# Project ExplAIn

## Explaining decisions made with AI

Project status
**Finished**

Related programmes

**Public Policy**

# Introduction

AI and machine learning technologies are helping people do remarkable things. From assisting doctors in the early detection of diseases and supporting scientists who are wrestling with climate change to bringing together diverse groups from around the globe through real-time speech-to-speech translation, AI systems are enabling humans to successfully confront an ever-widening range of societal challenges.

This progress has, however, brought with it a new set of difficulties. Many machine learning applications, such as those in natural language processing and computer vision, complete their assigned tasks by identifying subtle patterns in large datasets. These systems accomplish this by linking together many hundreds, thousands—or sometimes even millions—of data points at a time. Humans don't think this way and because of this have difficulty understanding and explaining how these sorts of AI systems reach their results.

This gap in AI explainability becomes crucial when the outcomes of AI-assisted decisions have a significant impact on affected individuals and their communities. If an AI system is opaque then there is no way to ensure that its data processing is robust, reliable, and safe. Similarly, in cases where social or demographic data are being used as inputs in AI decision-support systems—for instance, in domains such as criminal justice, social care, or job recruitment—the employment of 'black box' models leaves designers and deployers no way to properly safeguard against possibilities of lurking biases that may produce inequitable or discriminatory results.

To respond to these challenges and gaps in AI explainability, the best practice document, ***Explaining Decisions Made with AI <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/>*** . Since its publication, the guidance has served as an essential reference point for public and private sector organisations around the world who are trying to navigate the complicated terrain of AI explainability.

# Project aims

Increasingly, organisations are using AI to help them make decisions. Where they are processing personal data to do this, they have to comply with certain parts of the General Data Protection Regulation. Moreover, where their AI-assisted decisions raise possibilities of discrimination against protected characteristics such as age, disability or race, organisations must comply with the 2010 Equality Act.

But beyond this, an organisation's capacity to explain its AI-assisted decisions to those affected by them builds trust among the public. It also improves the transparency and accountability of internal governance processes by having an informed workforce that can then maintain oversight of what these systems do and why. Society benefits too, as the priority of designing explainable AI models can improve their reliability, safety, and robustness. It can also help surface the existence of potential issues of bias within these AI systems and in the data they use, which can then be addressed and possibly mitigated.

Project ExplAIn is a collaboration between the Information Commissioner's Office and The Alan Turing Institute to provide guidance to organisations on the key principles, concepts, and tools that can help provide explanations in practice. In the second phase of the project Manchester Metropolitan University was a key collaborator and helped to produce workbooks for organisations to help communicate this guidance.

## Where did this come from?

The project underpinning this work, Project ExplAIn, came about as a result of Professor Dame Wendy Hall and Jérôme Pesenti's 2017 **independent review on growing the AI industry in the UK <https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk>** . This was followed in 2018 by the **Government's AI Sector Deal <https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal>** , which tasked the ICO and the Turing to "…work together to develop guidance to assist in explaining AI decisions."

In February 2019, two five-day-long citizens' juries on AI explanation were staged in Coventry and Manchester. These were designed to elicit public preferences about what people expect from explanations of AI-assisted decisions. The juries used a deliberative format with the assistance of expert witnesses, who provided jurors with background information about the technical, legal and ethical dimensions of AI explainability. The juries were followed by three roundtables, where the feedback from the citizens were presented to and then discussed by a range of academic and industry stakeholders, from data scientists and researchers to data protection officers, C-suite executives and lawyers. The results of these public engagement activities as well as extensive desk research have provided the basis for the guidance.

## How will it help?

Wherever organisations use personal data to make AI-assisted decisions, they should be able to explain those decisions to the people affected by them. The guidance we have produced provides an accessible overview of the key principles, concepts and tools that can help organisations provide explanations in practice.

## What's in the guidance?

At the heart of the guidance is a series of related questions: What makes for a good explanation of decisions supported by AI systems? How can such explanations be reliably extracted and made understandable to a non-technical audience? How should organisations go about providing meaningful explanations of the AI-supported decisions they make? What do the people affected by these decisions deserve, desire and need to know?

The main focus of the guidance is the need to tailor explanations to the context in which AI systems are used for decision-making. This vital contextual aspect includes the domain or sector in which an organisation operates, and the individual circumstances of the person receiving the decision.

The guidance also stresses a principles-based approach to the governance of AI explanations. We present four principles of explainability that provide ethical underpinnings for the guidance and that steer the practical recommendations contained in it:

- **Be transparent**: Be open and candid regarding how and where your organisation uses AI decision-support systems and provide meaningful explanations of their results.
- **Be accountable**: Ensure appropriate oversight of AI decision-support systems and be answerable to others in your organisation, to external bodies, and to the individuals affected by AI-assisted decisions.
- **Consider context**: Choose AI models and explanations that are appropriate to the settings and potential impacts of their use-cases, and tailor governance processes to the structures and management processes of your organisation.
- **Reflect on impacts**: Weigh up the ethical purposes and objectives of your AI project at the initial stages of formulating the problem and defining the outcome, and think about how the system may affect the wellbeing of individuals and wider society.

Building off these principles, we identify a number of different explanation types, which cover various facets of an explanation, and will often be used in concert with each other:

- **Responsibility**: who is involved in the development and management of an AI system, and who to contact for a human review of a decision.
- **Rationale**: the reasons that led to a decision, delivered in an accessible way.
- **Fairness**: steps taken to ensure that AI decisions are generally unbiased and fair, and whether or not an individual has been treated equitably.
- **Safety and performance**: steps taken to maximise the accuracy, reliability, security and robustness of the decisions the AI system helps to make.
- **Impact**: the effect that the AI system has on an individual, and on wider society.

- **Data**: what data has been used in a particular decision, and what data has been used to train and test the AI model.

For organisations, the emphasis is on how to set up and govern the use of AI systems to be suitably transparent and accountable, and that they prioritise, where appropriate, using inherently explainable AI models before choosing less interpretable models, such as 'black box' systems. We outline the art of the possible in these considerations, to help the governance and technical teams in organisations think about how to extract explanations from their AI systems.

When delivering an explanation to the individual affected, there are a number of contextual factors that will inform what they should be told first, and what information to make available separately. We call this 'layering' explanations, which is designed to avoid information overload. These contextual factors are:

- **Domain**: the setting or sector in which the AI system is deployed to help make decisions about people. What people want to know in the health sector will be very different to the explanation they will want in the criminal justice domain.
- **Impact**: the effect an AI-assisted decision can have on an individual. Varying levels of severity and different types of impact can change what explanations people will find useful, and the purpose the explanation serves.
- **Data**: the data used to train and test an AI model, and the input data used for a particular decision. The type of data used can influence an individual's willingness to accept or contest an AI-assisted decision, and the actions they take as a result of it.
- **Urgency**: the importance of receiving, or acting upon, the outcome of a decision within a short timeframe.
- **Audience**: the individuals the explanation is being given to will influence what type(s) of explanation will be useful.

# Research outputs

The Alan Turing Institute and the Information Commissioner's Office collaborated to produce guidance for organisations; and The Alan Turing Institute subsequently collaborated with Manchester Metropolitan University to produce workbooks to communicate the guidance to organisations through case studies.

## *Explaining Decisions Made with AI* workbooks and workshops

At the beginning of 2021, the project team assembled two workbooks to help support the uptake of the guidance. The goal of the workbooks was to summarise

the main themes from *Explaining Decisions Made with AI* in a non-technical way. Additionally, each workbook served as the basis of a workshop exercise built around one of two use cases, created to help organisations and individuals gain a flavour of how to put the guidance into practice.

The workbooks were written to support the second phase of Project ExplAIn, centred on stakeholder outreach and practice-based evaluation. This included a series of engagement activities held in January 2021 to assess the usability, accessibility and clarity of the guidance, as well as the readiness levels of organisations to put explainable AI principles into practice. In partnership with Manchester Metropolitan University (MMU) and the ICO, two workshops – one with SMEs from advertising, AI development, finance, recruitment, health, education, fraud protection, media and insurance sectors, and a second with public sector organisations – were held virtually. The workshops engaged participants from a variety of different backgrounds, levels of seniority, and roles across the public and private sectors. We are extremely grateful to them for their energy, enthusiasm, and tremendous insight.

We hope that our workbooks will allow for more widespread use and dissemination of the guidance. The workbooks begin with a truncated form of the *Explaining Decisions Made with AI* guidance, presenting the four principles of AI explainability, the basics of an explanation-aware approach to AI innovation, and the practical tasks needed for the explanation-aware design, development and use of AI systems. They then provide some reflection questions, which are intended to be a launching pad for group discussion. The appendices of the workbooks are primarily focused on both the workshop setting and the case studies. Appendix A provides a structure for how to use the workbook in a workshop setting, including details on necessary resources, personnel, and recommended timelines. These recommendations are based on the workshops co-hosted with ICO and MMU in January 2021. Appendix B contains the case study, followed by appendix C which consists of a checklist for one or more of the explanation types to be used in tandem with the case study.

**Case studies found in the workbooks:**

- The '**AI-assisted recruitment tool <https://zenodo.org/record/4624711#.YUiUzy1Q1ao>** ' case study depicts a company considering the use of an AI-assisted recruitment tool to support HR personnel with future job vacancies. The tool uses a variety of personal and professional criteria to determine which candidates would be the best fit for the organisation. The case study provides detail on the data available to the organisation, model type considerations, and the roles and responsibilities within the organisation. Participants are asked to focus on responsibility explanation, data explanation, and fairness explanation by applying a checklist of tasks to the case study.

- The '**Machine learning for children's social care <https://zenodo.org/record/4624733#.YUiVAy1Q1ao>** ' case study describes an organisation contemplating the use of a machine learning algorithm within a children's social care setting. This algorithm would be used to identify children at risk. The case study gives detailed explanations of the variables available and the

exploratory data analysis that took place. It then illustrates what a logistic regression model trained on the data would look like and includes feature importance plots to help participants think through how each variable factors into an explanation of the model. Additionally, the case study presents a hypothetical scenario in which the model is applied to a specific family. Participants are asked to focus on rationale explanation by applying a checklist of tasks to the case study.

These workbooks would simply not exist without the commitment and keenness of all our collaborators and workshop participants, and we would like to thank them again for their involvement.

This infographic video produced by Fable Studios consists of an introduction to guidance on *Explaining Decisions Made with AI*. The video provides basic information about the importance of explainable AI. It includes an introduction to the four principles of AI explainability and a description of the six explanation types which are meant to assist organisations with delivering understandable explanations to relevant stakeholders. The purpose of the video is to provide an accessible entry point to the guidance and to direct towards the **complete version of the guidance <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-ai/>** , to learn more about how to implement it in practice.

**Read guidance <https://ico.org.uk/for-organisations/guide-to-data-prot**

**ection/key-dp-themes/explaining-decisions-made-with-ai/>**

# Insights from Phase 2 of Project ExplAIn

The ***Explaining Decisions Made with AI <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/?q=explainability>*** guidance was released in May 2020. Since its publication, the guidance has served as an essential reference point for public and private sector organisations around the world who are trying to navigate the complicated terrain of AI explainability.

In mid-2022, to assess the guidance's usability and uptake, our team hosted a series of workshops and a mini public - a diverse group of people brought together to deliberate on a topic and inform decision-making — on the content of the guidance — hosted in Bristol in collaboration with **Traverse <https://traverse.ltd/>** , a former public engagement company.

The aim of the mini public was to gather views and opinions on the guidance from a diverse and inclusive sample of people living in the UK. 40 participants were recruited to partake in our Bristol mini public, with 31 participants attending all four days. We prioritised diversity from the outset of the engagement design to ensure that an optimally inclusive range of voices were present in the room. In the final sample of 31 participants, 19 self-identified as female, 13 self-identified as members of an ethnic minority group, 7 were providers of unpaid care, and 17 had long-term health problems or disabilities. We believe the diversity of this set of participants contributed greatly to a rich dialogue surrounding possible improvements to the guidance itself.

While there was a vast array of interesting findings, key themes from the mini public include the following:

- Overarching concerns surrounding the fairness of AI systems
- Differing expectations for an explanation when comparing a decision made by an AI system versus a human decision-maker
- No complete consensus on the desired delivery method of the explanation
- Precedence given to considerations of responsibility and the importance of human involvement
- Widespread acknowledgement of the various benefits of an explanation

More about these themes can be read on our **associated blog post <https://www.turing.ac.uk/news/project-explain/insights-from-phase-two>** .

**Find out more <https://www.turing.ac.uk/news/project-explain/insights-from-phase-two>**

*This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/W006022/1, particularly the “Criminal Justice System” theme within that grant & The Alan Turing Institute.*

# Organisers

**Professor David Leslie**
Director of Ethics and Responsible Innovation Research

## Collaborators

Manchester Metropolitan University

ico.
Information Commissioner's Office

## Researchers and collaborators

**Morgan Briggs**
Research Associate for Data Science and Ethics

## Related content

## Launching guidance from Project ExplAIn

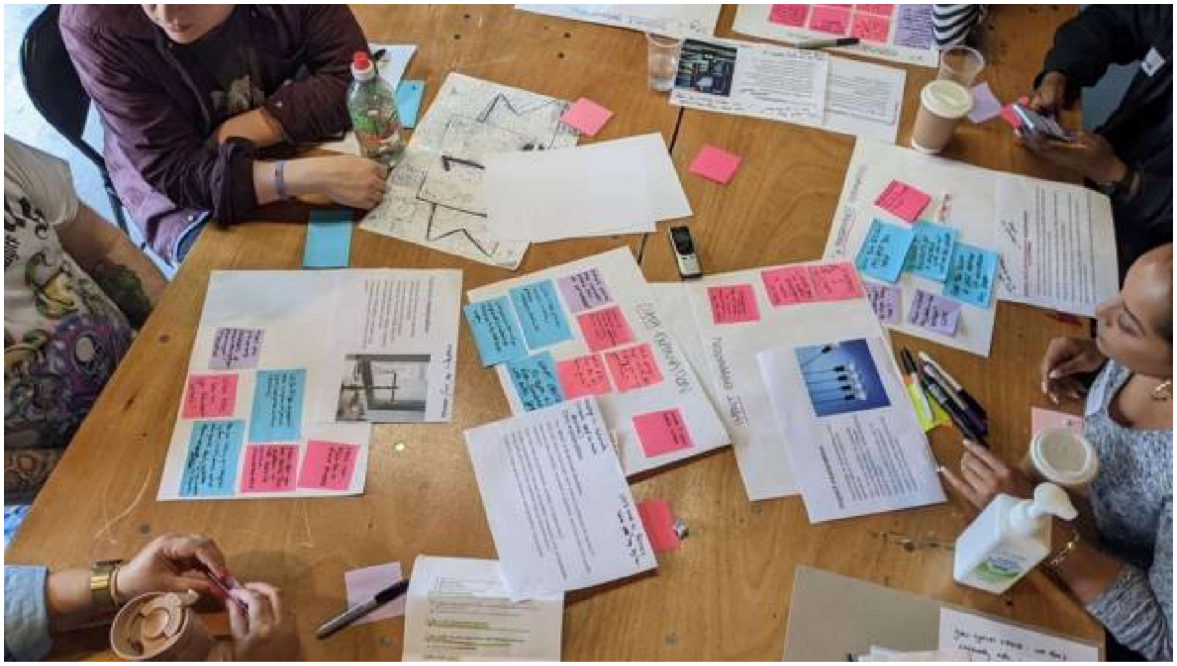
**At the cutting edge of practice-centred guidance on explainable AI**

Monday 02 Dec 2019

## Project ExplAIn enters its next phase

**The Turing and the Information Commissioner's Office continue to work on their first-of-its-kind guidance on AI explainability**

Tuesday 16 Nov 2021



**Insights from phase two of Project ExplAIn**

# Funders

The Alan Turing Institute, a charity incorporated and registered in England and Wales with company number 09512457 and charity number 1162533 whose registered office is at British Library, 96 Euston Road, London, England, NW1 2DB, United Kingdom.

For website-related enquiries email **website@turing.ac.uk**

## Awards