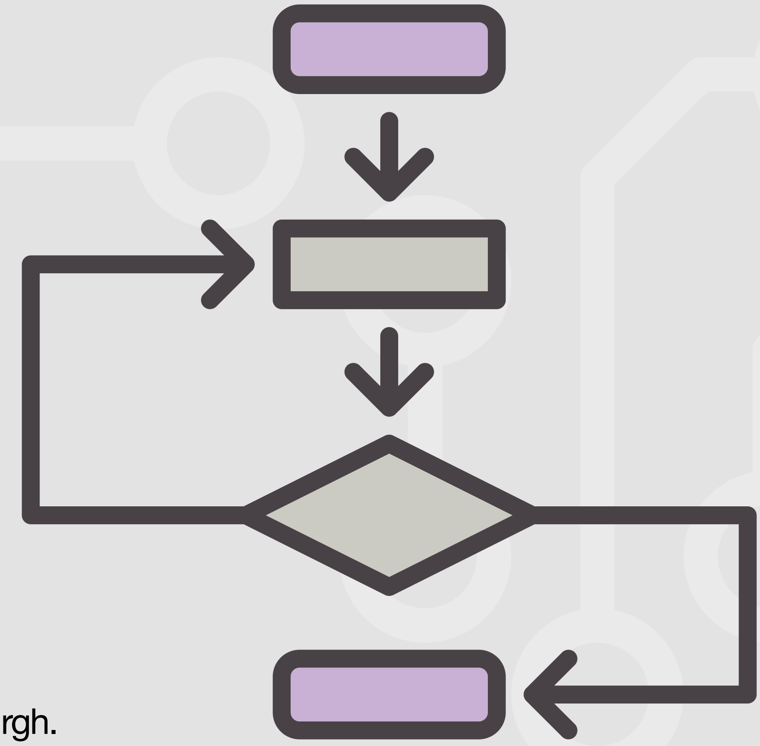
The Scottish Parliament
Pàrlamaid na h-Alba

Scotland's Futures Forum
Fòram Alba air Thoiseach

Scrutinising the use of artificial intelligence

A TOOLKIT



This paper provides a set of questions that can be used to ensure that the key issues in the design and use of AI have been considered and addressed.

It is intended to help ensure that AI is used appropriately and safely.

Robbie Scarff, University of Edinburgh.
In conjunction with Scotland's Futures Forum.

7 Key Questions

1 Intended Goals: What do you want to do and how will AI help you achieve that?

Having clear, measurable goals is crucial for AI systems in two respects and it requires careful thought to determine what these should be.

Firstly, there must be clear goals that the AI is being used to achieve, and a strong justification for using AI tools to achieve them.

Secondly, the AI system must be deployed with clear, technical goals by which to measure its performance.

2 Scientific Validity: What evidence do you have that this AI system can do what you intend it to do?

It is critical that the AI system is scientifically valid – that it can measure, analyse and produce results which are consistent with relevant scientific research and within AI's true capabilities.

If established scientific research does not back up the approach taken by the AI system, there is a greater chance that it will not work as intended.

3 Training Datasets: Which dataset or sets were used to train this AI system?

Most AI systems work by taking in new data from the real world, including from human users, analysing it according to historical data from the dataset it was trained on, and using this information to predict or classify the input data, or generate a new output.

AI therefore relies fundamentally on the quality, relevance and adequacy of the dataset used to train it. This has a significant impact on the quality and accuracy of the output.

4 Legal Issues: What legal issues might the use of this AI system raise?

The use of AI invariably raises legal issues due to the various contexts in which it is used (with potentially significant impacts on individuals), its use of personal data, and its impact on human rights.

These must be anticipated and appropriately dealt with in order to reduce the chance of individuals or groups being harmed.

5 Ethical Issues: What ethical issues might the use of this AI system raise?

Just because something is legal does not mean it is ethical, and AI raises many ethical questions over and above legal issues. These include unjust bias, low accuracy, environmental impacts, opacity (of the AI system or its training dataset), and impact on particularly vulnerable and/or marginalised groups.

Due to the relatively novel nature of AI, its complexity, and its use in a range of sectors, it often requires concerted effort from technical, ethical and sector specific experts to identify and attempt to resolve the complex ethical questions AI raises.

6 Transparency: How will you scrutinise and review the operation and impacts of this AI system?

Due to their complex nature, AI systems can be inherently difficult to scrutinise, even for experts. This can cause problems such as it being difficult to spot when the AI system is not working as intended, potentially leading to harmful impacts on people.

Even if people are not being harmed, improper operation of the AI system can result in the intended goals not being met.

7 Complaints and Redress: If people want to complain, particularly if they feel they have suffered harm, how will they do so?

AI systems can result in unintended, sometimes harmful, consequences for individuals or groups. It is therefore essential that people have a means of raising and resolving complaints.

This may be done by raising a complaint through a dedicated person within an organisation, such as a data protection officer, or through a public body or ombudsman, such as the Information Commissioner’s Office. This helps to ensure people are treated as fairly as possible, and to identify problematic issues early.

Introduction

At its simplest, artificial intelligence [AI] refers to using data processing and computer science techniques to solve problems. AI systems work by taking new data from the real world, analysing it according to the historical data it was trained on, and using this information to classify the input, produce new results or make predictions.

For example, an image recognition AI system for analysing mammograms would be trained using historical mammogram data before analysing new mammograms and classifying them by risk of malignancy based on the historical data.[1] AI can be used to make or inform decisions. In this case, the AI is used as a tool to help inform doctors' decisions regarding future care.

The number of ways in which AI is being used throughout the world is already significant and continues to grow. AI is being used in education to assess and grade students, and to monitor their behaviour. It is used in the workplace to evaluate job candidates and inform employee performance reviews, as well as being used in policing and security to identify wanted individuals via facial recognition and to assess risk of reoffending. AI is also being used in the public sector for tasks such as determining where resources are best allocated and improving public services.

In 2021, the Scottish Government, supported by The Data Lab, published its AI Strategy, which sets out guiding principles and actions for the development of AI that is trustworthy, ethical, and inclusive.

While aiming to use the AI Strategy to develop regulation, Scotland does not currently have any AI-specific regulation in place. However, the Scottish Government adheres to relevant existing UK regulatory frameworks, and all public bodies in Scotland adhere to the ICO's GDPR legal requirements on the use of data[2].

This toolkit provides a set of potential questions for Members of the Scottish Parliament [MSPs] and others with an interest in the use of AI to scrutinise the decision to implement, or potentially implement, an AI system. It can also be used to regularly assess an AI system for as long as it is in operation.

Seven key areas are covered: intended goals, scientific validity, training datasets, legal and ethical issues, transparency, and complaints/redress. For each area, an initial question is provided along with follow-up questions and background information.

This is the first published version of this toolkit and the author's thanks go to those who have informed its development.

To provide feedback on it, suggest improvements or find out more about the work, please visit **www.scotlandfutureforum.org** or email **ScotlandsFuturesForum@parliament.scot**.

Robbie Scarff, University of Edinburgh.
In conjunction with Scotland's Futures Forum.

[1] An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload | Radiology (rsna.org)

[2] Policies and standards for the evaluation of Artificial Intelligence (AI) in government decision making: FOI release – gov.scot (www.gov.scot)

What do you want to do and how will AI help you achieve that?

Follow-up questions

- Could you achieve your goals without using AI? Have you compared using AI to possible alternatives by conducting an options appraisal?
- Have you publicly stated your goals in implementing this AI system?
- Have you considered possible unintended consequences and how you would deal with them?
- Who have you consulted about the use of this AI system?
- Have you consulted or included in the implementation process the group of stakeholders who will be most affected by your deployment of this system?
- Who does this AI system benefit most, and how? Who is most vulnerable to risk from this AI system's deployment or failure?
- Who else is involved in implementing this AI system? What are their goals and motivations, and could they result in conflicts of interest? Have you vetted or assessed stakeholders in terms of their previous delivery of AI systems and/or services?
- How will you regularly assess whether your AI system is achieving its intended goals without causing unintended or unduly harmful consequences?

Background

Having clear, measurable goals is crucial for AI systems in two respects and it requires careful thought to determine what these should be.

Firstly, there must be clear goals that the AI is being used to achieve, and a strong justification for using AI tools to achieve them. Secondly, the AI system must be deployed with clear, technical goals by which to measure its performance.

Technosolutionism refers to the view that human, social problems require or can have a technical solution. However, not all social problems require a technical solution, and most cannot be solved with technology alone. This means that it is crucial that full and careful consideration is given to alternative options.

Publicly stating the goals that the AI system is intended to achieve improves transparency and is a way for those implementing the AI system to hold themselves accountable.

When a new AI system is introduced, unintended consequences are almost inevitable due to the system's exposure to real world data, the complexity of which is only partially captured in the training data.

For example, Amazon developed AI for evaluating CVs with the intention of finding the best applicants. However, the developers eventually realised it was not ranking candidates for technical positions in a gender-neutral way because it was trained using the CVs submitted to Amazon over 10 years, which were primarily from males.

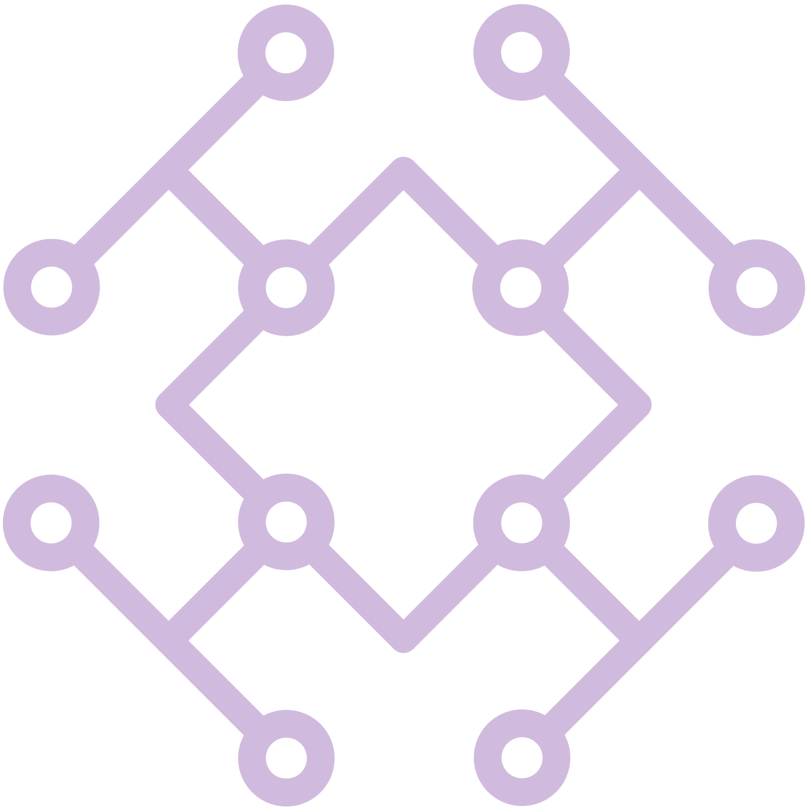
Amazon abandoned that tool, but it serves as an example that AI developers setting out to do one thing – find talented candidates – could unintentionally do something else: discriminate against women.[3]

It is therefore best to anticipate potential unintended consequences as much as possible and to consider ways of resolving them should they arise.

One way to identify possible unintended consequences is consulting the public, especially those most likely to be impacted by the system, and appropriate experts. This is also important for getting a sense of how the public feels about the proposed AI system.

If there is widespread and strong opposition to the proposed AI system, this may suggest a need for increased scrutiny. It is vital that any such public consultations should be conducted with a representative sample that gives appropriate weight to the impacted population.

Finally, when possible, it is good practice to conduct a thorough review of the outcomes of previous, similar AI deployments, including assessing the past performance of firms providing AI services. This should highlight any controversial or otherwise negative incidents, alerting decision-makers to the possibility of such things happening again.



[3] Amazon scraps secret AI recruiting tool that showed bias against women | Reuters

2 What evidence do you have that this AI system can do what you intend it to do?

Follow-up questions

- How did you determine if the AI system can do what it claims and you intend?
- How accurate is the AI system? In general? For specific populations? In different settings?
- How did you assess the accuracy of the AI system? How does this compare to how accuracy/performance is measured in the domain in question (even for human experts)?
- Have you decided upon an acceptable level of accuracy for the AI system and how was this level determined?
- How will you assess the validity and accuracy of the AI system while it is in use?

Background

It is critical that the AI system is scientifically valid – that it can measure, analyse, and produce results which are consistent with relevant scientific research and within AI's true capabilities.

If established scientific research does not back up the approach taken by the AI system, there is a greater chance that it will not work as intended.

There will be some variance in what is considered sufficient research to support or refute an AI's approach, but the users of AI must be prepared to back up their assertions on its scientific validity, respond to any concerns, and review any changes in the research base.

For example, it is claimed that some AI applications can determine emotional state from facial expressions, something which, according to established scientific research, they are simply unable to do.[4]

Taking the word of AI providers on what their systems can do is not enough. They are ultimately selling a product and it is not always in their interest to highlight potential flaws. The best practice is therefore to use independent experts who have experience in auditing such systems.

In terms of AI, accuracy refers to the number of data points the AI predicts correctly. In other words, it is a measure of how often the AI's predictions match some predefined benchmark measure for the task, which is assumed, not always correctly, to represent the ground truth.

For example, a "smile detection" AI can be 98% accurate if it correctly matches 98% of the images labelled "smile" in the benchmark set. However, if the benchmark dataset is poorly curated, the "98% accurate system" can be bad at detecting real smiles.

[4] Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements – Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, Seth D. Pollak, 2019 (sagepub.com)

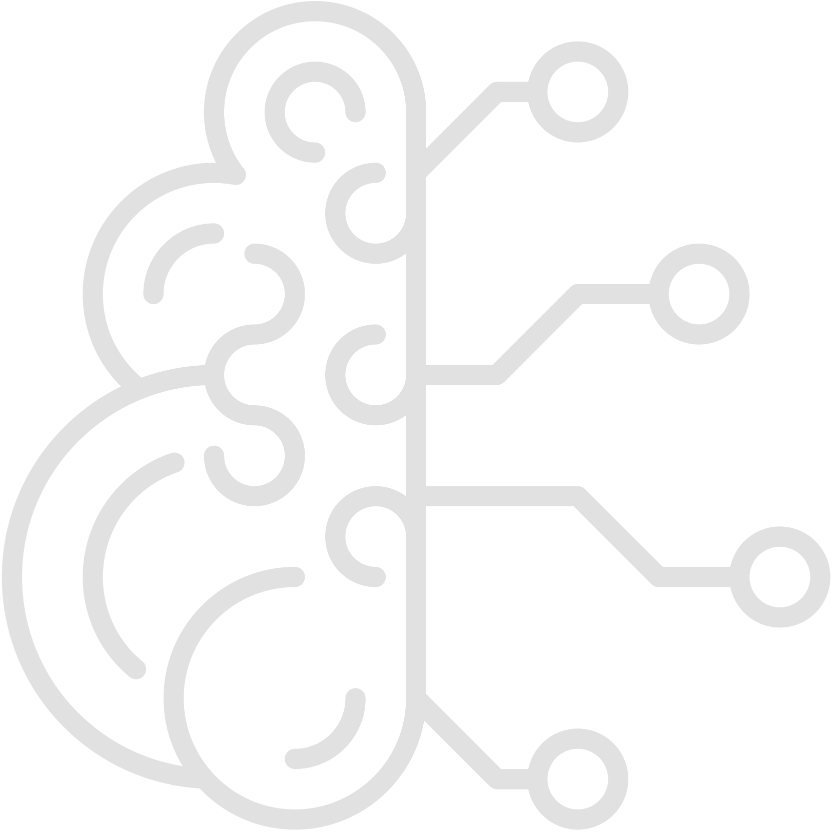
Determining the accuracy of AI is a developing and highly contested field. But when AI is used to make or support decisions about people that affect their lives in important, consequential ways, accuracy is a very important factor. It is therefore crucial that the acceptable level of accuracy for an AI system used in a specific context is given careful consideration.

Accuracy can also differ for different sections of the population. If the consequences for individuals and groups are trivial, it is less important that high levels of accuracy are ensured. However, if AI is used in contexts where the consequences for individuals, groups, or society are more significant or lasting, it is far more important that high levels of accuracy are ensured.

A clear example of this is facial recognition systems having poor accuracy rates in successfully identifying people from different ethnic and racial backgrounds.[5] This, in turn, can lead to issues like the wrongful arrest of people incorrectly identified, personal harm caused to individuals, and loss of community trust in key public services like policing.

It is worth bearing in mind that AI systems change and adapt over time, which can lead to problems developing down the line. This is especially so for machine learning systems which constantly alter their own process for working things out as they learn more about their environment and identify patterns in the data they analyse.

Ultimately, this means that AI systems that work well to begin with may develop inaccuracies over time (sometimes known as “model drift”) resulting in people potentially suffering harm. This is another reason why high-impact AI systems often warrant auditing both before and after they are deployed.



[5] Racial Influence on Automated Perceptions of Emotions by Lauren Rhue: SSRN

3 Which dataset or sets were used to train this AI system?

Follow-up questions

- How accessible is the training dataset? What is its source or provenance?
- Have you audited the training dataset for quality, including: bias, representativeness, accuracy, completeness, timeliness, and appropriateness of dataset to specific task?
- How will you mitigate or address quality issues with the data, such as unfair bias?
- Have you considered the security and anonymity risks that arise due to increased collection and use of data?
- Human labelling of datasets can lead to the introduction of biases and human error. Have you considered the impact of these issues on the quality of the training dataset(s)?
- Have you considered the problematic ethical issues raised by the working conditions of data labellers, such as poor pay and being exposed to extremely graphic, obscene images?

Background

AI systems work by taking new data from the real world, analysing it according to historical data from the dataset it was trained on, and using this information to predict or classify the input or produce new results.

AI therefore relies fundamentally on the dataset used to train it. This has a significant impact on the quality and accuracy of the output.

A poor-quality training dataset will result in a system which produces inaccurate and/or biased results, which can lead to discrimination. But even a dataset which is of good relative quality can contain unfair biases or gaps that need to be identified and mitigated.

It is therefore crucial to be able to audit the datasets used to train the AI system, as this will provide the best chance of spotting any causes for concern. However, such training datasets are often generated and owned by large companies, meaning they cannot always be easily scrutinised by external actors.

While not always the case, some AI systems are developed using the labour of poorly paid workers. A simple example is an image recognition AI system which can tell a cat from a dog from a fish and so on. Such systems can only do so after workers have manually labelled pictures of cats, dogs and fish, so that the system learns the distinguishing characteristics of each and can apply that in the real world.

A similar problem highlights further ethically problematic issues. Human workers who are poorly paid and have poor working conditions are often used to train the AI systems which detect obscene, graphic content posted on social media sites. Thus, they are also exposed to mentally distressing images.

Moreover, in cases which are not extreme and clear-cut, classifying images is a subjective exercise in which the individual worker’s cultural, social and religious environment may influence what is deemed permissible or not.

4 What legal issues might the use of this AI system raise?

Follow-up questions

- What steps will be taken to resolve any legal issues that arise?
- Who is accountable for addressing any legal issues that arise, especially where harm occurs? Who has a duty of care? Are the people responsible appropriately trained and resourced?
- Have you conducted a human rights impact assessment? Have you conducted a data protection impact assessment? If not, why?
- Have you consulted the appropriate regulator(s) regarding the requirements and safeguards that need to be put in place?
- Will implementing the AI system conflict with any of Scotland's or the UK's obligations under international law, such as the International Covenant on Civil and Political Rights, the Convention on the Elimination of Racial Discrimination, and the Convention on the Rights of the Child?
- How will you regularly review the legal issues raised by the AI system?

Background

The use of AI invariably raises legal issues due to the various contexts in which it is used (with potentially significant impacts on individuals), its use of personal data, and its impact on human rights.

These must be anticipated and appropriately dealt with in order to reduce the chance of individuals or groups being harmed.

AI may violate people's human rights in a wide variety of ways. AI used for predictive policing or hiring workers may result in discrimination. AI used to filter and moderate content online may violate the right to freedom of expression. AI used to make sentencing decisions may impact upon the right to liberty.

Harm may even occur before the AI system is deployed if people's personal data is used in a way that is not compliant with the GDPR, thus violating their data protection rights, as well as privacy.

AI can often be an adaptable tool. It can be deployed in one sector, demonstrate successful results, and then be used in other sectors. While this is not necessarily problematic, it is important to recognise that different contextual settings often come with different challenges and responsibilities for those deploying the AI system.

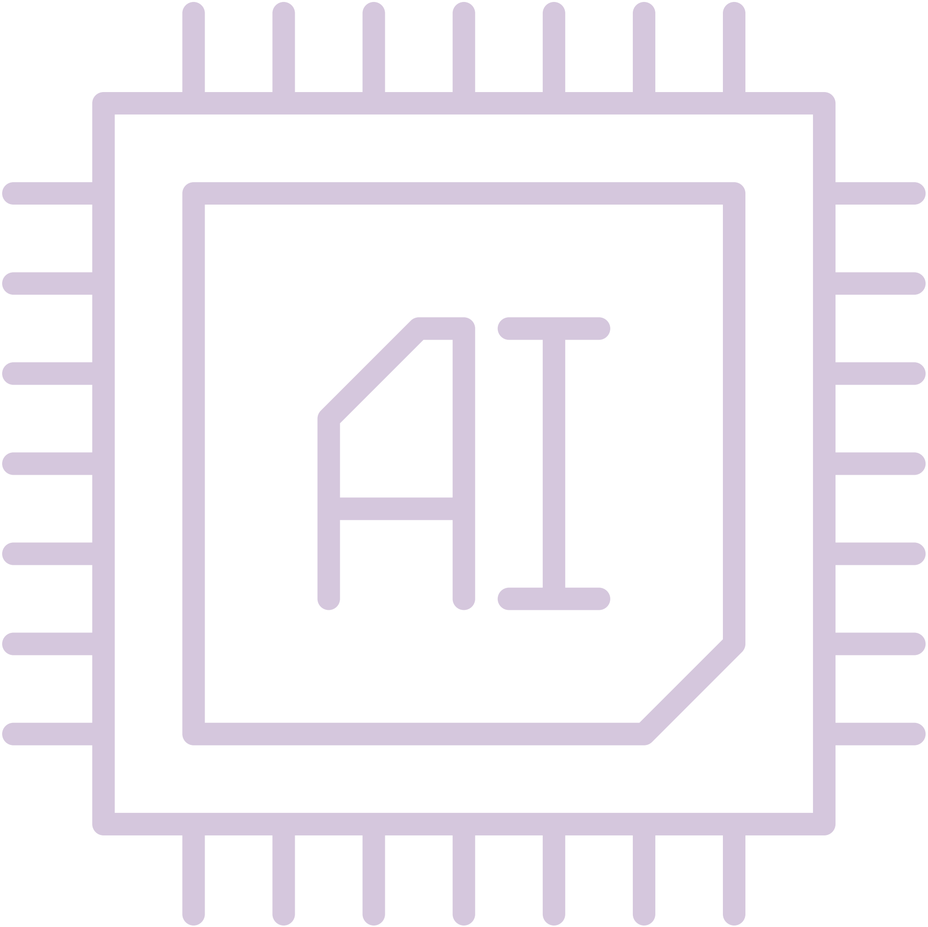
For example, schools, hospitals, prisons or public spaces all present unique contexts. Certain factors differ between each context, such as the capacities of those interacting with an AI system, expectations of privacy, and the appropriateness of use for the given context.

In essence, the fact that AI has been successfully deployed in one setting does not automatically mean it can or should be deployed in another. Rather, a careful case-by-case evaluation is required prior to deployment.

Conducting impact assessments are a very useful way of working through the potential issues that may arise in a methodical and comprehensive way. However, it is important that these are done not as a tick-box exercise but by appropriately trained and knowledgeable individuals who are supported to be as thorough as possible.

The legal issues raised by using an AI system are liable to change over time as it is used with different people, in different contexts and for different purposes. Equally, the AI system itself may also change in adapting to new data.

It is therefore appropriate to establish a system for regularly reviewing the legal issues raised by an AI system. Such a system could consist of strong mechanisms for post-deployment monitoring, incident reporting, and ecosystem level interventions around safety, as is the case in the aviation sector. This will help ensure any problematic issues are identified early and steps are taken to rectify them.



5 What ethical issues might the use of this AI system raise?

Follow-up questions

- What process was used to identify and assess ethical issues in the design and development of the AI system?
- Who is potentially harmed by any ethical issues or risks, and what is the nature, scale and likelihood of these harms?
- What is likely to be the impact on populations who are uniquely vulnerable? Have you conducted an equalities impact assessment?
- Who does this system benefit most, and how? Who is most vulnerable to risk from this system's deployment or failure?
- Have you consulted those most likely to be directly affected by the AI system?
- What environmental impact does the AI system have and how does this compare to alternatives?
- How will you ensure that humans retain meaningful control over the operation and impacts of the AI system?
- How will you continue to assess and address the ethical issues raised by the AI system?

Background

Just because something is legal does not mean it is ethical, and AI raises many ethical questions over and above legal issues. These include unjust bias, low accuracy, environmental impact, opacity (of the AI system or its training dataset), and impact on particularly vulnerable and/or marginalised groups.

Due to the relatively novel nature of AI, the complexity of how it operates, and its use in a range of sectors which each present their own challenges, it often requires concerted effort from technical, ethical, and sector specific experts to identify and attempt to resolve the complex ethical questions AI raises.

AI systems often require external, expert advice from people who have technical training in how the AI system operates and from people who are familiar with the context in which the AI system will be deployed. One of the best ways to identify ethical issues is to ask those who are likely to be directly affected by the system for their views on it.

Some ethical issues may only become apparent after the AI system has been put in place and been functioning for some time. Therefore, regular review of the impact and issues raised by the system is required to ensure it is working properly and not causing harm to those interacting with it.

AI systems can have varying levels of autonomy to make decisions and influence their environment. When humans are involved in this process in terms of overseeing and influencing what the AI does, this is generally referred to as having a "human in the loop". In contrast, having no human in the loop means an AI system can operate fully autonomously in the sense that it operates and impacts its environment without any real-time input or oversight from humans.

Responsible AI systems require consideration of when and how to implement human-in-the-loop protocols, or other forms of effective human control.

6 How will you scrutinise and review the operation and impacts of this AI system?

Follow-up questions

- Will people be told that they are interacting with an AI system?
- If the AI system makes decisions about people that affect important parts of their lives, will you provide an explanation to them about how that decision was made? Do you have the resources and expertise required to do this?
- If you identify a problem with the AI system which is or may lead to unfair, harmful outcomes for people, how will you assess potential impacts and inform the individuals affected if required?
- How will you make the procurement and implementation process as transparent as possible?
- Will your organisation have the access and expertise needed to examine, test or audit the system, or to understand its limitations?
- Will public expectations regarding the ability of AI be managed so that it is understood that AI can make errors?

Background

Due to their complex nature, AI systems can be inherently difficult to scrutinise, even for experts. This can present problems such as an inability to spot when the AI system is not working as intended, or to know why or how a certain result was reached, potentially leading to negative impacts on people.

Even if people are not being harmed, improper operation of the AI system can result in the intended goals not being met.

AI technology is not always well understood among the general public. As its use becomes more prevalent in different walks of life, public awareness will increase and people will have certain expectations regarding how it ought to be fairly and responsibly used, in ways they can trust. In order to scrutinise exactly how AI is delivered and operated, the process by which it is acquired and deployed must be as transparent as possible.

While it may be obvious in some cases, in others it can be hard for end users to know that they are interacting with an AI system. This may be because the system is so good at mimicking human behaviour, or simply that the AI operates in the background. For instance, if AI is used to assess someone's suitability for a loan, the individual may deal with a human operator, but it is AI that conducts the suitability assessment in the background.

As well as providing people with an explanation in the interests of fairness, it is also worth bearing in mind that under the UK GDPR people have the right not to be subjected to “solely automated decisions” which have a “legal or similarly significant effect” on them.[6]

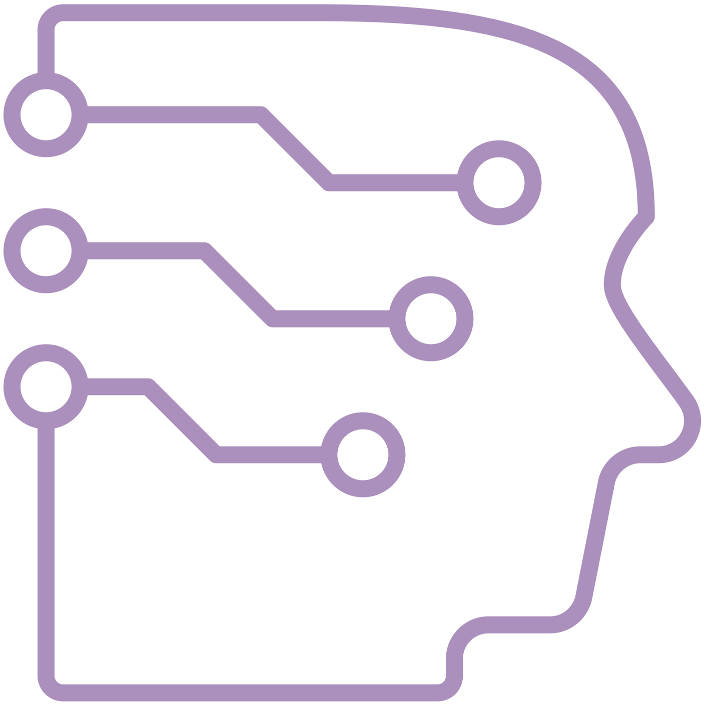
A solely automated decision would be when there is no human involved in the decision-making process at all. A legal effect is something which affects someone’s legal rights, while a significant effect is tricky to define but would include the likes of e-recruiting practices or automatic refusal of an online credit application.

Ultimately, for people to enforce their rights in this case they need to know that AI had been used, before they could contest exactly how it had been used.

Even if people are not harmed, improper operation of the AI system can result in the intended goals not being met. This can occur even if people are, in theory, overseeing the operation of the system if they defer to the results of the AI due to its complexity and inscrutability. This effect is known as automation bias.

A lack of transparency in AI can also arise due to an organisation’s inability to examine or audit proprietary third-party software, or from failures of an organisation to disclose an AI deployment.

The goal is not to achieve full technical transparency, as this is difficult and rarely useful anyway. Rather, the goal should be to achieve a clearer notion of accountability for how the AI system is used. This entails greater scrutiny of the humans involved and tasking them with creating tools which can provide useful evidence regarding the AI system’s operation.



[6] Rights related to automated decision making including profiling | ICO

7 If people want to complain, particularly if they feel they have suffered harm, how will they do so?

Follow-up questions

- How will you make people aware of their rights (e.g. human rights, data protection rights) in relation to the use of AI?
- Will individuals be able to seek redress if they feel they have been unduly harmed? What will this redress process consist of?
- Should the AI system receive a high number of complaints, how will you review how it is operating?

Background

AI systems can result in unintended, sometimes harmful, consequences for individuals or groups. It is therefore essential that people have a means of raising and resolving complaints.

This may be done by raising a complaint through a dedicated person within an organisation, such as a data protection officer, or through a public body or ombudsman, such as the Information Commissioner's Office. This helps to ensure people are treated as fairly as possible, and to identify problematic issues early.

It is important to remember that people have a range of rights, most notably data protection and human rights, when they are impacted by an AI system. People may often be unaware of what these rights are, when they may be being infringed, and how to enforce them.

AI systems can make mistakes and people can end up suffering a range of harms which can have a significant material impact on their life. In such cases, when things go wrong, it is essential that there are established processes through which people can raise complaints and have them fairly and comprehensively assessed.

While not definitive, a reasonably reliable indicator that an AI system is causing problems is a high number of complaints about it. This should signal to those deploying and operating of the system to investigate the issue and, if there is a problem, to rectify it. For this to happen, a complaints system must be in place and be visible and accessible to stakeholders.

There could be a process by which a certain number of complaints will trigger a review of the system. This need not be a predetermined number of complaints. Rather, this may be done in an ad hoc manner, applying the professional insight from those operating the system.

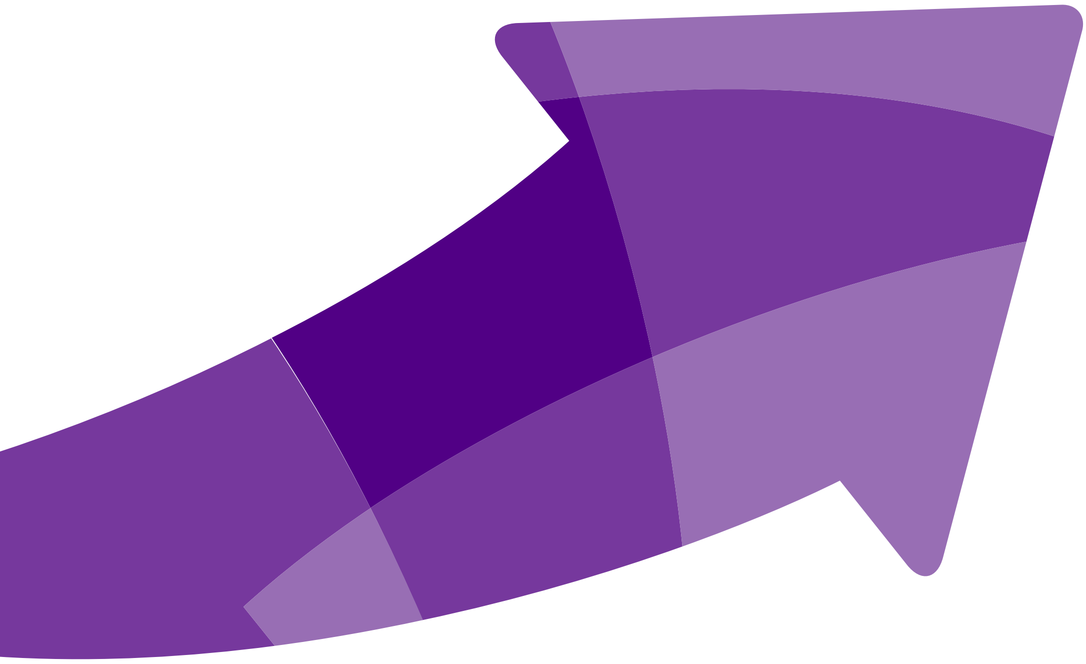
It may also be more appropriate to look at the specific nature of the complaints, such as assessing the severity of the impacts on people.

Thanks

The author sincerely thanks the following contributors who provided valuable advice during the development of this toolkit.

- **Morag Campsie and colleagues (Audit Scotland)**
- **Angus Evans (Scottish Parliament Information Centre)**
- **Rob Littlejohn (Scotland's Futures Forum)**
- **Professor Ram Ramamoorthy (University of Edinburgh)**
- **Lynn Russell (Clerk, Scottish Parliament)**
- **Professor Burkhard Schafer (University of Edinburgh)**
- **Professor Shanon Vallor (University of Edinburgh)**
- **Alison Wilson (Clerk, Scottish Parliament)**
- **Seán Wixted (Clerk, Scottish Parliament)**

The toolkit was produced for Scotland's Futures Forum during a PhD internship part-funded by the Scottish Graduate School of Social Science.



Website **www.scotlandfutureforum.org**

Twitter **@ScotFutures**

Email **ScotlandsFuturesForum@parliament.scot**