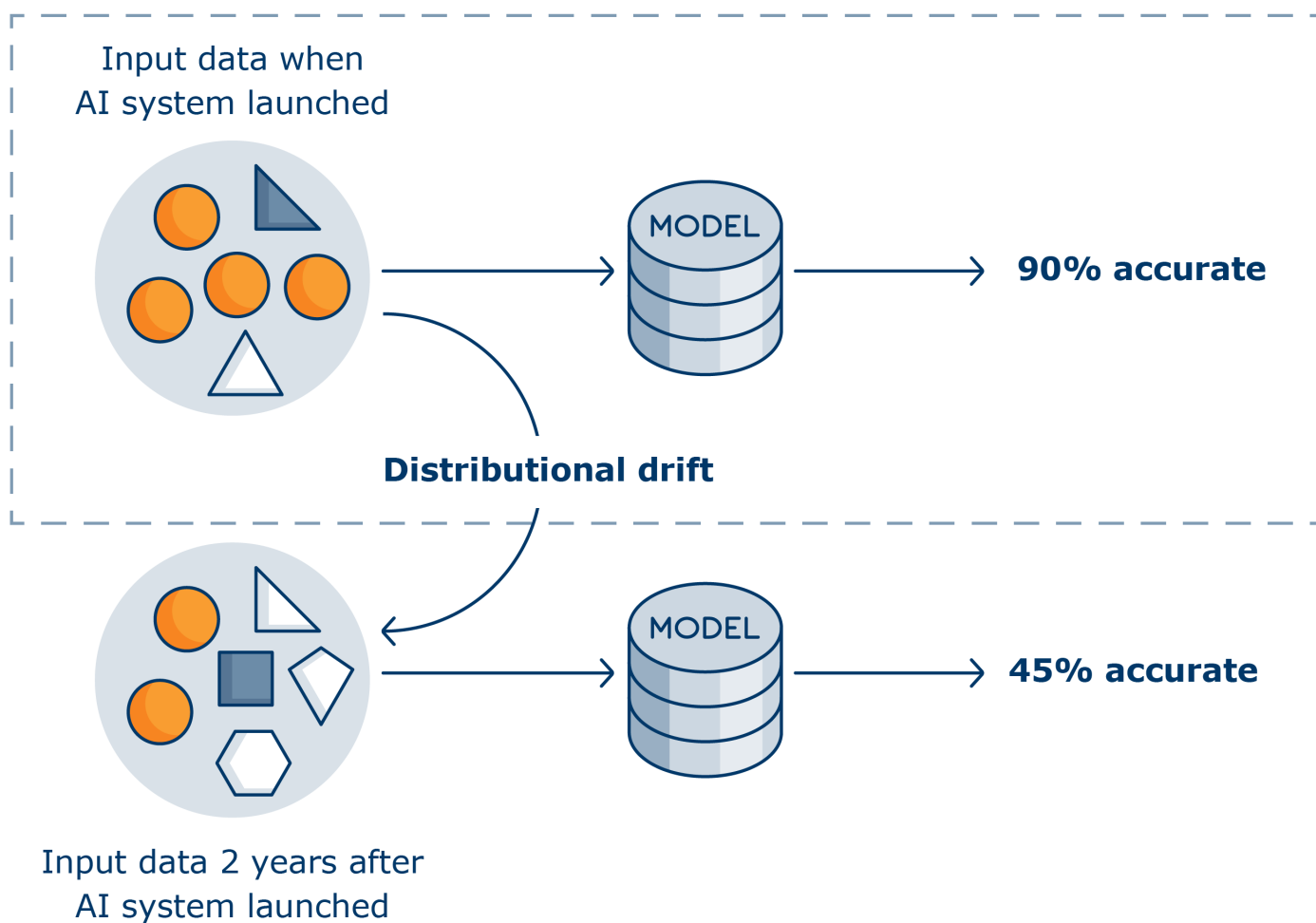
[Implementation fairness](#) (Explaining decisions made with AI)

Distributional drifts

Statistical accuracy is not a static measure. Ultimately, you deploy your AI system in changing populations or environments. As a result, the system may become less statistically accurate over time, representing distributional drifts.

There are three main types of distribution drift:

- Covariate drift, when the distribution of input data between the training and the live environment changes. This can lead to model misclassifications. For example, a speech recognition system may have been trained on data from English speakers from a particular area where a specific accent is prevalent. The system may become less statistically accurate when encountering increasing input data reflecting dialects or accents that represent a new distribution compared to the original training dataset. Covariate shift may be a sign the model cannot generalise adequately.
- Label drift, when the distribution of the target variables changes but the conditional distribution of features given the target remains the same. For example, an increasing ratio of the approval predictions to non-approval predictions in the context of loans may be an example of label drift.
- Concept drift, when the domain in which you use your AI system changes over time in unforeseen ways. This can lead to changes in the relationship between the features and the target variable, which may lead to the outputs becoming less statistically accurate. A specific kind of concept drift is an ontology drift, where the definition of classes in a categorical variable changes. An example would be the expansion of medical data taxonomies to include nonbinary genders and intersex identities.



For processing to be fair, it has to be sufficiently statistically accurate based on your context, and avoid discrimination. This is something that you can address by a robust risk management framework that records classification errors over time.

Supplementary reading in this guidance

[How should we define and prioritise different statistical accuracy measures?](#)

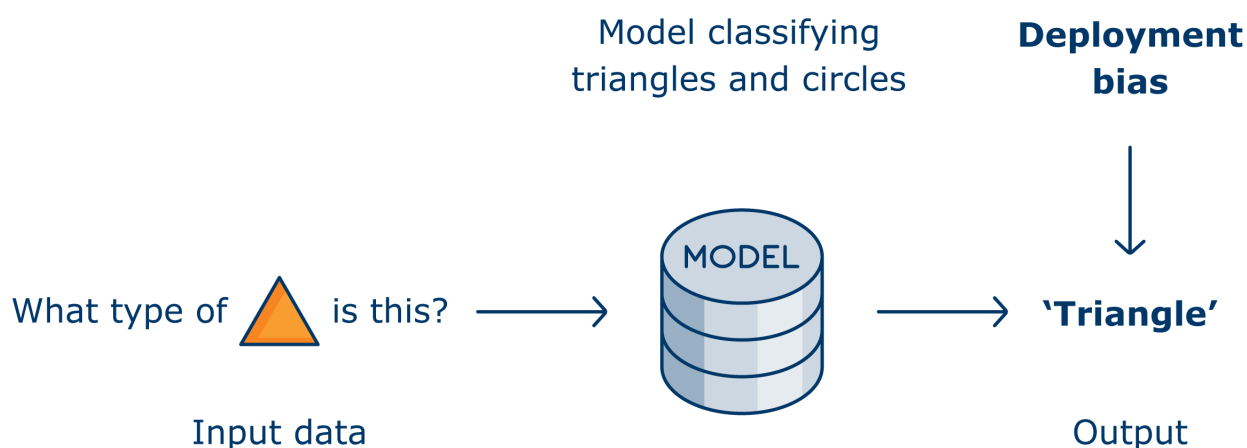
Function or scope creep

The term denotes the expansion of the functionality of a system beyond what it was originally created for. For example, a divergence of its initial purpose and actual use. Function creep can lead to **deployment bias**, where there is an inappropriate use of a model in a live environment.

Your training data needs to represent the population your AI model will be applied to. Therefore, deploying

an AI system that someone else created for a specific use in another domain can lead to statistical inaccuracies or discrimination.

Even when you deploy the model on the same population but for a different purpose, you must still comply with data protection. In particular the transparency principle and data subject's rights. Otherwise, you may also violate the purpose limitation principle.



8. Decommissioning or replacing an AI system

You should consider how you will deprecate the functionality and the backend infrastructure of your system, along with any personal data processed, if necessary. For example, you may need to decommission or replace the system, if it:

- ends up not meeting the benchmarks for its use;
- stops fulfilling its initial purpose; or
- has resulted in material or non-material harms.

In the case of mission-critical systems that cannot be allowed to fail, you need to have a transition or replacement plan in place. For example, systems processing interbank payments. You should plan for this at the design stage or as part of the development roadmaps. You should consider how you:

- incorporate decommissioning, including code, datasets, liabilities, risks, and responsibilities into your roadmap;
- take steps to erase or anonymise any personal data once or if you decide to withdraw the AI system; and
- ensure your decommissioning process is independently verifiable and auditable.

Further reading outside this guidance

The UK government's advice on retiring a service can be useful. See '[Retiring your service](#)'

Glossary

■ [Latest updates](#)

15 March 2023 - This is an old chapter with old and new content.

Affinity groups, algorithmic fairness, algorithmic fairness constraints, bias mitigation algorithm, causality, confidence interval, correlation, cost function, dataset labellers, decision, construct and observed space, decision boundary, decision tree, downstream effects, ground truth, inductive bias, in-processing, hyperparameters, multi-criteria optimisation, objective function, post-processing bias mitigation, regularisation, redundant encodings, reward function, use case, target variable, variance.

Term	Meaning
Affinity groups	Groups created on the basis of inferred interests rather than the personal traits of the individuals comprising them. They are also often described as “cohorts” or “ad-hoc groups”.
AI development tools	A service that allows clients to build and run their own models, with data they have chosen to process, but using the tools and infrastructure provided to them by a third-party.
AI prediction as a service	A service that provides live prediction and classification services to customers.
Algorithmic fairness	An active field of research that involves developing mathematical techniques to measure how ML models treat individuals from different groups in potentially discriminatory ways and reduce them.
Algorithmic fairness constraints	These are constraints you put in place while training a model in order to embed algorithmic fairness into its objective function.
Application Programming Interface (API)	A computing interface which defines interactions between multiple software intermediaries.
Automation bias	Where human users routinely rely on the output generated by a decision-support system and stop using their own judgement or stop questioning whether the output might be wrong.
Bias mitigation algorithms	Processes to remove unwanted bias in data or models.
Black box	A system, device or object that can be viewed in terms of its inputs and outputs, without any knowledge of its internal workings.
Black box attack	Where an attacker has the ability to query a model and observe the relationships between inputs and outputs but does not have access to the model itself.
Black box problem	The problem of explaining a decision made by an AI system, which can be understood by the average person.
Causality	The principle that one variable (X) - an independent variable - produces change in another variable (Y) which is called a dependent variable. To

establish causation, the two variables must be associated or correlated with each other and non-causal, 'spurious' explanations for the relationship must be eliminated. Depending on the context and because events in the real world are too complex to be explained just by one causal relationship, the principle of multiple causation needs to be considered, which says that a combination of causal relationships are more often than not in operation.

Confidence interval	A range of values that describes the uncertainty surrounding an estimate for an unknown parameter - the variable your AI system is trying to predict.
Concept/model drift	Where the domain in which an AI system is used changes over time in unforeseen ways leading to the outputs becoming less statistically accurate.
Constrained optimisation	A number of mathematical and computer science techniques that aim to find the optimal solutions for minimising trade-offs in AI systems.
Correlation	The relationship between two variables, where we can predict one variable from the other.
Cost function	An aspect of a learning process which attaches a cost to certain kinds of behaviours (eg errors) to help with achieving its objective function.
Dataset labellers	Individuals that label the training data so that an ML algorithm can learn from it.
Decision space, construct space and observed space	<p>These spaces denote levels of examining a problem:</p> <ul style="list-style-type: none">• the construct space relates to unobservable variables;• the observed space relates to observed features; and• the decision space is the set of actions available to a decision-maker.
Decision boundary	A threshold that separates data into different classes. For example, the boundary that separates loan applicants that will be rejected from those that will be accepted.
Decision tree	A model that uses inductive branching methods to split data into interrelated decision nodes which end in classifications or predictions. Decision trees move from starting 'root' nodes to terminal 'leaf' nodes, following a logical decision path that is determined by Boolean-like 'if-then' operators that are weighted through training.
Deep learning	A subset of machine learning where systems 'learn' to detect features that are not explicitly labelled in the data.
Differential privacy	A system for publicly sharing information about a dataset by describing the patterns of groups within the dataset while withholding information about individuals in the dataset.
Downstream effects	The impact(s) of an AI system on individuals once it is deployed.
False negative ('type II') error	When an AI system incorrectly labels cases as negative when they are positive.

False positive ('type I') error	When an AI system incorrect labels cases as positive when they are negative.
Feature selection	The process of selecting a subset of relevant features for in developing a model.
Federated learning	A technique which allows multiple different parties to train models on their own data ('local' models). They then combine some of the patterns that those models have identified into a single, more accurate 'global' model, without having to share any training data with each other.
Ground truth	At a high level, the reality a model intends to predict.
Inductive bias	The assumptions an algorithm is built on. It plays a role in the ability of a model to generalise when faced with new data.
In-processing	A series of techniques to intervene in the model during its training process, such as by adding additional constraints or regularization terms to its learning process.
Hyperparameters	Hyperparameters are configurations of parameters of an algorithm or learning process. Parameters are rules, constraints or assumptions that developers provide to an algorithm for training in order to deliver a functioning model for their specific use case. Hyperparameters are configurations of these parameters.
'K-nearest neighbours' (KNN) models	An approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are. KNN models contain some of the training data in the model itself.
Lack of interpretability	An AI system which has outputs that are difficult for a human reviewer to interpret.
Local Interpretable Model-agnostic Explanation (LIME)	An approach to low interpretability which provides an explanation of a specific output rather the model in general.
Machine learning (ML)	The set of techniques and tools that allow computers to 'think' by creating mathematical algorithms based on accumulated data.
Membership inference attack	An attack which allows actors to deduce whether a given individual was present in the training data of a machine learning model.
Model inversion attack	An attack where attackers already have access to some personal data belonging to specific individuals in the training data, but can also infer further personal information about those same individuals by observing the inputs and outputs of the machine learning model.
Multi-criteria optimisation	A mathematical approach trying to satisfy multiple criteria in the process of decision-making.
Objective function	The goal that a machine learning algorithm is trying to achieve (eg 'minimise errors').
Perturbation	Where the values of data points belonging to individuals are changed at random whilst preserving some of the statistical properties of those features in the overall dataset.

Post-processing bias mitigation	A series of techniques applied to the initial model after its original training.
Precision	The percentage of cases identified as positive that are in fact positive (also called 'positive predictive value').
Pre-processing	The process of transforming data prior to using it for training a statistical model.
Privacy enhancing technologies (PETs)	A broad range of technologies that are designed for supporting privacy and data protection.
Programming language	A formal language comprising a set of instructions that produce various kinds of outputs that are using in computer programming to implement algorithms.
Query	A request for data or information from a database table or combination of tables.
Recall (or sensitivity)	The percentage of all cases that are in fact positive that are identified as such.
Redundant encodings	Patterns encoded in complex combinations of features.
Regularisation	A method to reduce overfitting to training data, particularly when the training data is scarce or known to be incomplete.
Reward function	The term is used within the context of a reinforcement learning model where you provide a reward if it is able to learn and penalise it by imposing a cost function.
Statistical accuracy	The proportion of answers that an AI system gets correct.
Supervised machine learning	A machine learning task of learning a function that maps an input to an output based on examples of correctly labelled input-output pairs.
Support Vector Machines (SVMs)	A method of separating out classes by using a line (or hyperplane) to divide a plane into parts where each class lay in either side.
Upstream effects	The impact(s) of an AI system on individuals at the early stage(s) of its development.
Use case	An AI application or the problem an AI system intends to solve.
Target variable	The outcome that an AI system seeks to predict.
Variance	The extent to which a model is overfitted to the data it is trained on. High variance means the model is more likely to fail when presented with new examples that are different from the training data. Ultimately, variance is used to understand how reliable a model is in its performance.
'Virtual machines' or 'containers'	Emulations of a computer system that run inside, but isolated from the rest of an IT system.
'White box' attack	Where an attacker has complete access to the model itself, and can inspect its underlying code and properties. White box attacks allow additional information to be gathered (such as the type of model and parameters used) which could help an attacker infer personal data from

the model.
