SLAUGHTER AND MAY/

DIGITAL
Part of the Horizon Scanning series

EXPLAINING AI: THE IMPORTANCE OF TRANSPARENCY AND EXPLAINABILITY

June 2020

Explaining AI

As Artificial Intelligence (or "AI") solutions become more prevalent, customers and regulators are demanding increased information regarding what this new technology does and how it is being used. In Europe there is also a strong focus at governmental level on the ethical deployment of AI, and transparency forms an important part of this. Being able to explain AI, particularly where it is used to make decisions about people (for example, whether or not credit is given to an individual) is often seen by regulators as essential for those organisations wishing to bring their customers, regulators and supply chain along with them on their AI journey. But is it always possible (or sensible) to explain AI? In this briefing we look at why explaining AI is important, and how (according to the UK's data regulator) organisations should go about explaining their AI use.

Why explaining AI is important

Being able to explain the AI being used, at least at some level, is seen by many as good business sense. For example, it:

- enables organisations to build consumer and regulator trust in their offering, and is in certain circumstances a regulatory requirement (e.g. it is a General Data Protection Regulation ("GDPR") requirement in the UK and EU – see below);
- improves an organisation's internal governance. Explaining the AI to affected individuals requires those within the organisation to understand the models, choices and processes used along with any AI decisions which are made. This gives the organisation more oversight and helps it ensure the AI systems meet the organisation's objectives; and
- can lead to better outcomes, as organisations identify and mitigate discriminatory outcomes which may be present in traditional systems and human decision making.

What is AI and what are AI assisted decisions or outputs?

AI is an umbrella term used to describe a range of technologies and approaches that try to mimic human thought to solve tasks. Examples include machine learning (a sub-set of AI) and natural language processing.

AI-based systems can be purely software based, acting in the virtual world (for example, voice assistants, search engines and speech or facial recognition), or can be embedded in hardware devices. Examples of these include autonomous cars, wearable technology and other internet of things applications.

There are several ways to build an AI system but each involves the creation of an algorithm that uses data to model an aspect of the world. It then applies this model to new data to make predictions.

Big data is therefore often intrinsically linked to AI. In its 2017 guidance on big data, AI and machine learning, the UK's data regulator (the Information Commissioner's Office, or "ICO") described big data as "an asset that is difficult to exploit" and AI as "the key to unlocking" its value.

The field of AI is generally divided into two categories: (i) general AI, which has broad applicability and could solve any tasks requiring human intelligence – this is not yet a reality; and (ii) narrow AI, which is basically algorithms that are designed to solve one or more particular problem.

Decisions and outputs made using AI can also be divided into categories. Outputs can be classed as predictions (e.g. you will not default on a loan), recommendations (e.g. you will like this advert) or classifications (e.g. this is spam), whereas decisions are either fully automated or involve human intervention (often referred to as having a "human in the loop").

That said, explainability does bring with it some challenges. Industry engagement carried out by the UK's data regulator (the Information Commissioner's Office or ICO)[1] in the UK has highlighted a number of issues which could limit the information organisations are willing to share regarding their use of AI. There are concerns that sharing too much information can actually lead to distrust due to the complex and sometimes opaque nature of AI. There may also be sensitivities around inadvertently disclosing commercially sensitive information about how an AI model or system works or that disclosing too much information may enable individuals to exploit the AI model, particularly where AI is used to identify wrongdoing or misconduct (such as fraud detection). In addition, trade-offs may need to be made, for example between a system's accuracy and its transparency. The ICO considers that these challenges can be mitigated, for example by using a data protection impact assessment (see below). In its view, organisations should start with the assumption that they will be as transparent as possible about the rationale of an AI system and work back from there, justifying and documenting where they consider it necessary to limit information.

How organisations can explain their use of AI – the ICO Approach

In the UK, the ICO has collaborated with the UK's national institute for data science and AI (The Alan Turing Institute or "the Turing") to look at how organisations can explain their AI use.

In April 2018, it was tasked (in the UK Government AI sector deal) to develop guidance with the Turing to assist in explaining AI decisions. They subsequently launched Project ExplAIn and published draft guidance for "Explaining decisions made with AI" in December 2019.[2] This draft guidance was open for consultation until 24 January 2020, and the final version was published in May 2020. Although not a statutory code of practice, it sets out good practice for explaining AI decisions to individuals and discusses the data protection provisions associated with this.

While the guidance is primarily relevant to those organisations caught by the GDPR (which can include non-EU organisations given the GDPR's extra-territorial scope), its practical approach means it is of interest to any organisation that decides (or is required) to explain its AI decision making processes. We therefore set out below details of why the GDPR, and the Project ExplAIn guidance, is relevant, and what the guidance covers.

Why is the GDPR relevant?

Within the EU and UK, the GDPR applies whenever an AI model processes personal data (see box "What is personal data?"). While some AI models do not use personal data, many use or create personal data both during the development phase and when in operation.

What does the GDPR require regarding AI explainability?

The GDPR is drafted in a technology-neutral manner, and so does not explicitly reference AI. However, it does contain specific provisions on large-scale automated processing of personal data (including profiling), which means it will apply where AI is used to make a prediction or recommendation about someone. For example, it gives individuals:

- a right to be informed of the existence of solely automated decision-making (including profiling) producing legal or similarly significant effects. An example may include where AI is used to determine whether or not an individual is granted credit. In such circumstances, the individual is entitled to receive meaningful information about the logic involved in the decision, as well as the significance and the envisaged consequences of such processing for them;
- a right of access in relation to that information, which includes the right to obtain an explanation of a solely automated decision after it has been made;
- a right to object to the processing of their personal data, specifically including profiling, in certain situations. For example, they have an absolute right to object to profiling for direct marketing purposes; and
- a right not to be subject to a solely automated decision producing legal or similarly significant effects, subject to certain exemptions. Where an organisation is relying on one of the exemptions, that organisation must adopt suitable measures to safeguard individuals, including the rights to obtain human intervention, to express their view and to contest the decision. There are also separate provisions in Parts 3 and 4 of the Data Protection Act 2018 for solely automated decision-making carried out for law enforcement purposes or by the intelligence services. For example, individuals have a right to obtain human intervention in these cases.

Even where an AI-assisted decision is not part of a solely automated process (because there is meaningful human involvement), the GDPR imposes general requirements to provide information to individuals whose data is being processed about that processing (Articles 12–14). In addition, the main GDPR principles (Article 5) will still apply, with the principles of fairness, transparency and accountability having particular relevance to explainability:

[1] Project ExplAIn Part1.

[2] See https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-and-the-turing-consultation-on-explaining-ai-decisions-guidance/.

- Fairness involves considering how an individual's interests are affected. If a decision is made using AI (whether solely automated, or merely AI-assisted) without some form of explanation about the decision, this is unlikely to be fair.
- Transparency is also about being clear and open with individuals about how and why their personal data is being used. The ICO considers it unlikely that processing will be considered transparent if an organisation is not open with individuals about how and why an AI-assisted decision about them was made, or where their personal data is being used to train and test an AI system. Privacy notices are often used to provide some of the necessary transparency, together with the general information that must be provided whenever personal data is processed (for example, around the purpose and duration of the processing).
- Accountability includes demonstrating compliance with the GDPR principles. One way to demonstrate that you have treated an individual fairly and in a transparent manner when making an AI decision about them is to provide them with an explanation of the decision and to document this.

In addition, the GDPR requires organisations to carry out a data protection impact assessment ("DPIA") when they are processing data using new technologies (like AI) which is likely to have a high risk to individuals (Article 35). DPIAs are also required where there is any systematic and extensive profiling or other automated processing of individual's personal aspects which are used for decisions which produce legal or similarly significant effects. The ICO considers that carrying out a DPIA may help organisations mitigate some of the challenges (mentioned above) around explaining AI.

What is personal data?

Personal data is defined in the GDPR as any information relating to an identified or identifiable natural person. An identifiable person is one who can be identified directly or indirectly, in particular by reference to an identifier such as a name, ID number or online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person.

Personal data is often processed when an AI model is being trained and operated. AI can also determine whether information falls within the definition of personal data, as the ability of AI to recognise patterns in data, or link data sets, can potentially enable data that would not normally be considered personal data to become "identifiable".

What Does the Project ExplAIn Guidance Cover?

The guidance is set out in three parts:

- Part 1 covers the basics of explaining AI and is an introductory section aimed at all stakeholders within an organisation.
- Part 2 looks at explaining AI in practice. It is aimed at technical teams but may also be of interest to compliance teams and data protection officers.
- Part 3 examines what explaining AI means for organisations. This is aimed at senior executives in an organisation and outlines the different roles that should be involved in providing an explanation to the relevant individuals.

It also contains a number of checklists to help organisations apply the guidance.

Part 1: Explanation types and principles

Part 1 of the guidance explains some of the basic terminology, GDPR provisions and risks associated with AI explainability. It also lists a set of AI principles which should be applied when explaining AI and a number of different ways in which AI decisions can be explained (explanation types).

The AI principles

The following four principles underpin how organisations should explain AI-assisted decisions to individuals and should be used together with the explanation types listed below them:

1. Be transparent – this is an extension of the transparency aspect of the lawfulness, fairness and transparency principle in the GDPR (see above). It is about making the use of AI for decision making obvious, and explaining the decisions you make to individuals in an appropriate way and at an appropriate time.
2. Be accountable – again, this is linked to the GDPR principle of the same name. GDPR accountability means: (i) taking responsibility for complying with the other data protection principles and demonstrating that compliance; and (ii) implementing appropriate technical and organisational measures and data protection by design and default. In an AI context, this means ensuring appropriate oversight of AI decision systems and being answerable for the decisions made (within the organisation, but also to regulators and relevant individuals). Organisations must, for example, identify those within their organisation who manage

and oversee the "explainability" requirements of an AI decisions system and assign responsibility for this. It also means showing that you have considered how to design and deploy explainable AI (and can justify this), have provided explanations to individuals and have a "capable human point of contact" to manage queries.

3. Consider context – the importance of context was one of the key findings of the Project ExplAIn research, as set out in its interim report released in June 2019.[3] The guidance lists five key contextual factors which affect why people want explanations of AI-assisted decisions and how explanations should be delivered (for example, which to prioritise). These are the: (i) domain (setting or sector) in which you operate and deploy the AI; (ii) impact or effect of the decision; (iii) data used; (iv) urgency of the decision; and (v) audience to whom it is being presented.

4. Reflect on impacts – many decisions made by AI will previously have been made by humans. The guidance describes AI as "increasingly serving as trustees of human decision-making" but notes that "individuals cannot hold these systems directly accountable for the consequences of their outcomes and behaviours". The principle of reflecting on impacts helps organisations to explain to individuals that the AI will not harm their wellbeing, which involves considering questions about the ethical purposes and objectives of the AI project. This aligns with the focus at UK and EU level on the ethical deployment of AI, although arguably stretches beyond the ICO's remit of data protection compliance.

Explanation types

As we have seen, context is a key aspect of explaining decisions involving AI. Several factors about the decision, the individual involved, type of data and the setting will all affect what information an individual would find useful or expect to receive as part of an explanation. The guidance therefore recognises that different types of explanation are required. It sets out six different types, which can be combined into an explanation in different ways, depending on the particular decision in question and the intended audience. They are:

1. Rationale explanation – the "why" of the decision which helps people understand the reasons that led to a decision or outcome. These should be delivered in a non-technical, accessible way. Part 2 of the guidance contains detailed information for technical teams on how to do this in practice.

2. Responsibility explanation – this focuses more on who is involved in the development, management and implementation of an AI system and who to contact for a human review of that decision.

3. Data explanation – what data has been used in a particular decision, and how. For example, what data was used to train and test the AI model and how it was used.

4. Fairness explanation – what steps have been taken (and will continue to be taken) in the design and implementation of the AI systems to ensure decisions are generally fair and unbiased. This also gives people an understanding of whether or not they have been treated equitably.

5. Safety and performance explanation – what steps have been taken across the design and implementation of an AI system to maximise the security, robustness, accuracy and reliability of its decisions and behaviours.

6. Impact explanation – what impact will the use of an AI system and its decisions have on an individual (and what broader societal effects may it have).

These are not intended to be an exhaustive list, but rather to identify what the ICO and the Turing consider to be the key types of explanations people will need. Each of these explanation types can be further subcategorised into "process" or "outcome" based explanations; the guidance discusses, for each explanation type, what information the process and outcome based explanations provide:

- Process-based explanations of AI systems are about demonstrating that you have followed good governance processes throughout the design and use.

- Outcome-based explanations of AI systems are about clarifying the results of a specific decision (i.e. explaining the reasoning behind an algorithmically-generated outcome in understandable language).

Part 2: Explaining in practice

As well as providing explanation types and principles to follow, the guidance provides some practical assistance on how to apply these. Part 2 is aimed primarily at technical teams, although the content is accessible and useful (albeit a long read) for compliance and risk advisors. It shows you how to:

- select the appropriate explanation for your sector and use case;

- choose an appropriately explainable model, which includes looking at some of the issues which arise with black box models; and

[3] https://ico.org.uk/about-the-ico/research-and-reports/project-explain-interim-report/

- use certain tools to extract explanations from less interpretable models.

It also sets out six tasks that organisations can undertake, which aim to provide a systematic approach to developing AI models with explainability in mind and selecting, extracting and delivering explanations regarding AI decisions. These are:

Task 1: Prioritise – get to know the different explanation types and select priority explanations by considering the domain, use case and impact on the individual. This will often involve prioritising the rationale and responsibility explanations, although all relevant explanations should be made available to the relevant individuals.

Task 2: Collect and pre-process your data in an explanation-aware manner (for example by using the PROV data model). How you collect and pre-process the data you use in your AI model will impact the quality of explanation you can provide.

Task 3: Build your system to ensure you are able to extract relevant information for a range of explanation types. This requires an understanding of the AI. Ensure that you have selected an AI model/system with an appropriate level of interpretability for your use case and for the impact it will have on the decision recipient. If you use a "black box" model, make sure you use supplementary explanation techniques which accurately represent the system's behaviour (see box "Black box issues and hybrid methods" below).

Task 4: Translate the rationale of the AI system's results into usable and easily understandable reasons. There must be a simple way to explain the model's statistical results to an individual. Where a decision is fully automated, the use of software may be needed to do this.

Task 5: Prepare implementers to deploy the AI systems/models – when human decision makers are involved in an AI-assisted outcome, they must be trained to use the model's results responsibly and fairly.

Task 6: Consider how to build and present your explanation – gather together and review the information gained when implementing tasks 1-4 and determine how this provides an evidence base for process or outcome-based explanations. Considering context should help you decide how to deliver appropriate information to an individual (what sort of, and how much, information to give and when.) A layered approach may avoid information overload, providing individuals with the information you have prioritised first while still making additional information available.

Black box issues and hybrid methods

Black box models

The black box effect of some AI models or systems has traditionally been seen as a barrier to explainability. The guidance defines a black box model as "an AI system whose inner workings and rationale are opaque or inaccessible to human understanding".

It may not always be possible to avoid black box models. For example, the most effective machine learning approaches will likely be opaque (for example, when recognising speech) as the feature spaces of these types of AI systems grow exponentially. However, such models should only be used if the potential impacts and risks have been thoroughly considered in advance, and it has been determined that the use case and organisational capacities/resources support the responsible design and implementation of these systems. In addition, appropriate supplementary interpretability tools should be used which provide a "domain-appropriate level of explainability" that is "reasonably sufficient to mitigate its potential risks and… a solid basis for providing affected decision recipients with meaningful information about the rationale of any outcome."

Hybrid methods – use of challenger models

The Project ExplAIn research found that, while some organisations in highly regulated sectors like banking and insurance are using interpretable models in their customer-facing AI decision-support applications, they are starting to use more opaque "challenger" models in parallel. Provided this is done in a transparent and responsible manner (and is documented), it can provide useful insights and comparisons. However, if the insights from the challenger model's processing are incorporated into the actual decision making, then they must be treated as core and held to the same explainability standards as the main model.

Organisations should keep a record of any deliberations that go into their selection of a black box or challenger model.

Part 3: What explaining AI means for your organisation

The final section of the guidance focuses on what this all means in practice for an organisation. It is aimed primarily at senior executives and looks at the various roles, policies, procedures and documentation that can be put in place to ensure an organisation is prepared to provide meaningful explanations to its customers and other individuals. It is also of use to compliance teams and risk advisors as it lists, albeit at a high level, what should be covered in the organisation's relevant policies and procedures, what documentation is legally required under the GDPR and what documentation can help the organisation demonstrate the explainability of its AI systems.

The first action point for organisations is to identify everyone involved in the decision-making pipeline and where they are responsible for providing an explanation of the AI system. In terms of the role of senior management, the guidance confirms that this is the team with overall responsibility for ensuring the AI system used by their organisation (whether developed in-house or procured) is appropriately explainable to the affected individuals. The guidance suggests that compliance teams, including DPOs and senior management, should expect assurances from the AI system's product manager (who, amongst other things, defines the product's requirements) that the system provides the appropriate levels of explanation to individuals. These assurances should give them a high level understanding of the system and types of explanations it produces.

Where AI is procured, organisations are still primarily responsible for ensuring the AI system is capable of producing explanations even where it is brought from a third party. Where off-the-shelf products are procured which do not contain inherent explainability, the organisation may need to use another model in parallel.

Comment

Project ExplAIn provides some useful guidance for organisations looking to increase the transparency of their AI use. It is, however, not the only source of guidance in this area. In the UK, the ICO also discusses AI transparency in its AI auditing framework and general big data and AI guidance[5]. The Turing is also working on a similar project with the UK's financial regulator and the UK's Centre for Data Ethics and Innovation ("CDEI") and the European Commission are also considering this point. For example, themes of transparency and explainability came through in responses to the CDEI's review into bias in algorithmic decision-making, and the Ethics Guidelines for Trustworthy AI published by the Commission's High-Level Expert Group on AI include transparency as one of seven key requirements that AI systems should meet. They state: "AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations."[6]

[4] https://www.gov.uk/government/publications/interim-reports-from-the-centre-for-data-ethics-and-innovation/interim-report-review-into-bias-in-algorithmic-decision-making

[5] FCA and Turing collaboration on AI transparency - https://www.fca.org.uk/insight/ai-transparency-financial-services-why-what-who-and-when

[6] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

This article was written by Rob Sumroy and Natalie Donovan. Rob is co-head of Slaughter and May's Emerging Tech Group and Natalie is a lawyer (PSL) in the team. The article is based on an article that first appeared in the International Comparative Legal Guide - Fintech 2020.

Rob Sumroy
T +44 (0)20 7090 4032
E rob.sumroy@slaughterandmay.com

Natalie Donovan
T +44 (0)20 7090 4058
E natalie.donovan@slaughterandmay.com