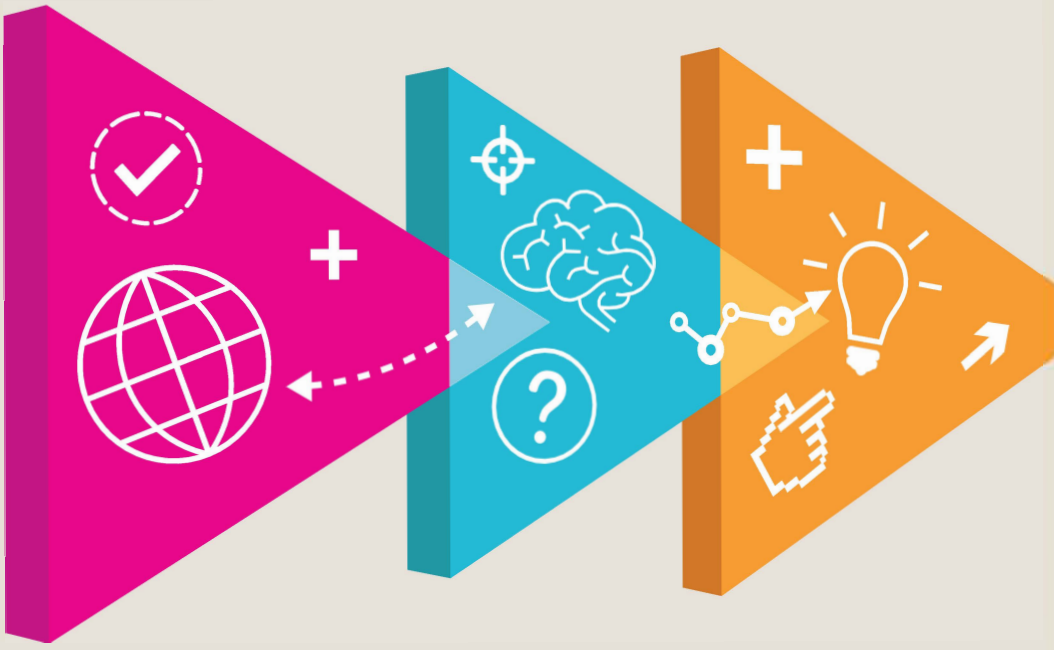
Guidance on the AI auditing framework

Draft guidance for consultation



Information Commissioner's Office

Contents

About this guidance

At a glance

Applications of artificial intelligence (AI) increasingly permeate many aspects of our lives. We understand the distinct benefits that AI can bring, but also the risks it can pose to the rights and freedoms of individuals.

This is why we have developed a framework for auditing AI, focusing on best practices for data protection compliance – whether you design your own AI system, or implement one from a third party. It provides a solid methodology to audit AI applications and ensure they process personal data fairly. It comprises:

- auditing tools and procedures that we will use in audits and investigations; and
- this detailed guidance on AI and data protection, which includes indicative risk and control measures that you can deploy when you use AI to process personal data.

This guidance is aimed at two audiences:

- those with a compliance focus, such as data protection officers (DPOs), general counsel, risk managers and the ICO's own auditors; and
- technology specialists, including machine learning experts, data scientists, software developers and engineers, and cybersecurity and IT risk managers.

The guidance clarifies how you can assess the risks to rights and freedoms that AI can pose; and the appropriate measures you can implement to mitigate them.

While data protection and 'AI ethics' overlap, this guidance does not provide generic ethical or design principles for your use of AI. It corresponds to different data protection principles, and is structured as follows:

- part one addresses accountability and governance in AI, including data protection impact assessments (DPIAs);
- part two covers fair, lawful and transparent processing, including lawful bases, assessing and improving AI system performance, and mitigating potential discrimination;
- part three addresses data minimisation and security; and
- part four is about how you can facilitate the exercise of individual rights in your AI systems, including rights related to automated decision-making.

In detail

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. That's why AI is one of our top three strategic priorities, and why we decided to develop a framework for auditing AI compliance with data protection obligations.

The framework:

- gives us a solid 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 framework to guide our own activity, we also wanted to share our thinking behind it. The framework therefore has two distinct outputs:

1. Auditing tools and procedures which will be used by our investigation and assurance teams when assessing the compliance of organisations using AI.
2. This detailed guidance on AI and data protection for organisations, which outlines our thinking and also includes indicative risk and control tables at the end of each section to help organisations audit the compliance of their own AI systems.

This guidance aims to inform you about what we think constitutes best practice for data protection-compliant AI.

This guidance is not a statutory code. It contains advice on how to interpret relevant 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.

Further reading – ICO guidance

Technology Strategy 2018-2021

What do you mean by 'AI'?

The term 'AI' has a variety of meanings. Within the AI research community, it refers to various methods 'for using a non-human system to learn from experience and imitate human intelligent behaviour', whilst in the context of data protection it has been referred to as 'the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages'.

However, the use of large amounts of data to make predictions or classifications about individuals has existed in sectors like insurance since the 19th century, long before the term 'AI' was coined in the 1950s. In these traditional forms of statistical analysis, statistical models are calculated using pen and paper (and later, with a calculator).

With modern machine learning techniques, it is now much easier to create statistical models using computationally intensive techniques on much larger, multi-dimensional datasets. This increase in the complexity of such models, combined with the decreasing costs of creating them, has heightened concerns about their risks to the rights and freedoms of individuals.

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 the context of 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 differ.

If you are processing large amounts of personal data for any purpose, data protection law applies. If you are processing this data in the context of statistical models, and using those models to make predictions about people, this guidance will be relevant to you regardless of whether you classify those activities as ML (or AI).

We use the umbrella term 'AI' because it has become a mainstream way for organisations to refer to a range of technologies that mimic human thought. Similar technologies, which have similar sources of risk, are likely to benefit from the same set of risk measures. So, whether you call it AI, machine learning, complex information processing, or something else, the risks and controls identified here should be helpful. Where there are important differences between different types of AI, for example, simple regression models and deep neural networks, we will refer to these explicitly.

Other resources

See the International Working Group on Data Protection in Telecommunications' 2018 working paper on privacy and artificial intelligence (PDF, external link)

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 explAIn guidance)

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 in the last three years, 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 Turing Institute we have produced guidance on how organisations can best explain their use of AI to individuals. This resulted in the ExplAIn guidance, which was published in draft form for consultation last year. We are now in the process of finalising the explAIn guidance in light of feedback from stakeholders and will update links in this guidance when it is completed.

While the explAIn 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 both in tandem.

Further reading – ICO guidance

Big data, artificial intelligence, machine learning and data protection

ICO and The Turing consultation on Explaining AI decisions guidance

Who is this guidance for?

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

First, those with a compliance focus, including:

- data protection officers;
- general counsel;
- risk managers; and
- the ICO's own auditors – in other words, we will utilise this guidance ourselves in the exercise of 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.

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.

Why is the ICO focusing on 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 and/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 designers 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. These may not have direct application to developers who don't go on to play a role in the processing – however, you should also note that:

- you have a responsibility to put in place appropriate technical and organisational measures designed to implement the data protection principles in an effective manner; and
- if you are a developer and you use personal data to train your models, then you will be a controller for that processing and data protection law applies.

Certain design choices are more likely to result in AI systems which infringe data protection in one way or other. This guidance will help designers 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. As such, more general data protection considerations, are not addressed except in so far as they relate to and are challenged by AI.

Other resources

Read the Global Privacy Assembly's 'Declaration on ethics and data protection in artificial intelligence' (PDF, external link) for more information on how ethics and data protection intersect in the AI context.

What legislation applies?

This guidance deals with the challenges that AI raises for data protection. As such, the most relevant piece of UK legislation is the Data Protection Act 2018.

The DPA 2018 sets out the UK's data protection framework, alongside the General Data Protection Regulation (GDPR). It comprises the following data protection regimes:

- Part 2 – supplements and tailors the GDPR in the UK;
- 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 different regimes, these are explicitly addressed in the text.

You should also review our guidance on how Brexit 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 – ICO guidance

For more information on the different regimes, read the 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 corresponding to different data protection principles and rights.

Part one addresses issues that primarily relate to the accountability principle. This requires you to be responsible for complying with the data protection principles and for demonstrating that compliance. Sections in this part deal with the AI-specific implications of accountability including data protection impact assessments (DPIAs), controller / processor responsibilities, and assessing and justifying trade-offs.

Part two covers the lawfulness, fairness, and transparency of processing personal data in AI systems, with sections covering lawful bases for processing personal data in AI systems, assessing and improving AI system performance and mitigating potential discrimination to ensure fair processing.

Part three covers the principle of security and data minimisation in AI systems.

Part four covers how you can facilitate the exercise of individuals' rights about their personal data in your AI systems, and rights relating to solely automated decisions. In particular, part four covers how you can ensure meaningful human input in non-automated or partly-automated decisions, and meaningful human review of solely automated decisions.

What are the accountability and governance implications of AI?

At a glance

The accountability principle makes you responsible for complying with data protection and for demonstrating that compliance in any AI system. In an AI context, accountability requires you to:

- be responsible for the compliance of your system;
- assess and mitigate its risks; and
- document and demonstrate how your system is compliant and justify the choices you have made.

You should consider these issues as part of your DPIA for any system you intend to use. You should note that you are legally required to complete a DPIA if you use AI systems that process personal data. DPIAs offer you an opportunity to consider how and why you are using AI systems to process personal data and what the potential risks could be.

Due to the complexity and mutual dependency of the various kinds of processing typically involved in AI supply chains, you also need to take care to understand and identify controller / processor relationships.

Additionally, depending on how they are designed and deployed, AI systems will inevitably involve making trade-offs between privacy and other competing rights and interests. You need to know what these trade-offs may be and how you can manage them, as otherwise there is a risk that you fail to adequately assess them and strike the right balance. However, you should also note that you always have to comply with the fundamental data protection principles, and cannot 'trade' this requirement away.

In detail

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 Data Protection Officers (DPOs), are accountable for understanding and addressing them appropriately and promptly.

To do so, in addition to their own upskilling, they need diverse, well-resourced, teams to support them in discharging 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. Your governance and risk management capabilities need to be proportionate to your use of AI. This is particularly true now as 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 still developing quickly.

We are also currently developing a more general accountability toolkit. This is not specific to AI, but provides a baseline for demonstrating your accountability under the GDPR, on which you could build your approach to AI accountability. We will update the final version of this guidance to refer to the final version of the accountability toolkit when it is published.

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.

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

This applies to 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. See 'What are trade-offs and how should we manage them?' below.

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

- assess the risks to individuals' rights that your use of AI poses;
- determine how you need to 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. Ultimately 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 a mere box ticking compliance exercise. They can effectively act as roadmaps for you to identify and control the risks to rights and freedoms that use of AI can pose. They are also a perfect opportunity for you to consider and demonstrate your accountability for the decisions you make in the design or procurement of AI systems.

Why do we need to carry out DPIAs under the data protection law?

Using AI to process personal data is likely to result in high risk to individuals' rights and freedoms, and therefore triggers the legal requirement to undertake a DPIA.

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 instance, public sector organisations are required to undertake equality impact assessments, while other organisations voluntarily undertake 'algorithm impact assessments'. 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.

Relevant provisions in the legislation

See Articles 35 and 36 and Recitals 74-77, 84, 89-92, 94 and 95 of the GDPR (external link)

See Sections 64 and 65 of the DPA 2018 (external link)

How do we decide whether to do a DPIA?

As AI is on the list of types of processing that are likely to result in a high risk, if you use it to process personal data, then you must carry out a DPIA. In any case, if you have a major project that involves the use of personal data it is 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 – ICO guidance

See 'How do we decide whether to do a DPIA?' in our 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.

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 '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
    - how far individuals are likely to expect the processing.

A 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 a complex AI system. It may be appropriate 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.

A 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 GDPR; and if so, collaborate in the DPIA process as appropriate.

Where 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.

Relevant provisions in the legislation

See Article 35(7)(a) and Recitals 84, 90 and 94 of the GDPR (External link)

Do we need to consult anyone?

You should:

- seek and document the views of individuals 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, where appropriate.

Unless there is a good reason not to do so, you should seek and document the views of individuals, or their representatives, on the intended processing operation during a DPIA. It is therefore important that you can describe the processing in a way those who are consulted can understand.

Relevant provisions in the legislation

See Article 28(3)(f) and Article 35 (9) of the 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 the proven ability of that system to fulfil a specific and legitimate purpose,

not by the 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 its 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?

The DPIA process will help you to objectively identify the relevant risks. 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, a DPIA should consider the potential impact of material and non-material damage or harm on individuals.

For instance, 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 such contexts, you should consider the relevant legal frameworks beyond data protection.

Relevant provisions in the legislation

See Articles 35(7)(c) and Recitals 76 and 90 of the GDPR (External link)

How do we identify mitigating measures?

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

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

It is important that DPOs and other information governance professionals 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 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 appreciation for 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 measures have been introduced to mitigate the risks 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 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 instance, 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.

Relevant provisions in the legislation

See Articles 35(11), 36(1) and 39(1)(c) and Recital 84 of the GDPR (external link)

Further reading – ICO guidance

Read our guidance on DPIAs in the Guide to the 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 GDPR.

WP29 produced guidelines on data protection impact assessments (WP248 rev.01), which have been endorsed by the EDPB.

Other relevant guidelines include:

- Guidelines on Data Protection Officers ('DPOs') (WP243 rev.01)
- Guidelines on automated individual decision-making and profiling (WP251 rev.01)

How should we understand controller/processor relationships in AI?

Why is controllership important for AI systems?

In many cases, the various processing operations involved in AI may be undertaken by a number of different organisations. It is therefore crucial that you determine who is a controller, joint controller or processor if your use of AI involves multiple organisations.

A first step to understanding these relationships is to identify the distinct sets of processing operations and their purposes (see 'What should we assess in our DPIA?). For each of these, you and the organisations you work with then need to assess whether you are a controller, processor or joint controller. You should consult our guidance on controller/processor for assistance in this assessment, but in essence:

- if you decide on the purposes and means of processing, you are a controller;
- if you process personal data under the instruction of another organisation, you are a processor; and
- if you jointly determine the purposes and means of processing with another organisation, you are joint controllers.

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

While the means of processing are decided on by controllers, processors may have some discretion to decide on some of the *non-essential* details of those means. While it is not possible to generalise about every context, the following examples show the kinds of decisions about means of AI-related processing that may indicate you are a controller, and those decisions about non-essential details which could be taken by a processor.

What type of decisions may make us a controller?

The type of decisions that are likely to make you a controller include deciding:

- to collect personal data in the first place
- the purpose(s) for that processing
- which individuals to collect the data about, and what to tell them about the processing;
- how long you will retain the data for; and
- how to respond to requests made in line with individuals' rights.

For more specifics on the circumstances that may make you a controller, read our guidance on controllers and processors. In the context of AI, the type of decisions made by controllers can include:

- the source and nature of the data used to train an AI model;
- the target output of the model (ie 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.

Whilst non-exhaustive, the above list constitutes decisions about purpose and means. If you make any of these decisions, you are likely to be a controller; if you and another organisation jointly determine them then you are joint controllers.

What decisions can processors take?

Conversely, if you don't make any of the above decisions, and you only process data on the basis of agreed terms, you are more likely to be a processor. Processors have different obligations, and depending on the terms of any contract you have with your controller, you are able to take certain decisions such as:

- 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.

In an AI context, processors may be able to decide (depending on the terms of their contract):

- 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.

If you only make these kinds of decisions, you may be a processor rather than a controller.

In practice, this means that you can provide a variety of AI services without necessarily being considered a controller. However, the key is to remember that the overarching decisions about what data is processed for what purposes can only be taken by a controller.

Providing AI development tools

For instance, you might provide 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 your clients to build and run their own models, with data they have chosen to process, but using the tools and infrastructure you provide them in the cloud.

In this example, the clients are more likely to be controllers, whilst you are a processor.

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

Your clients are controllers as long 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.

If you use your clients' data for any other purposes than those decided on by them, you become a controller for such processing.

Providing AI prediction as a service

Some companies provide live AI prediction and classification services to customers. For instance, you might develop your own AI models and allow customers to send queries to them via an API (eg 'what objects are in this image?') and get responses (eg a classification of the objects in the image).

In this case, as an AI prediction service provider, you are likely to be a controller for at least some of the processing.

First, there is the processing necessary to create and improve the models that power your services; if the means and purposes of this processing are principally decided on by you as the service provider, you are likely to be the controller for such processing. Second, there is the processing necessary to make the predictions and classifications about particular examples on behalf

of your clients; for this kind of processing, the client is more likely to be the controller with you as the processor.

However, you may even be considered a controller, or joint controller, for the latter kind of processing if you design your services in such a way that your customers do not have sufficient influence over the essential elements and purposes of the processing involved in the prediction. For instance, if you do not enable your customers to set the balance between false positives and negatives according to their own requirements, or to add or remove certain features from a model, you may in practice be a joint controller. This is because you are exerting a sufficiently strong influence over the essential means of processing.

Furthermore, if you initially process data on behalf of a client as part of providing them a service, but then process that same data from your clients to improve your own models, then, you are a controller for this processing. This is because you have decided to undertake it for your own purposes. For instance, a recruitment consultant might process job applicants' data on behalf of their clients using an ML system which provides scores for each applicant, while also retaining that data to further improve their ML system.

In such cases, you should explicitly state these purposes from the outset and seek your own lawful basis. See the section on purposes and lawful bases for further information.

Additionally, data protection legislation states that anyone acting under the authority of the controller or the processor shall not process personal data except on instructions from the controller, unless required by law. You need to consider how this applies in your particular circumstances. For example, a processor is only permitted to process on instructions of the controller – so if it further processes personal data that it only due to its role as a processor, that processing may still infringe the GDPR even if the purposes for fair and compatible. The original controller needs to agree to disclose that data to you as a third-party controller, ie in a data sharing operation (which itself needs to comply with data protection law).

Relevant provisions in the legislation

See Article 29 of the GDPR (external link)

See Sections 60 and 106 of the DPA 2018 (external link).

What are our responsibilities when procuring AI models for local deployment?

Some companies sell (or provide for free) pre-trained AI models as standalone pieces of software which their customers can install and run.

In order to develop and market a robust AI system that organisations such as yours would purchase, these third parties are likely to have processed personal data in some form (eg to train the system). However, once you

integrate the system into your processing environment the developer may play no further part in the processing itself.

So, where these providers make decisions about the processing of personal data for the purposes of **training** the models they sell, they are a controller for such processing.

However, third party developers of AI systems may not intend to process personal data in their own right (as controllers) or on your behalf (as processors) once you have procured their models. Once you deploy the model, if provider does not process personal data, then the you are the controller for any processing you undertake using that model.

If you are a provider of such AI services, you should identify which processing operations you are the controller or processor for and ensure this is clear with your clients. If you are procuring these services, you have different responsibilities. As part of your considerations when selecting an AI system to process personal data, you should:

- remember that compliance with data protection law remains your responsibility, whether you procure a system or build your own;
- check how the developer has designed the AI system, including whether they did so in line with the data protection principles; and
- as a result, assess whether using this service is going to help you meet both your processing objectives, as well as your data protection obligations.

As a controller, when writing contracts and service level agreements, you also need to remember that while your contract might stipulate your status as controller or processor, what matters from a data protection perspective is who **in practice** decides the purposes and essential means of processing.

Similarly, if you are a provider of such AI services, you should identify which processing operations you are the controller or processor for and ensure this is clear with your clients.

Relevant provisions in the legislation

See Articles 4(7), 4(8), 5(1), 5(2), 25, 28 and Recital 78 of the GDPR (external link)

Further reading – ICO guidance

Read our guidance on controller/processor and contracts/liabilities in the Guide to the GDPR.

Other resources

See the Court of Justice of the European Union's (CJEU) judgment in the case of Unabhängiges Landeszentrum für Datenschutz (ULD) Schleswig-Holstein against Wirtschaftsakademie Schleswig-Holstein GmbH.

See also the CJEU's judgment in the case of Fashion ID GmbH & Co. KG against Verbraucherzentrale NRW eV.

What are AI-related trade-offs and how should we manage them?

Your use of AI must comply with the requirements of data protection law. However, there can be a number of different values and interests, which may at times pull in different directions. The risk-based approach of data protection law can help you navigate potential 'trade-offs' between privacy on the one hand and other competing values and interest on the other.

If you are using AI, you therefore need to identify and assess these interests, and strike an appropriate balance between them given your context whilst continuing to meet your obligations under the law.

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

The following sections provide a short overview of some of the most notable trade-offs that you are likely to face when designing or procuring AI systems.

Privacy vs statistical accuracy

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.

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. Accuracy in AI (and, more generally, in statistical modelling) refers to how often an AI system guesses the correct answer – and in many cases, these answers will be personal data.

While the '**accuracy principle**' applies to the personal data you process, an AI system does not need to be 100% '**statistically accurate**' in order to comply with that principle. However, the more statistically accurate an AI

system is, the more likely that your processing will be in line with the fairness principle.

So, for clarity, in this guidance:

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

For more information on the differences between 'data protection' accuracy and statistical accuracy, see the section 'What do we need to do about 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.

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 instance, 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.

Further reading – ICO guidance

Read our guidance on data minimisation in the Guide to the GDPR.

Statistical accuracy and discrimination

As discussed in the 'How should we address risks of bias and discrimination?' section, the use of AI systems can lead to biased or discriminatory outcomes. These may in turn pose compliance risks in terms of the fairness principle. You need to implement appropriate technical measures to mitigate this risk.

Your considerations should also include 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.

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 minority 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 posed to the other rights and freedoms of those individuals.

Explainability and statistical accuracy

Part of your fairness considerations also include a trade-off between the explainability and statistical accuracy of AI systems.

For very complex systems, such as those based on deep learning, it may be hard to follow the logic of the system and therefore difficult for you to adequately explain how they work. This is sometimes characterised as the 'black box problem'.

Depending on the circumstances, you may view these complex systems as the most statistically accurate and effective, especially when it comes to problems like image recognition. You may therefore think that you face a trade-off between explainability and statistical accuracy.

However, for many applications, where simpler models perform well, the trade-offs between explainability and statistical accuracy may actually be relatively small. These issues are considered in greater depth in the ExplAIn project guidance, but as a general point, you should only use 'black box' models if:

- you have thoroughly considered their potential impacts and risks in advance, and the members of your team have determined that your use case and your organisational capacities/resources support the responsible design and implementation of these systems; and
- your system includes supplemental interpretability tools that provide a domain-appropriate level of explainability.

Explainability, exposure of personal data, and commercial security

Providing individuals with meaningful information about the logic of an AI-driven decision can potentially increase the risk of inadvertently disclosing information you need to keep private – including personal data but also other information such as proprietary logic of the system –in the process.

Recent research has demonstrated how some proposed methods to make ML models explainable can unintentionally make it easier to infer personal data about the individuals whose data you used to train the model. (See the

sections on 'What are model inversion attacks?' and 'What are membership inference attacks?').

Some research also highlights the risk that in the course of providing an explanation to individuals, you may accidentally reveal proprietary information about how an 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.

Our research and stakeholder engagement so far indicate this risk is quite low. However, in theory at least, there may be cases where you will 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.

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 proportional 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 reasonable 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 (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, and should capture them in your DPIA with an appropriate level of detail.

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 an appropriate trade-off between two or multiple 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 ex post.

Recital 78 of the 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 you have the authority to modify the systems before and during deployment, either directly or via the third party provider, so that they align with what you consider to be the appropriate trade-offs at the outset and as new risks and considerations arise.

For instance, 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 in order to make 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. See the section on data minimisation.

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.

Assessing trade-offs: a worked example

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.

One possible approach to help you identify the best possible trade-offs is to visually represent the choices between different designs on a graph.

You can plot possible choices about how a system could be designed on a graph, with two criteria being balanced – for example statistical accuracy and privacy – on the X and Y axis. Statistical accuracy can be given a precise measurement in the ways described in the section 'What do we need to do about statistical accuracy?'.

Privacy measures are likely to be less exact and more indicative in nature, but could include:

- the amount of personal data required;
- the sensitivity of this data;
- the extent to which it might uniquely identify the individual;
- the nature, scope, context and purposes of the processing;
- the risks to rights and freedoms the data processing may present to individuals; and
- the number of individuals the AI systems will be applied to.

A graph can reveal what is known as the 'production-possibility frontier' and is one way to help decision makers understand how system design decisions may impact the balance between different values.

We have used this method in Figure 1 to visualise what a trade-off between privacy and statistical accuracy might look like. This can be presented to a senior decision maker responsible for approving the choice of a particular system to help them understand the trade-offs.

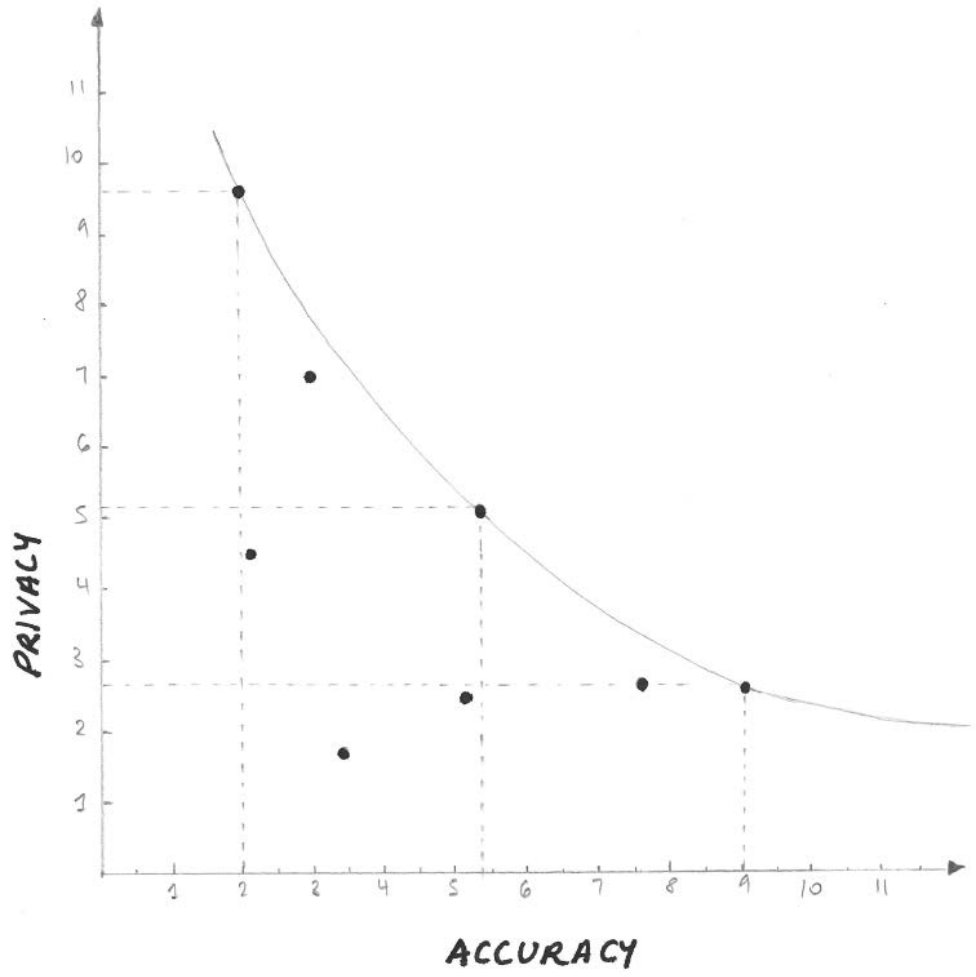


Figure 1

The data points in the graph represent the different possible technical configurations of an AI system, resulting from different design choices, ML models, and the amount and types of data used. Each point represents a possible end result of a particular trade-off between statistical accuracy and privacy.

In the scenario in Figure 1, none of the proposed systems can achieve both high statistical accuracy *and* high privacy, and the trade-off between privacy and statistical accuracy is significant.

For a different use case, the trade-offs may look very different, as visualised in Figure 2.

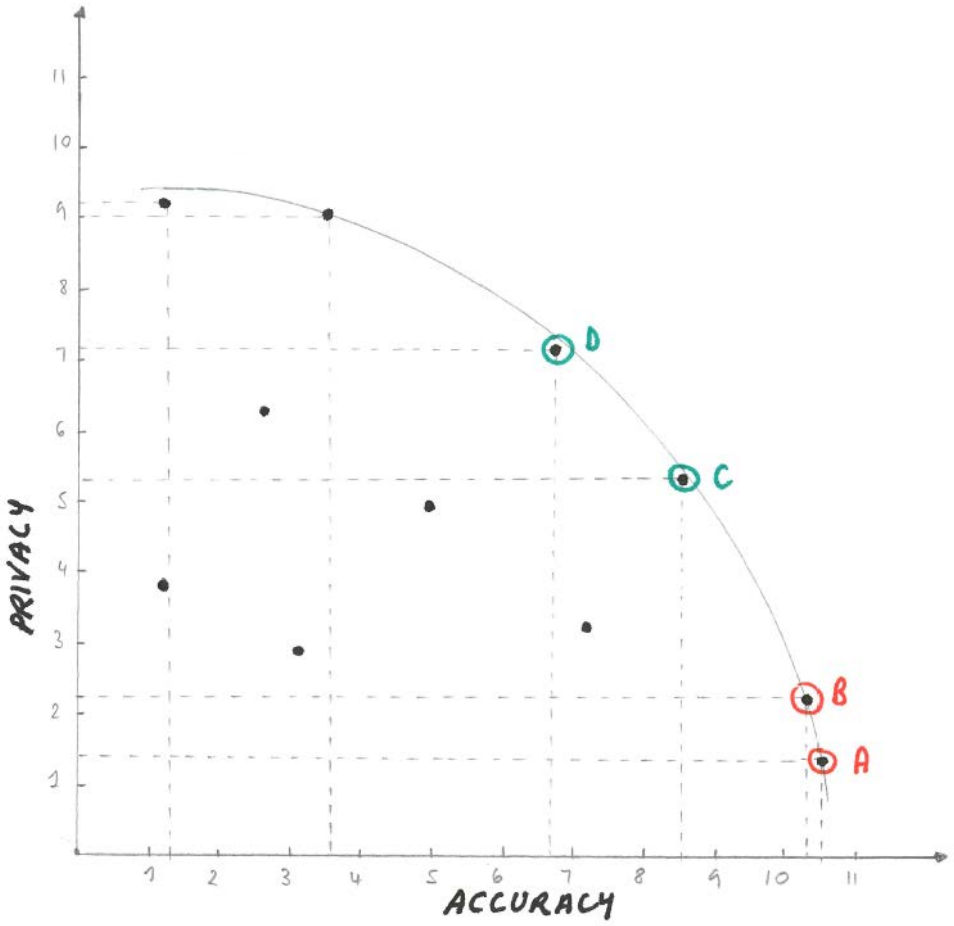


Figure 2

In this scenario, it may be easier to achieve a reasonable trade-off between statistical accuracy and privacy. The graph also shows that the cost of sacrificing either privacy or statistical accuracy is lower for those AI systems in the middle of the curve (C and D), than for those at the edge (A and B).

In this example, there are diminishing returns on statistical accuracy for the possible systems at the bottom right. If you intend to choose system A over B, you would have to justify why placing a higher value on statistical accuracy is warranted in the circumstances. You need to take into account the impact on the rights and freedoms of individuals, and be able to demonstrate that the processing is still fair and proportionate overall.

Visual representation of trade-offs can also include lower limits for either variable below which you are not willing to go (Figure 3).

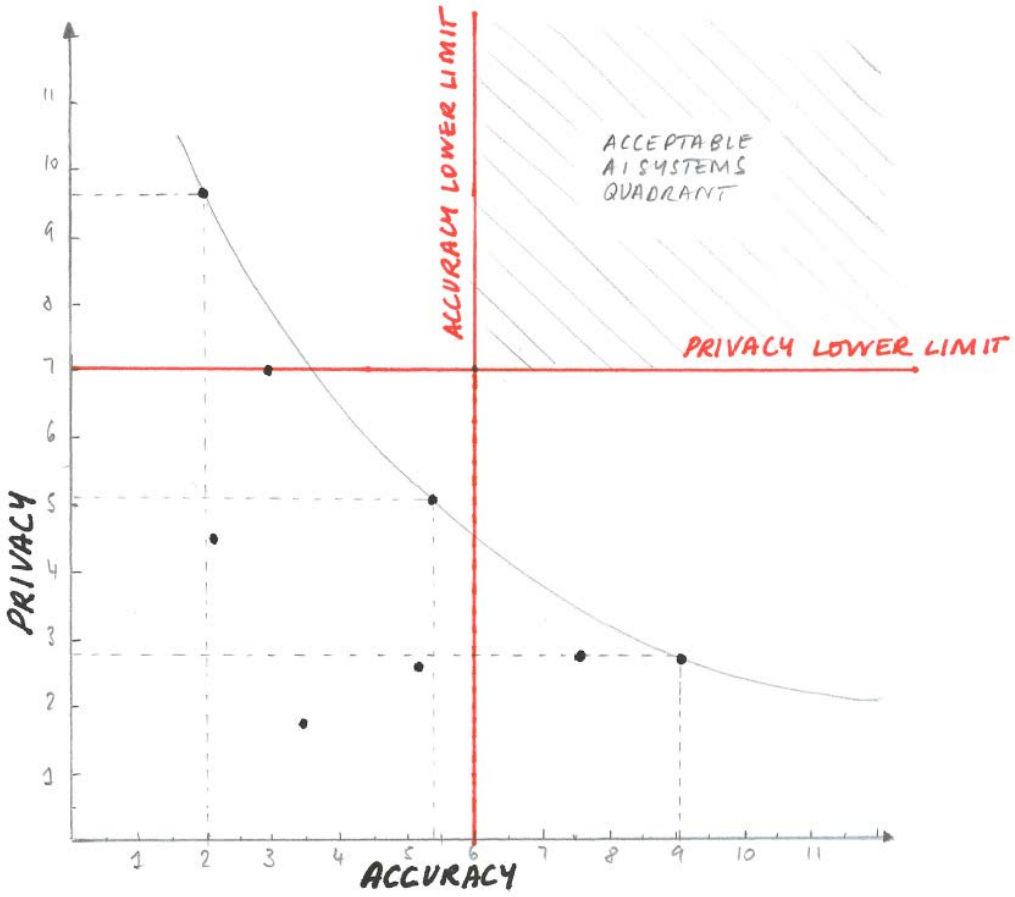


Figure 3

In the scenario in Figure 3, there is no possible system which meets both the lower limits for statistical accuracy and for privacy, suggesting that you should not pursue the deployment of any of these systems. This may mean looking into other methods and data sources, reformulating the problem, or abandoning the attempt to use AI to solve the problem.

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 such trade-offs. A technical specialist, such as an ML engineer, should assess the viability of these techniques in your particular context.

For instance, 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 measure the likelihood of an individual being uniquely identified from a particular dataset, but not the sensitivity of that identification. Therefore, 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.

Example of controls

Risk statement

Inadequate or inappropriate trade-off analysis / decisions lead to AI systems that incorrectly prioritise one criterion over another more important criteria.

Preventative

- Clearly document the purpose of the model and the most important criteria in your model specification.
- Ensure this specification is signed off by appropriate management.
- Senior management review various models (trade-off analysis) and approve a particular model for use.
- Systematically review the trade-off options and provide justification as to why the specific model was selected.
- Ensure reviews have been completed and action taken as a result.
- Complete training to ensure AI system designers are up to date with latest techniques.

Detective

- Do a periodic review of trade-off given new data available since date of deployment.
- Periodically re-analyse trade-offs.

Corrective

- Select a more appropriate model, and include a thorough justification for the change.
- Retrain the AI system developers

What do we need to do to ensure lawfulness, fairness, and transparency in AI systems?

At a glance

When you use AI to process personal data, you must ensure that it is lawful, fair and transparent. Compliance with these principles may be challenging in an AI context.

As AI systems process personal data in various stages for a variety of purposes, there is a risk that if you fail to appropriately distinguish each distinct processing operation and identify an appropriate lawful basis for it, this could lead to a failure to comply with the data protection principle of lawfulness.

This section presents considerations that will help you find an appropriate lawful basis for the various kinds personal data processing involved when creating or using AI and ensure such processing is fair.

In detail

How does the principle of lawfulness, fairness and transparency apply to AI?

First, the development and deployment of AI systems involve processing personal data in different ways for different purposes. You must identify these purposes and have an appropriate lawful basis in order to comply with the principle of lawfulness.

Second, if you use an AI system to infer data about people, in order for this processing to be fair, you need to ensure that:

- the system is sufficiently statistically accurate and avoids discrimination; and
- you consider the impact of individuals' reasonable expectations.

Finally, you need to be transparent about how you process personal data in an AI system, to comply with the principle of transparency. The core issues regarding AI and the transparency principle are addressed in the explAIn guidance, so are not discussed in detail here.

How do we identify our purposes and lawful basis when using AI?

What should we consider when deciding lawful bases?

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.

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 – ICO guidance

Read our guidance on lawful basis for processing in the Guide to GDPR

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 **development** or **training** of AI systems from their **deployment**. This is because these are distinct and separate purposes, with different circumstances and risks.

Therefore, you should consider whether different lawful bases apply for your AI **development** and **deployment**. For example, you need to do this where:

- the AI system was **trained** for a general-purpose task, and you then **deploy** it in different contexts for different purposes. For instance, 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;
- in cases where 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 to that for which you intend to use the system, 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 **training** and **deployment** stages.

The following sections outline some AI-related considerations for each of the GDPR's lawful bases. They do not consider Part 3 of the DPA 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 for training and deploying your model.

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. 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 instance, 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.

Relevant provisions in the GDPR

See Articles 4(11), 6(1)(a) 7, 8, 9(2)(a) and Recitals 32, 38, 40, 42, 43, 171 (external link)

Further reading – ICO guidance

Read our guidance on consent in the Guide to the GDPR for more information.

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 GDPR.

WP29 adopted Guidelines on consent, which the EDPB endorsed in May 2018.

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 **train** 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.

Similarly, even if you can use performance of a contract as a lawful basis for processing personal data with AI to provide a quote prior to a contract, this does not mean you can also use it to justify using that data to train 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).

Relevant provisions in the legislation

See Article 6(1)(b) and Recital 44 of the GDPR (external link)

Further reading – ICO guidance

Read our guidance on contracts in the Guide to the GDPR for more information.

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 GDPR.

The EDPB published Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) in the context of online services in November 2019.

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** (such as for the purposes of detecting anti-money laundering under Part 7 of the Proceeds of Crime Act 2002).

Similarly, where an organisation uses AI as part of the exercise of its 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.

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 training an AI system, because this would rarely directly and immediately result in protecting the vital interests of individuals, even if the models that are eventually built might later be used to save lives. For the training 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 GDPR for provisions on using legal obligation (external link)

See Article 6 (1)(e) and 6(3), and Recitals 41, 45 and 50 of the GDPR for provisions on using Public Interests.

See Article 6(1)(d), Article 9(2)(c) and Recital 46 of the GDPR for provisions on using 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 – ICO guidance

Read our guidance on legal obligation, vital interests and public task in the Guide to the GDPR for more information.

Can we rely on legitimate interests?

Depending on your circumstances, you could base your processing of personal data for both training 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.

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, and 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).

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.

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.

Relevant provisions in the legislation

See GDPR Article 6(1)(f) and Recitals 47-49 (external link)

Further reading – ICO guidance

Read our guidance on legitimate interests in the Guide to the GDPR.

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 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.

A number of these conditions will also require to you meet additional requirements and safeguards set out in Schedule 1 of the DPA 2018.

You must:

- determine, and document, your condition for processing before you start it;
- ensure you have an appropriate policy document in place, where required; and
- complete a DPIA for any processing likely to be high risk.

For data about criminal offences, you need a lawful basis under Article 6 of the GDPR, and either lawful or official authority under Article 10.

The DPA 2018 sets out specific conditions that provide lawful authority. You can also process this type of data if you have official authority to do so (ie you are processing the data in an official capacity).

As with special category data, you must determine your condition for processing (or identify your official authority) before you start it, and you should also document this.

Relevant provisions in the legislation

See Articles 9 and 10 of the GDPR (external link)

Further reading – ICO guidance

Read our guidance on special category data and on criminal offence data in the Guide to the GDPR.

What is the impact of Article 22 of the GDPR?

Data protection law applies to all automated individual decision making and profiling. Article 22 of the 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 have application in the AI context, eg where you are using an AI system to make these kinds of decisions.

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.

Further reading – ICO guidance

Read our guidance on rights related to automated decision making including profiling in the Guide to the GDPR.

Example controls

Risk Statement

Reliance on an inappropriate lawful basis for processing results in potential failure to fulfil the necessary requirements and non-compliance with DP legislation.

Preventative

- Ensure AI system developers have completed training and associated competency assessments.
- Document training for key stakeholders and how the relevant personnel were identified (eg senior management, risk managers, audit).
- Thoroughly assess and justify your lawful basis for processing in your DPIA.
- Consult with DP specialists within your model design workforce.
- Ensure the requirement for a DPIA is documented and AI developers are provided with clear guidance on the assessment criteria.
- Complete a legitimate interests assessment where there is a reliance on legitimate interests as a lawful basis.

Detective

- Monitor individual rights requests and complaints from individuals, including the action taken as a result (at both individual level and boarder analysis).
- Conduct a periodic DPIA review to ensure it remains accurate and up to date.
- Periodically assess the model usage to ensure purpose remains the same and necessity and legitimate interests (LI) still valid.
- Conduct a periodic review of records of processing to ensure validity of lawful basis.

Corrective

- Implement corrective measures to AI system in order to satisfy the original lawful basis
- Select a new lawful basis and associated actions. For example, carrying out a legitimate interests assessment or obtaining consent.
- Retrain AI system developers / individuals involved in the assessment of lawful bases.

What do we need to do about statistical accuracy?

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

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

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

As said in the section 'What are trade-offs and how should we manage them?', accuracy has slightly different meanings in data protection and AI contexts.

In data protection, accuracy is one of the fundamental principles. 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.

In AI, accuracy refers to how often an AI system guesses the correct answer. In many contexts, the answers the AI system provides will be personal data. For instance, an AI system might infer someone's demographic information or their interests from their behaviour on a social network.

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** in order 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. In order 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 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 – including 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 (see next section).

Relevant provisions in the legislation

See GDPR Articles 5(1)(d), 22 and Recital 71 (external link)

Further reading – ICO guidance

Read our guidance on accuracy in our Guide to the GDPR as well as our guidance on the rights to rectification and erasure.

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 GDPR.

WP29 published Guidelines on automated decision making and profiling in 2017. The EDPB endorsed these guidelines in May 2018.

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 instance, 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 to assess how good a system is. 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 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 instance, 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 instance, 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 measures such as the 'F-1' statistic - see link below). 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.

Example

If a CV filtering system selecting qualified candidates for an interview produces a false positive, then an unqualified candidate will be invited to interview, wasting the employer 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 instance, 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, 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 instance, 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, 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 – ICO guidance

See our guidance on the accuracy principle in the Guide to the GDPR.

Other resources

See 'How should we define and prioritise different statistical accuracy measures?'

See 'Learning under concept drift: an overview' for a further explanation of concept drift.

What should we do?

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 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 else should we do?

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.

Example of controls

Risk Statement

Inaccurate output or decisions made by AI systems could lead to unfair / negative outcomes for individuals and a failure to meet the fairness principle.

Preventative

- Put in place a data governance framework, which describes how all personal data used for ongoing training, testing or evaluation of an AI system or service is as correct, accurate, relevant, representative, complete and up-to-date as possible.

- Provide training for key stakeholders and document how the relevant personnel were identified (eg senior management, risk managers, audit).
- Document your access management controls and segregation of duties for the development and deployment of AI systems to ensure changes affecting statistical accuracy are made and signed off by authorised persons. Maintain evidence of how these controls are monitored.
- Document levels of approval authority for the development/use of AI systems. Maintain evidence of appropriate approval.
- In your DPIA, include thorough assessment of the impact/importance of different errors.
- Maintain documented policies / processes for dealing with third parties and evidence of the due diligence completed. In particular, when procuring AI systems / services, ensure they meet statistical accuracy requirements and allow regular re-testing.
- Maintain and document a policy / process for performing pre-implementation testing of any AI systems or changes prior to go-live. Maintain evidence that testing was completed prior to the deployment of the AI system and the results of the test(s).

Detective

- Do post-implementation testing, document the results of the testing and action(s) taken as a result.
- Monitor the output of reports/performance against expectations.
- Conduct a human review of a sample of AI decisions for statistical accuracy, including how the sample was selected / criteria used.
- Document individual rights requests and complaints regarding statistically inaccurate outputs from AI systems from individuals, in particular, any relating to Article 22, including the action taken as a result (at both individual level and broader analysis).
- Ensure there is continuous oversight of any third-party suppliers/processors including regularly reviewing performance against expectations and adherence to contractual requirements.
- Test the AI system on new data set(s) to confirm the same outcome is reached.

Corrective

- Retrain the AI system (eg by improving input data, different balance of false positives and negatives, or using different learning algorithm).
- Retrain the AI system developers in relation to discriminatory model performance.
- Change the decision made by the AI and assess whether other individuals could have been impacted by the inaccuracy.

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.

Data protection law is intended to balance the right to the protection of personal data to its function in society. Processing of personal data which leads to discrimination and bias will impact on the fairness of that processing. This poses compliance issues with the fairness principle as well as risks to individuals' rights and freedoms – including the right to non-discrimination. Furthermore, the GDPR specifically notes that organisations should take measures to prevent 'discriminatory effects on natural persons'.

Additionally, the UK's anti-discrimination legal framework, notably the UK Equality Act 2010, sits alongside data protection law and applies to a range of organisations. This includes government departments, service providers, employers, education providers, transport providers, associations and membership bodies, as well as providers of public functions. It gives individuals protection from discrimination, whether generated by a human or an automated decision-making system (or some combination of the two).

In this section we explore what this means, in practice, in the context of AI, focusing on how machine learning (ML) systems used to classify or make a prediction about individuals may lead to discrimination. We also explore some of the technical and organisational measures that you can adopt to manage this risk.

Why might an AI system lead to discrimination?

Let's take a hypothetical scenario:

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.

There are two main reasons why this might be.

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.

The AI algorithm will generate a statistical model 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 instance, 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.

These issues can occur even if the training data does not contain any protected characteristics like gender or race. A variety of features in the training data are often closely correlated with protected characteristics, eg occupation. These 'proxy variables' enable the model to reproduce patterns of discrimination associated with those characteristics, even if its designers did not intend this.

These problems can occur in any statistical model. 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 also include patterns that reflect discrimination.

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, there are various approaches to mitigate AI-driven discrimination. Computer scientists and others have been developing different mathematical techniques to measure how ML models treat individuals from different groups in potentially discriminatory ways. This field is often referred to as algorithmic 'fairness'. While many of these techniques are at the early stages of development and may not be market-ready, there are some basic approaches which AI developers can and should take to measure and mitigate potential discrimination resulting from their systems.

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).

Alternatively, you could train separate models, for example one for men and another for women, and design them to perform as well as possible on each sub-group. However, in some cases, creating different models for different protected classes could itself be a violation of non-discrimination law (eg different car insurance premiums for men and women).

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.

In order for these techniques to be effective, you need to choose one or more mathematical 'fairness' measures against which you can measure the results.

These measures can be grouped in three broad categories:

- anti-classification;
- outcome / error parity; and
- equal calibration.

Anti-classification is where a model is fair if it excludes protected characteristics when making a classification or prediction. Some anti-classification approaches also try to identify and exclude proxies for protected characteristics (eg attendance at a single-sex school). This can be impractical as removing all possible proxies may leave very few predictively useful features. Also, it is often hard to know whether a particular variable (or combination of variables) is a proxy for a protected characteristic without further data collection and analysis.

Outcome / error parity compares how members of different protected groups are treated by the model. With **outcome parity**, a model is fair if it gives equal numbers of positive or negative outcomes to different groups. With **error parity**, a model is fair if it gives equal numbers of **errors** to different groups. Error parity can be broken down into parity of false positives or false negatives (see section 2.3 on statistical accuracy for more details).

Equal calibration – calibration measures how closely the model's estimation of the likelihood of something happening matches the actual frequency of the event happening. According to 'equal calibration' a model is fair if it is equally calibrated between members of different protected groups. For instance, if a classification model sorts loan applicants into those with a low, medium, or high chance of repayment, there should be equal proportions of male and female applicants who **actually repay** within each risk category. This doesn't mean there should be equal proportions of men and women across different risk categories. For instance, if women have actually had higher repayment rates than men, there may be more women than men in the low risk category.

Unfortunately, these different measures may often be incompatible with each other, and therefore you need to consider any conflicts carefully before selecting any particular approach(es). For example:

- equal calibration is incompatible with outcome or error parity, except in rare cases where the actual distribution of outcomes is equal between different protected groups; and
- attempting to achieve outcome parity while removing protected characteristics, as required by anti-classification measures, may result in the learning algorithm finding and using irrelevant proxies in order to create a model which equalises outcomes, which may be unfair.

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

Most of the techniques discussed above require you to have access to a dataset containing personal data of a representative sample of the population. For each person represented in the data, you need labels for the protected characteristics of interest, such as those outlined in the Equality Act 2010. You could then use this dataset containing protected characteristics to test how the system performs with each protected group, and also potentially to re-train the model so that it performs more fairly.

Before doing this kind of analysis, you need to ensure you have an appropriate lawful basis to process the data for such purposes. There are different data protection considerations depending on the kinds of discrimination you are testing for. If you are testing a system for discriminatory impact by age or sex / gender, there are no special data protection conditions for processing these protected characteristics, because

they are not classified as 'special category data' in data protection law. You will 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).

You should also note that when you are dealing with 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 your AI system uses biometric data for testing and mitigating discrimination, 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 does not come under Article 9. The data is still regarded as biometric data under the GDPR, but is not 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).

However, some of the protected characteristics outlined in the Equality Act **are** classified as special category data. These include race, religion or belief, and sexual orientation. They may also include disability, pregnancy, and gender reassignment in so far as they may reveal 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.

If you are testing an AI system for discriminatory impact on the basis of these characteristics, you are likely to need to process special category data. In order to do this lawfully, in addition to having a lawful basis under Article 6, you need to meet one of the conditions in Article 9 of the GDPR. Some of these also require additional basis or authorisation in UK Law, which can be found in Schedule 1 of the DPA 2018.

Which (if any) of these conditions for processing special category data are appropriate depends on your individual circumstances.

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 is discriminating by religious or philosophical beliefs.

It 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.

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 available on the market 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.

Finally, if the protected characteristics 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.

Relevant provisions in the legislation

See Article 9 and Recitals 51 to 56 of the GDPR (external link)

See Schedule 1 of the DPA 2018 (external link)

Further reading – ICO guidance

Read our guidance on purpose limitation and special category data in the Guide to the GDPR.

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 in order 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, in order to ensure members of potentially disadvantaged groups are protected.

For instance, if you were using an AI system to sort 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 discriminate against them. Not including disability status as an input to the automated decision could mean the system is more likely to 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.

This approach amounts to making decisions about individuals, in a solely automated way, with significant effects, using special category data. This is prohibited under the GDPR unless you have explicit consent from the individual, or you can 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 an 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 instance, 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 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 and/or special category data in order to make predictions, and actively monitor this possibility throughout the lifecycle of the system. If the potentially inferred characteristics are special category data, you should ensure that you have an appropriate Article 9 condition for processing.

As noted above, if such a model is being used 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 – ICO guidance

Read our guidance on special category data for more information.

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.

Establishing clear policies and good practices for the procurement and lawful processing of high-quality training and test data will be 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 requires trade-offs (see the section 'What are AI-related trade-offs and how should we manage them?'). 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. 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.

Other resources

Equality Act 2010 (External link)

European Charter of Fundamental Rights (External link)

Example of controls

Risk Statement

Discriminatory output or decisions made by AI systems could lead to statistically inaccurate /unfair decisions for individuals from certain groups.

Preventative

- Put in place a data governance framework, which describes how all personal data used for ongoing training, testing or evaluation of an AI system is correct, accurate, relevant, representative, complete and up-to-date as possible.
- Ensure AI developers have completed training and associated competency assessments so they can identify and address bias and discrimination in AI systems.
- Provide training for key stakeholders and document how the relevant personnel were identified (eg senior management, risk managers, audit).

- Document your access management controls and segregation of duties for the development and deployment of AI systems to ensure changes affecting statistical accuracy are made and signed off by authorised persons. Maintain evidence of how these controls are monitored.
- Document levels of approval authority for the development/use of AI systems. Maintain evidence of appropriate approval.
- In your DPIA, include a thorough assessment of the risk of discrimination and the mitigants / controls in place to prevent it
- Maintain documented policies / processes for dealing with third parties and evidence of the due diligence completed.
- Maintain a documented process for the cross-section / peer review of AI system design. Maintain evidence the review was completed.
- Maintain a documented policy / process for performing pre-implementation testing of any AI systems or changes prior to go-live. Maintain evidence that testing was completed and the results of the test(s).
- Document levels of approval and attestation to the diversity / representation of the training / test data prior to use within the AI system. Maintain evidence of the appropriate approval.

Detective

- Regularly monitor for algorithmic fairness using appropriate measures.
- Document levels of approval and attestation to the diversity / representation of the training / test data prior to use within the AI system. Maintain evidence of the appropriate approval.
- Regularly review model performance against most recent data.

Corrective

- Add or remove data about under / overrepresented groups, including thorough analysis / justification.
- Retrain the model with fairness constraints.
- Retrain model designers in relation to discriminatory model performance.

How should we assess security and data minimisation in AI?

At a glance

AI systems can exacerbate known security risks and make them more difficult to manage. They also present challenges for compliance with the data minimisation principle.

AI can exacerbate known security risks and make them more difficult to manage. Two security risks that AI can increase are:

- the potential for loss or misuse of the large amounts of personal data often required to train AI systems; and
- the potential for software vulnerabilities to be introduced as a result of the introduction of new AI-related code and infrastructure.

By default, the standard practices for developing and deploying AI involve processing large amounts of data. There is a risk that this fails to comply with the data minimisation principle. A number of techniques exist which enable both data minimisation and effective AI development and deployment.

In detail

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 X below).

Further reading – ICO guidance

Read our guidance on security in the Guide to the 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 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's 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 and/or relationships with suppliers. Also, your existing systems need to be integrated with several other new and existing IT components, which are also intricately connected.

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.

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

Case study: losing track of training data

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

Example

An organisation decides to use an AI system offered by a third-party recruiter as part of its hiring process. To be effective, the organisation needs to share data about similar previous hiring decisions (eg sales manager) with the recruiter.

Previously, the organisation used an entirely manual CV scanning process. This led to some sharing of personal data (eg candidates' CVs) but, it did not involve the transfer of large quantities of personal data between the organisation and the recruiter.

The organisation must ensure that it has an appropriate lawful basis for this processing.

Beyond this, sharing the additional data could involve creating multiple copies, in different formats stored in different locations (see below), which require important security and information governance considerations such as:

- the organisation may need to copy HR and recruitment data into a separate database system to examine and select the data relevant to the vacancies the recruitment firm is working on;
- the selected data subsets will need to be saved and exported into files, and then transferred to the recruiter in compressed form;
- upon receipt the recruiter could upload the files to a remote location, eg the cloud;
- once in the cloud, the files may be loaded into a programming environment to be cleaned and used in building the AI system;
- once ready, the data is likely to be saved into a new file to be used at a later time; and
- for both the organisation and the recruiter, each time data is copied and stored in different places, there is an increased risk of a personal data breach, including unauthorised processing, loss, destruction and damage.

In this example, all copies of training data will need to be shared, managed, and when necessary deleted in line with security policies. While many recruitment firms will already have information governance and security policies in place, these may no longer be fit-for-purpose once AI is adopted, and should be reviewed and, if necessary, updated.

What should we do in this circumstance?

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.

Case study: security risks introduced by 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 the organisation 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 circumstance?

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 need to ensure that your staff have appropriate skills and knowledge to address these security risks.

Additionally, if you develop ML systems you can further mitigate security risks associated with third party code by separating the ML development environment from the rest of your IT infrastructure where possible.

Two ways to achieve this are:

- by using 'virtual machines' or 'containers' - emulations of a computer system that run inside, but isolated from the rest of the IT system. These can be pre-configured specifically for ML tasks. In our recruitment example, if the ML engineer had used a virtual machine, then the vulnerability could have been contained; and
- many ML systems are developed using programming languages that are well-developed for scientific and machine learning uses, like Python, but are not necessarily the most secure. However, it is possible to train an ML model using one programming language (eg Python) but then, before deployment, convert the model into another language (eg Java) that makes making insecure coding less likely. To return to our recruitment example, another way the ML engineer could have mitigated the risk of a malicious attack on CV filtering model, would have been to convert the model into a different programming language prior to deployment.

Further reading – ICO 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.

Other resources

Guidance from the National Cyber Security Centre (NCSC) on maintaining code repositories may also assist you.

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 update we will 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 goes beyond generic inferences about individuals with similar characteristics.

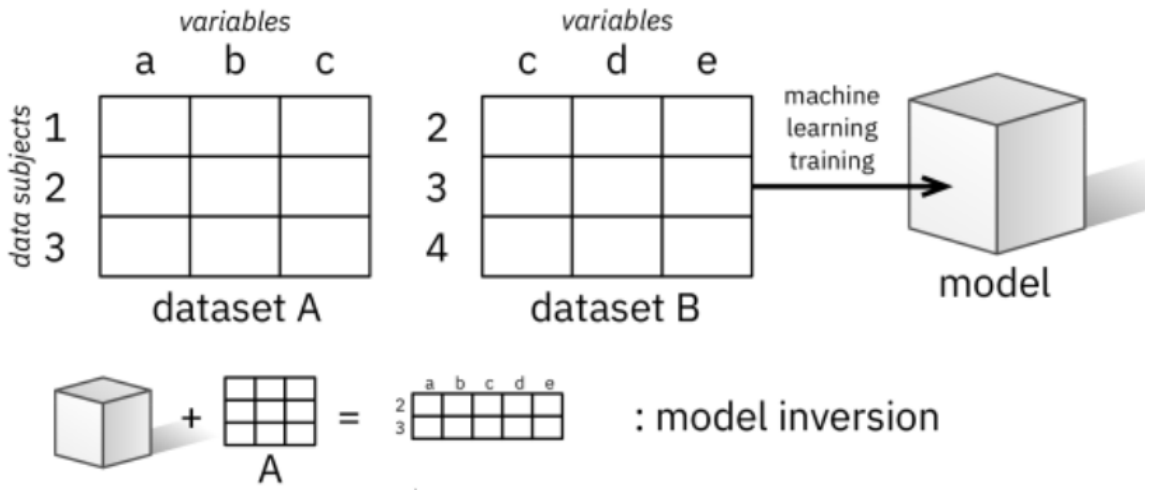


Figure 1. Illustration of model inversion and membership inference attacks, reproduced from [Veale et al. 'Algorithms that remember: model inversion attacks and data protection law']

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.

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'

Other resources

'Algorithms that remember: model inversion attacks and data protection law'

Simple demographics often identify people uniquely

'Model inversion attacks that exploit confidence information and basic countermeasures'

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 instance, 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 respectively.

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 instance, 'support vector machines' (SVMs) and 'k-nearest neighbours' (KNN) models contain some of the training data in the model itself.

In such 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 such models would be there by design and easily retrievable by the third party. Storing and using such models therefore constitutes processing of personal data and as such, the standard data protection provisions apply.

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. 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 would not be subject to white-box attacks in this way, because the API's users would 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 ensuring that the personal data used to train your models is not 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, mentioned above), this is a transfer of personal data, and you should treat it as such.

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 in and of themselves 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 instance, 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 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 NIS Directive 2018. 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 – ICO guidance

For more information about the NIS Regulations, including whether or not you qualify as a relevant digital service provider, read our Guide to NIS.

Example of controls

Risk Statement

The infrastructure and architecture of AI systems increases the likelihood of unauthorised access, alteration or destruction of personal data.

Preventative

- Subscribe to security advisories to receive alerts of vulnerabilities.
- Comply to or assess an AI system against external security certifications or schemes.

- Subject software to a quality review where one or more individuals view and read parts of its source code. At least one of the reviewers must not be the author of the code.
- Document policy / process for the separation of the AI development environment from the rest of the IT network / infrastructure. Evidence that the separation has been adhered to / happened.
- Have an approach for asset management to ensure a coordinated approach to the optimisation of costs, risks, service/performance and sustainability.
- Document contracts with third parties are clear about the role and responsibilities of third parties.
- Document policies / processes for dealing with third parties and evidence of the due diligence of information security completed.
- Document policy / processes for breach reporting and escalation.
- Adhere to the policy / process.
- Have a model governance policy.
- Assess more secure implementations of the trained model, and implement them as appropriate, post development but pre-deployment.
- Have processes in place to review the latest privacy enhancing techniques, assess the technique's applicability to their context, and implement it where appropriate.
- Document a DPIA, including thorough assessment of the security risks and the mitigants / controls to reduce the likelihood and impact of an attack.
- Have an API access policy in place which monitors volume and patterns of requests to identify and report suspicious activity.
- Ensure staff are trained to understand the breach reporting policy and what procedures to follow.

Detective

- Monitor API requests to detect suspicious requests and take action as a result.
- Regularly test, assess and evaluate the effectiveness of any security measures they have put in place (eg through techniques such as penetration testing).
- Monitor complaints monitoring and take action as a result, including broader analysis to identify other individuals who may be impacted.

Corrective

- Evidence changes made to AI system design, including analysis / justification to reduce the risk of future attacks.

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

What considerations of 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 GDPR says

Quote

'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 their 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 and/or implement systems operated by third parties, 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, eg making more statistically accurate and non-discriminatory ML models. See the trade-offs section for more detail.

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.

Relevant provisions in the legislation

See Article 5(1)(c) and Recital 39, and Article 16 (right to rectification) and Article 17 (right to erasure) of the GDPR (external link)

Further reading – ICO guidance

Read our guidance on the data minimisation principle in the Guide to the GDPR.

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 instance, 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, etc.). 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.

Further reading – ICO 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'; 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 measure the effectiveness of these privacy-enhancing techniques in balancing the privacy of individuals and the utility of a ML system, mathematically 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 in your particular context, before attempting to implement them.

- **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 instance, 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.

- **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

For more information on controllership in AI, read the section on controller/processor relationships.

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

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, and/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 instance, 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 facial recognition models that use them – and 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 mitigate the risks involved in sharing predictor variables 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 instance, 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

See 'Privad: practical privacy in online advertising' [external link] and 'Targeted advertising on the handset: privacy and security challenges' [external link] 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

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 instance 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

The European Union Agency for Network and Information Security (ENISA) has a number of publications about PETs, including research reports. (external link)

Example of controls

Risk Statement

AI developers do not properly assess the adequacy, necessity and relevance of the personal data used for AI systems for AI systems, resulting in non-compliance with the data minimisation principle.

Preventative

- Document levels of approval authority for the development/use of AI systems including the personal data sets included within the model. Evidence of appropriate approval.
- Review personal data relevance at each stage of model development, including detailed justification for the retention of data and confirmation that irrelevant data have been removed / deleted.
- Separate data during the different stages of the AI lifecycle based on the conditions of the data minimisation principle.
- Document a retention policy / schedule and evidence that the schedule is adhered to (personal data is deleted in line with the schedule or retention outside of schedule is justified and approved).
- Carry out an independent review of model input / output specifically regarding the relevance of the personal data inputs.
- Document a DPIA, including thorough assessment of the PETs considered and why they were or were not considered appropriate.

Detective

- Periodic review(s) of the features within an AI model to check they are still relevant eg testing against other systems with fewer features to see if same results can be achieved, with a view to reducing the amount of personal data being processed.
- Monitor individual rights requests and complaints from individuals, including the action taken as a result (at both individual level and boarder analysis).
- Periodically assess whether the model remains compliant with data minimization processes when used by third parties.

Corrective

- Remove / delete non-required features.
- Select a less invasive model, including thorough justification for the change.
- Remove / erase training data which is no longer required (eg, because it is out of data and no longer predictively useful).
- Implement appropriate PETs.

How do we enable individual rights in our AI systems?

At a glance

The way AI systems are developed and deployed means that personal data is often managed and processed in unusual ways. This may make it harder to understand when and how individual rights apply to such data, and more challenging to implement effective mechanisms for individuals to exercise those rights.

In detail

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 the considerations you may encounter when developing and deploying AI and attempting to comply with the individual rights of information, access, rectification, erasure, and to restriction of processing, data portability, object (rights referred to in Articles 13-21 of the GDPR). It does not cover each right in detail but discusses general challenges to facilitating 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 enable individual rights requests for training data?

When creating or using ML models, you invariably need to obtain data to train those models.

For instance, 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 enabling 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.

However, 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 instance, transforming numbers into values between 0 and 1) is often referred to as 'pre-processing'. This can create confusion regarding 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 instance, the training data in a purchase prediction model might include a pattern of purchases unique to one customer.

In this example, if a customer were to provide 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 that individual.

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.

You should consult our guidance on determining what is personal data for more information about identifiability.

- **Right of access**

You should not regard requests for access, rectification or erasure of training data as manifestly unfounded or excessive just because they may be harder to fulfil or the motivation for requesting them may be unclear in comparison to other access requests you might typically receive.

You do not have to collect or maintain additional personal data just to enable you to identify individuals within training data for the sole purposes of complying with the GDPR (as per Article 11). There may be times, therefore, when you are not able to identify the individual in the training data (and the individual cannot provide additional information that would enable their identification), and you therefore cannot fulfil a request.

- **Right to rectification**

The right to rectification may also apply to the use of personal data to train an AI system. The steps you should take with respect to rectification depend on the data you process as well as the nature, scope, context and purpose of that processing.

In the case of training data for an AI system, one purpose the processing may be to find general patterns in large datasets. In this context, individual inaccuracies in training data 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 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 an individual's personal data from the training data is unlikely to affect your ability to fulfil the purposes of training an AI system. 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 would therefore apply 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 would not be 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 should inform individuals if their personal data is going to be used to train an AI system. 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.

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 these decisions when you are taking them. However, you will 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 instance, 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 instance, you could provide public information explaining where you obtained the data from that you use to train your AI system.

How should we enable 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 instance, the product offers a customer sees on a website might be driven by the output of the predictive model stored in their profile. Where such data constitutes personal data, it is subject to the rights of access, rectification, and erasure. 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 is 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 instance, 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 – ICO guidance

Read our guidance on individual rights in the Guide to the GDPR, 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 section 3.2, this could happen for two reasons; by design or by accident.

How should we fulfil requests regarding 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 ends 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 such cases, it might 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.

If you have a well-organised model management system and deployment pipeline, accommodating such requests and re-training and redeploying your AI models accordingly should not be prohibitively costly.

How should we fulfil requests regarding 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 enable individual rights relating to solely automated decisions with legal or similar effect?

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 GDPR if the lawful basis for such 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. Guidance published by the European Data Protection Board (EDPB) states that human intervention should involve a review of the decision, which:

Quote

"must be carried out by someone who has the appropriate authority and capability to change the decision"

The review should also include:

Quote

"a thorough assessment of all the relevant data, including any additional information provided by the data subject."

The conditions under which human intervention qualifies as meaningful are similar to those that apply to those which render a decision non-solely automated (see previous section). 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.

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.

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 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 ExplAIn 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, to intervene on, to review, or to put an alternative point of view against.

For instance, 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.

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 is communicated 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 will 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 – ICO guidance

Read our guidance on Documentation in the Guide to the GDPR.

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 GDPR.

WP29 published guidelines on automated decision making and profiling in February 2018. The EDPB endorsed these guidelines in May 2018.

How should we explain the logic involved in an AI-driven automated decision?

Individuals also have the right to meaningful information about the logic involved in an AI-driven automated decision.

For more detail on how to comply with this right, please see our recent ExplAIn guidance, produced in collaboration with The Alan Turing Institute (The Turing).

Example of controls

Risk statement

Infrastructure and architecture of AI systems inhibits the ability to recognise, respond to and act upon individual rights requests.

Preventative

- Ensure your AI systems developers receive training, which includes the requirement to consider individual rights (IR) at the offset.
- Set and document levels of approval authority for the development/use of AI systems including consideration of IR in model. Maintain evidence of appropriate approval.
- Implement and document a policy / process for dealing with IR requests in the data processing pipeline, in particular defining the circumstances you would and wouldn't respond eg data used to train versus output and where there is an impact on the individual.
- Provide individual rights training for 'customer facing' individuals, including how to escalate more complex requests.
- In your DPIA, include a thorough data flow mapping.
- Consider data indexing and making your systems searchable using common identifiers as part of the system design if IR requests are anticipated.
- For fully automated decision making, ensure that humans who may be required to investigate a decision have the skills and tools and autonomy to investigate and override the decision.
- Ensure data subjects are informed about the processing of their data for purposes of training an AI system, or for profiling them. In cases where the data used for training is from another organisation and you do not have a direct relationship with the data subject, and where informing them directly would involve disproportionate effort, ensure that you make this information publicly available, and that the organisations you source the data from have processes in place to inform the data subjects about the processing.
- Ensure individuals are given the means to provide additional data in order to be identified within your AI systems.
- Maintain documented policies / processes for dealing with third parties, in particular, roles and responsibilities (controller / processor) are clear.

Detective

- Conduct peer reviews to ensure all actions have been completed as required.

- Conduct a periodic review of sample IR requests to ensure accurate and complete and where declined the justification (eg manifestly unfounded) was appropriate.
- Systematically monitor the time taken to respond to requests in order to identify systems which are potentially more complex.
- Submit 'dummy' IR requests to test the process, and measure the outcomes.

Corrective

- Re-design the AI system / data storage / indexing of training data.
- Consider if additional data could help to identify data subjects in the case of a request.
- Correct any inaccurate personal data and contextualise inferred data so that it is not misleading as to a matter of fact.
- Delete personal data if required.
- Re-train employees responsible for the identification and execution of IR requests.
- Retrain the humans / reassessment of resource requirements (eg if humans are pressured to make x decisions in y time.
- Redesign the AI, eg simplification / inclusion of warnings / pop-ups.
- Select a more appropriate model, including thorough justification for the change.
- Re-review / overturn decisions (eg one rogue reviewer), and any action taken as a result, including a broader assessment of the impacts on individuals.

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. This infringes Article 22 of the GDPR.

To mitigate this risk, you should ensure that people assigned to provide human oversight remain engaged and 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 can 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 fully automated, AI-driven decision making is generally more or less risky than AI-supported human 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 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 which only **support** human decision-making are not subject to these conditions. As a result of these restrictions and safeguards, automated decision-making arguably carries a higher risk than human decision-making (even though it may in some cases mitigate some of the risks of human decision-making).

If you decide to use AI only to **support** human decision-making, 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 human input needs to be **meaningful**. 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 oversight functions, 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.

Relevant provisions in the legislation

See GDPR Article 22 and Recital 71 (external link)

See DPA 2018 Sections 14, 49 and 50 (external link)

Further reading – ICO guidance

Read our guidance on automated decision making and profiling in the Guide to the GDPR

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 GDPR.

WP29 published guidelines on automated decision making and profiling in February 2018. The EDPB endorsed these guidelines in May 2018.

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 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 non-solely 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 partly or fully automated AI application will meet your needs best; or
- believe that a fully 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 (eg business owners, data scientists and oversight functions if you have them) should work together to develop design requirements that support a meaningful human review from the outset.

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 instance, the AI system could consider quantitatively measurable properties like how many years' 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 '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 process, and behaviours of human reviewers and enable them to effectively intervene. It may therefore be helpful to consult and test options with human reviewers early on.

However, the features 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 instance, 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. 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 non-solely 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 focussed 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 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 – ICO guidance

Read our explAIn draft guidance for more on methods of explaining and interpreting AI systems.

Example of controls

Risk Statement

AI systems incorrectly classified as not fully automated result in a lack of meaningful human oversight and the potential for non-compliance with DP legislation.

Preventative

- Implement and document a policy / process about the classification of AI systems in relation to Article 22, including level of approval authority. Maintain evidence of decision-making process and appropriate sign-off / approval.
- Provide training for humans employed to provide meaningful oversight including the ability to challenge the AI system decision and provide an independent review.
- Ensure AI system developers have understood the skills, experience and ability of human overseers when designing the AI system.
- Include in pre-implementation testing an assessment of human oversight to ensure it is meaningful.
- Set and document levels of approval authority for the development/use of AI systems, in particular in relation to model complexity to ensure human reviewers can interpret and challenge. Maintain evidence of appropriate approval.
- Conduct and document analysis of the time expected for a human to meaningfully review.

Detective

- Conduct post-implementation testing, and document the results of the testing and action(s) taken as a result.
- Test a sample of decisions to ensure the human is making the right decision. Document such tests, including how the sample was selected / criteria used.

- Monitor the decisions made by AI and compare them to human decisions, and document any action taken as a result of performance which goes outside of defined tolerances.
- Conduct and document 'mystery shopping' exercises, where you periodically provide deliberately misleading data that the human should disagree with the AI, to ensure their input is meaningful.
- Monitor individual rights requests and complaints from individuals, in particular, any relating to Article 22, including the action taken as a result (at both individual level and boarder analysis).
- Conduct a periodic assessment of human confidence in overturning the AI outcome.
- Monitor individuals' performance to identify outliers and the action taken as a result.

Corrective

- Re-train the human decision makers, reassess their resource requirements (eg if humans are pressured to make too many decisions in a short space of time).
- Re-design the AI, eg simplification / inclusion of warnings / pop-ups.
- Select a more appropriate model, and include a thorough justification for the change.
- Re-review or overturn decisions (eg if you have one rogue reviewer), and any action taken as a result, including a broader assessment of impacts to individuals.