

AI Explainability in Practice



Facilitator Workbook

Annotated to support facilitators in delivering the accompanying activities.

The
Alan Turing
Institute

Acknowledgements

This workbook was written by David Leslie, Cami Rincón, Morgan Briggs, Antonella Maia Perini, Smera Jayadeva, Ann Borda, SJ Bennett, Christopher Burr, Mhairi Aitken, Sabeegah Mahomed, Janis Wong, Madeleine Waller, and Claudia Fischer.

The creation of this workbook would not have been possible without the support and efforts of various partners and collaborators. As ever, all members of our brilliant team of researchers in the Ethics Theme of the Public Policy Programme at The Alan Turing Institute have been crucial and inimitable supports of this project from its inception several years ago, as have our Public Policy Programme Co-Directors, Helen Margetts and Cosmina Dorobantu. We are deeply thankful to Conor Rigby, who led the design of this workbook and provided extraordinary feedback across its iterations. We also want to acknowledge Johnny Lighthands, who created various illustrations for this document, and Alex Krook and John Gilbert, whose input and insights helped get the workbook over the finish line. Special thanks must be given to the Digital Office for Scottish Local Government, staff and residents of Camden Council, Justin Green (Northumbria Healthcare NHS Foundation Trust) and Postgraduate doctors in training (PGDiT) at NHS England Education Northeast & North Cumbria, and Anna Knack, Ardi Janjeva, Megan Hughes, Rosamunde Powell, and Samuel Stockwell (The Alan Turing Institute) for helping us test the activities and review the content included in this workbook.

This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/W006022/1, particularly the Public Policy Programme theme within that grant & The Alan Turing Institute; Towards Turing 2.0 under the EPSRC Grant EP/W037211/1 & The Alan Turing Institute; and the Ecosystem Leadership Award under the EPSRC Grant EP/X03870X/1 & The Alan Turing Institute

Cite this work as: Leslie, D., Rincón, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, SJ. Burr, C., Aitken, M., Mahomed, S., Wong, J., Waller, M., and Fischer, C. (2024). *AI Explainability in Practice*. The Alan Turing Institute.

Contents

About the Workbook Series



- 4 Who We Are
- 4 [Origins of the Workbook Series](#)
- 5 About the Workbooks
- 6 Intended Audience
- 7 [Introduction to This Workbook](#)

Key Concepts



10 Part One: Introduction to AI Explainability

- 11 [Transparency](#)
- 15 [Process-Based and Outcome-Based Explanations](#)
- 16 [Maxims of AI Explainability](#)
 - 16 [Maxim 1: Be Transparent](#)
 - 18 [Maxim 2: Be Accountable](#)
 - 19 [Maxim 3: Consider Context](#)
 - 22 [Maxim 4: Reflect on Impacts](#)
- 24 [High-Level Considerations for Building Appropriately Explainable AI Systems](#)
- 28 [Main Types of Explanation](#)
 - 29 [Rationale Explanation](#)
 - 31 [Responsibility Explanation](#)
 - 33 [Data Explanation](#)
 - 35 [Fairness Explanation](#)
 - 39 [Safety Explanation](#)
 - 42 [Impact Explanation](#)

44 Part Two: Putting the Principle of Explainability Into Practice

- 45 [Tasks for Explainability Assurance Management](#)
- 54 [Explainability Assurance Management Template](#)

Activities



- 68 [Activities Overview](#)
- 70 [Interactive Case Study: AI in Children's Social Care](#)
 - 72 [Details About the AI System Under Consideration](#)
 - 73 [Details About the Database](#)
 - 76 [Hypothetical Scenario: The Smith Family](#)

80 Content Review and Discussion

82 Information Gathering

84 Evaluating Explanations

Further Readings



- 86 [Appendix A: Algorithmic Techniques](#)
- 92 [Appendix B: Supplementary Models](#)
- 107 [Endnotes](#)

About the AI Ethics and Governance in Practice Workbook Series

Who We Are

The Public Policy Programme at The Alan Turing Institute was set up in May 2018 with the aim of developing research, tools, and techniques that help governments innovate with data-intensive technologies and improve the quality of people's lives. We work alongside policymakers to explore how data science and artificial intelligence can inform public policy and improve the provision of public services. We believe that governments can reap the benefits of these technologies only if they make considerations of ethics and safety a first priority.

Origins of the Workbook Series

In 2019, The Alan Turing Institute's Public Policy Programme, in collaboration with the UK's Office for Artificial Intelligence and the Government Digital Service, published the [UK Government's official Public Sector Guidance on AI Ethics and Safety](#). This document provides end-to-end guidance on how to apply principles of AI ethics and safety to the design, development, and implementation of algorithmic systems in the public sector. It provides a governance framework designed to assist AI project teams in ensuring that the AI technologies they build, procure, or use are ethical, safe, and responsible.

In 2021, the UK's National AI Strategy recommended as a 'key action' the update and expansion of this original guidance. From 2021 to 2023, with the support of funding from the Office for AI and the Engineering and Physical Sciences Research Council as well as with the assistance of several public sector bodies, we undertook this updating and expansion. The result is the AI Ethics and Governance in Practice Programme, a bespoke series of eight workbooks and a [digital platform](#) designed to equip the public sector with tools, training, and support for adopting what we call a Process-Based Governance (PBG) Framework to carry out projects in line with state-of-the-art practices in responsible and trustworthy AI innovation.

About the Workbooks

The AI Ethics and Governance in Practice Programme curriculum is composed of a series of eight workbooks. Each of the workbooks in the series covers how to implement a key component of the PBG Framework. These include Sustainability, Safety, Accountability, Fairness, Explainability, and Data Stewardship. Each of the workbooks also focuses on a specific domain, so that case studies can be used to promote ethical reflection and animate the Key Concepts.

Programme Curriculum: AI Ethics and Governance in Practice Workbook Series



1 AI Ethics and Governance in Practice: An Introduction
Multiple Domains



5 Responsible Data Stewardship in Practice
AI in Policing and Criminal Justice



2 AI Sustainability in Practice Part One
AI in Urban Planning



6 AI Safety in Practice
AI in Transport



3 AI Sustainability in Practice Part Two
AI in Urban Planning



7 AI Explainability in Practice
AI in Social Care



4 AI Fairness in Practice
AI in Healthcare



8 AI Accountability in Practice
AI in Education



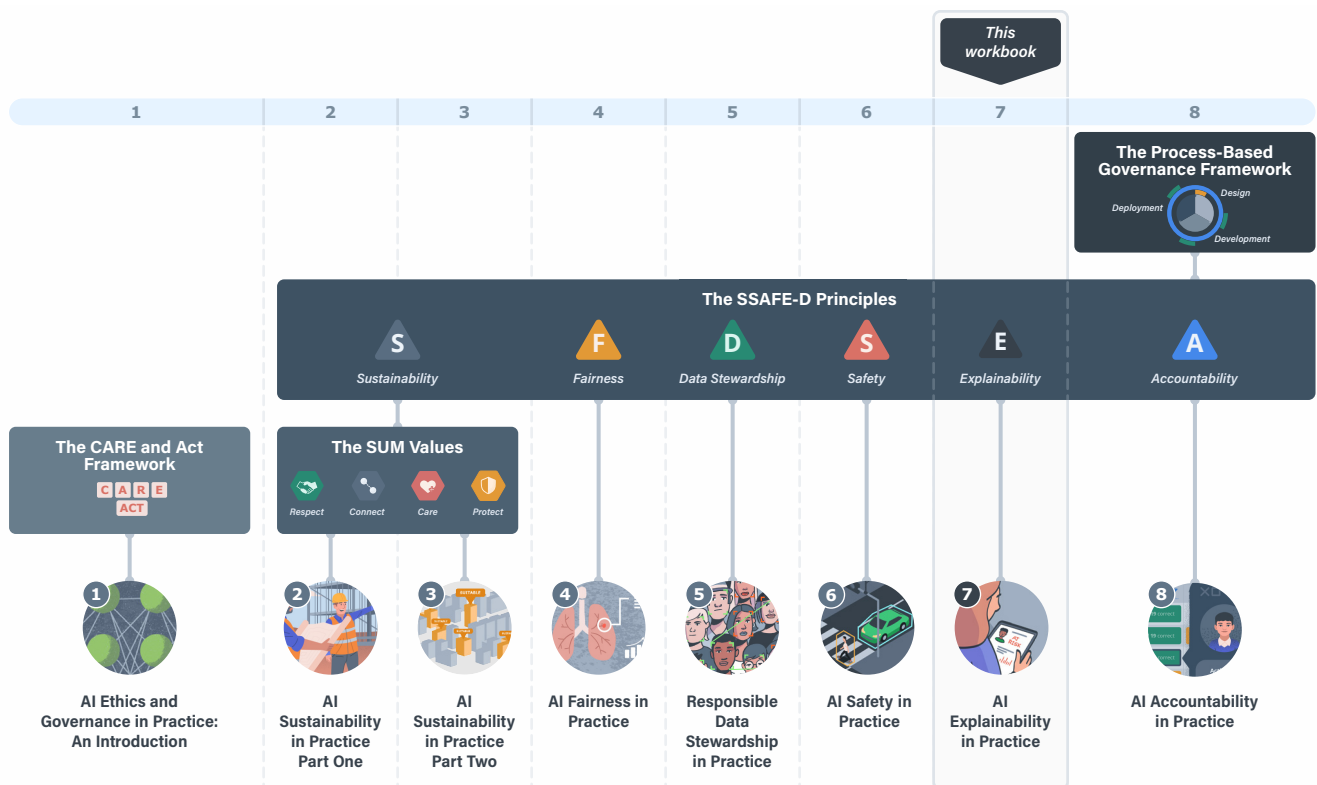
Explore the full curriculum and additional resources on the AI Ethics and Governance in Practice Platform at aiethics.turing.ac.uk.

Taken together, the workbooks are intended to provide public sector bodies with the skills required for putting AI ethics and governance principles into practice through the full implementation of the guidance. To this end, they contain activities with instructions for either facilitating or participating in capacity-building workshops.

Please note, these workbooks are living documents that will evolve and improve with input from users, affected stakeholders, and interested parties. We value your participation. Please share feedback with us at aiethics@turing.ac.uk.

Programme Roadmap

The graphic below visualises this workbook in context alongside key frameworks, values and principles discussed within this programme. For more information on how these elements build upon one another, refer to [AI Ethics and Governance in Practice: An Introduction](#).



Intended Audience

The workbooks are primarily aimed at civil servants engaging in the AI Ethics and Governance in Practice Programme — whether as AI Ethics Champions delivering the curriculum within their organisations by facilitating peer-learning workshops, or as participants completing the programmes by attending these workshops. Anyone interested in learning about AI ethics, however, can make use of the programme curriculum, the workbooks, and resources provided. These have been designed to serve as stand-alone, open access resources. Find out more at aiethics.turing.ac.uk.

There are two versions of each workbook:

- Facilitator Workbooks** (such as this document) are annotated with additional guidance and resources for preparing and facilitating training workshops.
- Participant Workbooks** are intended for workshop participants to engage with in preparation for, and during, workshops.

Introduction to This Workbook

The purpose of this workbook is to introduce participants to the principle of AI Explainability. Understanding how, why, and when explanations of AI-supported or -generated outcomes need to be provided, and what impacted people's expectations are about what these explanations should include, is crucial to fostering responsible and ethical practices within your AI projects. To guide you through this process, we will address essential questions: What do we need to explain? And who do we need to explain this to? This workbook offers practical insights and tools to facilitate your exploration of AI Explainability. By providing actionable approaches, we aim to equip you and your team with the means to identify when and how to employ various types of explanations effectively. This workbook is divided into two sections, Key Concepts and Activities:

Key Concepts Section

This section provides content for workshop participants and facilitators to engage with prior to attending each workshop. It first provides definitions of key terms, introduces the maxims of AI Explainability and considerations for building appropriately explainable AI systems, and gives an overview of the main types of explanations. The section then delves into practical tasks and tools to ensure AI Explainability. Topics discussed include:

Part One: Introduction to AI Explainability



Introduction to AI Explainability



Process-Based and Outcome-Based Explanations



Maxims of AI Explainability



Considerations for Building Appropriately Explainable AI Systems



Types of Explanation

Part Two: Putting the Principle of Explainability Into Practice



Tasks for Explainability Assurance Management



Explainability Assurance Management Template

Activities Section

This section contains instructions for group-based activities (each corresponding to a section in the Key Concepts). These activities are intended to increase understanding of Key Concepts by using them.

Case studies within the AI Ethics and Governance in Practice workbook series are grounded in public sector use cases, but do not reference specific AI projects.



Content Review and Discussion

Go over Key Concepts and review the case study for this workshop.



Information Gathering

Practise gathering relevant information for building explanations of AI systems.



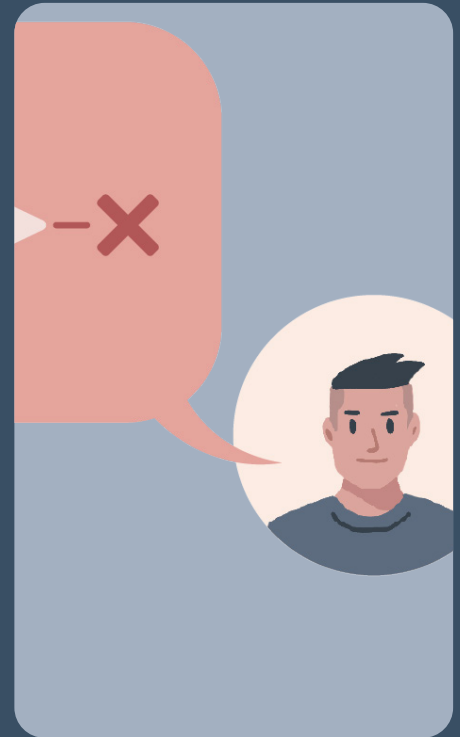
Evaluating Explanations

Practise evaluating the extent to which AI explanations meet their purpose and align with the Maxims of AI Explainability.

Note for Facilitators

Additionally, you will find facilitator instructions (and where appropriate, considerations) required for facilitating activities and delivering capacity-building workshops.

Key Concepts



10 Part One: Introduction to AI Explainability

11 Transparency

15 Process-Based and Outcome-Based Explanations

16 Maxims of AI Explainability

16 Maxim 1: Be Transparent

18 Maxim 2: Be Accountable

19 Maxim 3: Consider Context

22 Maxim 4: Reflect on Impacts

24 High-Level Considerations for Building Appropriately Explainable AI Systems

28 Main Types of Explanation

29 Rationale Explanation

31 Responsibility Explanation

33 Data Explanation

35 Fairness Explanation

39 Safety Explanation

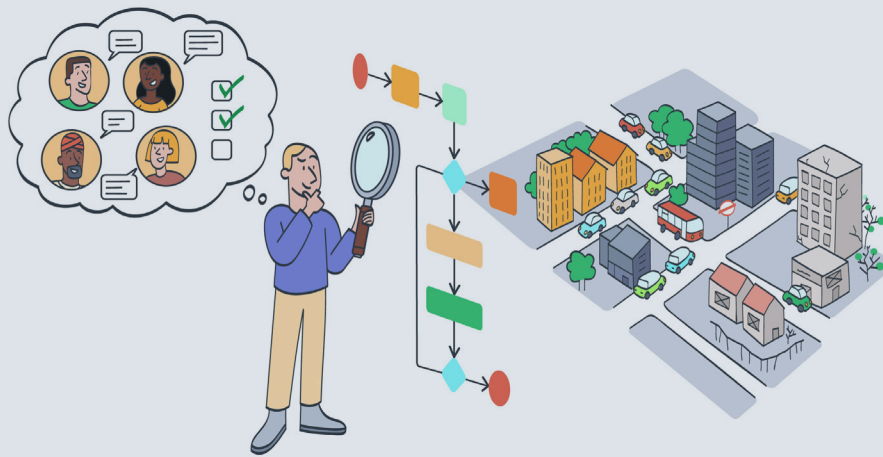
42 Impact Explanation

44 Part Two: Putting the Principle of Explainability Into Practice

45 Tasks for Explainability Assurance Management

54 Explainability Assurance Management Template

Part One: Introduction to AI Explainability



KEY CONCEPT

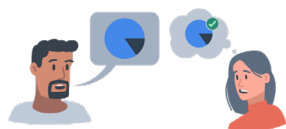
AI Explainability

In this workbook, we define AI Explainability as the degree to which a system or a set of governance practices and tools support a person's ability: (1) to explain and communicate the rationale underlying the behaviour of the system or (2) to demonstrate and convey that the processes behind its design, development, and deployment have been undertaken in ways that ensure its sustainability, safety, fairness, and accountability across its particular contexts of use and application.

The workbook operationalises concepts from [Explaining decisions made with AI](#), a co-badged guidance by the Information Commissioner's Office (ICO) and The Alan Turing Institute. This guidance aims to give organisations practical advice to help explain the processes, services and decisions delivered or assisted by AI, to the individuals affected by them.

Sections of the workbook also draw from the [Responsible Research and Innovation in Data Science and AI Skills Track of Turing Commons](#).^[1]

The ability to explain and justify AI project processes and AI-supported outcomes is central to ensuring AI projects that are sustainable, fair, safe, accountable, and maintain data quality, integrity, and protection.



Explainability entails an emphasis on communicability, and 'clear' and 'accessible' explanations.



The depth, breadth, and content of explanations of AI-supported decisions (and what makes such explanations appropriate, clear, and accessible) **will vary depending on the sociocultural context in which they are being delivered and the audience to whom they are being offered.**



Explainability aims to give reasons for:

1. the outcomes of the algorithmic model (which may be used for automated decisions or as inputs for human decision-making); and
2. the processes by which a model and its encompassing system/interface are designed, developed, deployed, and/or deprovisioned.

Transparency

A neighbouring concept to AI Explainability is that of AI Transparency. Transparency holds multiple meanings dependent on the context and discipline it is being used within. The common dictionary understanding of transparency defines it as either:

1. the quality an object has when one can see clearly through it, or
2. the quality of a situation or process that can be clearly justified and explained because it is open to inspection and free from secrets.^[2]

The principle of AI Transparency encompasses both of these meanings:

1. Interpretability of an AI system or the ability to know how and why a model performed the way it did in a specific context and therefore to understand the rationale behind its decision or behaviour. This sort of transparency is often referred to by way of the metaphor of 'opening the black box' of AI. It involves content clarification and intelligibility.^[3]
2. Transparent AI asks that the designers and developers of AI systems demonstrate that their processes and decision-making, in addition to system models and outputs, are sustainable, safe, fair, and driven by responsibly managed data.^[4]

Developing an explanation requires a certain degree of transparency. Transparency of outcomes and processes, for instance, through documentation about how an AI system

was designed, developed, and deployed, can 'help explain and justify the actions and decisions undertaken throughout a project's lifecycle'.^[5]

In this workbook, rather than zeroing in on AI Transparency, we focus primarily on the more practice-centred concept of AI Explainability, because we are concerned with providing guidance on how to operationalise the transparency of both of AI-supported outcomes and of the processes behind the design, development, and deployment of AI systems. The principle of AI Explainability will be used to refer directly to practices of outcome-based and process-based explanation, as we will explain in the next section.

What is the UK's national Algorithmic Transparency Recording Standard?

The Algorithmic Transparency Recording Standard (ATRS) was co-developed by the Central Digital and Data Office and the Responsible Technology Adoption Unit from 2021 to 2023 and updated in 2024. It was produced through a co-design process with UK citizens, with support from other stakeholders across academia and civil society. The ATRS has been trialled across the public sector, including with central and local governments, police forces and regulators.

The ATRS is a framework for capturing information about algorithmic tools, including AI systems. It is designed to help public sector bodies openly publish information about the algorithmic tools they use in decision-making processes that affect members of the public.

Transparency is a key component of the safe, fair, just, and responsible use of algorithmic tools. However, many public sector organisations are unsure how to be transparent when using algorithms to deliver services. The ATRS provides a clear and accessible means of communication, and contributes to a

more effective and accountable approach to public service delivery.

Since its initial publication, the ATRS has been tested, reviewed, and refined. With its 2024 update, it includes a new repository for ATRS records, with the ability to search for and filter published records, including by organisation, sector, geography and model capability, as we scale and have more records published.

For further information, visit these links.

- <https://www.gov.uk/government/publications/guidance-for-organisations-using-the-algorithmic-transparency-recording-standard/algorithmic-transparency-recording-standard-guidance-for-public-sector-bodies>
- <https://rtau.blog.gov.uk/2021/06/21/engaging-with-the-public-about-algorithmic-transparency-in-the-public-sector/>
- <https://rtau.blog.gov.uk/2024/03/07/algorithmic-transparency-recording-standard-getting-ready-for-adoption-at-scale/>



Consideration: Trade-offs of Security and Explainability

In high-stakes contexts, such as national security, transparency around how an AI system works may create vulnerabilities that could be exploited by malicious actors. This situation creates an incentive for more secure AI systems that protect their algorithms and data. However, where high-stake decisions are made with incomplete information and rely on discretion and professional judgment, human operators still need to demonstrate the necessary and proportionate basis for the decision.^[6] In addition, this lack of transparency and ability to explain the AI system can be problematic, as it raises concerns about bias, fairness, and accountability and can lead to unintended consequences.^[7]

Balancing these two aspects is essential for building responsible AI systems. Project teams need to address considerations around the potential AI Safety risks, how they manage the information generated about those risks, and how these are shared or protected. They will also need to integrate transparency considerations into those decisions, and consider the extent to which explanations about the model, and the processes of the AI project, will be made available.^[8]



Considerations for Child-Centred AI

There are a host of risks AI applications create for children. Among many challenges, there is a long-term concern of the potential transformative effects of these technologies on the holistic development of children into socialised members of their communities. Furthermore, there are ongoing risks to the privacy of children and their families in increasingly data extractive, intrusive, and digitally networked social environments.

Throughout this workbook, we will flag specific considerations related to explainability that must take place when children are the impacted data subjects. These considerations are included as call-out-boxes and have been built on The Alan Turing Institute's ongoing research with United Nations Children's Fund (UNICEF), The Scottish AI Alliance, and the Children's Parliament, including their piloting of the [UNICEF Policy guidance on AI for children](#). The considerations discussed are underpinned by the [United Nations Convention on the Rights of the Child](#) signed by 196 countries which consists of 54 articles that outline how governments must work to meet children's needs and their full potential through the understanding that all children have basic fundamental rights.

The Age Appropriate Design Code^[9] ^[10]

The UK's Information Commissioner's Office (ICO) has developed a statutory code aimed at protecting children's online data. This code comprises 15 standards that online services, deemed 'likely to be accessed by children', must adhere to, in line with the Convention on the Rights of the Child (UNCRC) and the obligations outlined in the General Data Protection Regulation (GDPR) and the Privacy and Electronic Communications Regulations (PECR). Among its provisions, the code mandates that online services:

- provide a high level of default privacy settings;
- present their services in a language suitable for children;
- do not use nudge techniques;
- refrain from sharing children's data with third parties, disable geolocation services;
- provide tools to empower children in exercising their data rights; and
- prioritise choices that serve the best interests of the child.

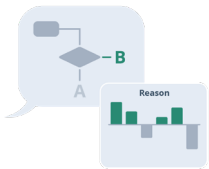
Policy Guidance on AI for Children^[11]

UNICEF published in 2021 policy guidance to promote human-centric AI through a child rights lens. The guidance recommends that developers, policymakers, and businesses should meet the following nine requirements:

1. Support children's development and wellbeing.
2. Ensure inclusion of and for children.
3. Prioritise fairness and non-discrimination for children.
4. Protect children's data and privacy.
5. Ensure safety for children.
6. Provide transparency, explainability, and accountability for children.
7. Empower governments and businesses with knowledge of AI and children's rights.
8. Prepare children for present and future developments in AI.
9. Create an enabling environment.

Process-Based and Outcome-Based Explanations

Explanations can be provided for outcomes of systems (the content and justification of that outcome) or for the process behind the systems (the project design, model development, and system deployment practices that lead to an algorithmically supported outcome). It is important that you provide explanations to impacted stakeholders that demonstrate how you and all others involved in the development of your system acted responsibly when choosing the processes behind its design and deployment, and make the reasoning behind the outcome of that decision clear.



Outcome-based explanations include the components and reasoning behind model outputs while delineating contextual and relational factors. These explanations should be made accessible to impacted stakeholders through plain, easily understandable, and everyday language.^[12]

- In human-in-the-loop systems (where a person periodically reviews decisions taken by AI), you should also explain to the affected stakeholders if, how, and why the AI-assisted human judgement was reached.
- In offering an explanation to affected stakeholders, you should be able to demonstrate that the specific decision or behaviour of your system is sustainable, safe, fair, and driven by data that has been responsibly managed.



Considerations for Child-Centred AI

When delivering an explanation of an algorithmic decision, it is critical that the specific needs and capabilities of children are considered. This includes training the implementers and users of the AI system to deliver accessible and understandable explanations to children. As such it is important to engage with children across the project lifecycle so that teams can better understand how best to craft explanations on the use of children's data and subsequent AI outputs.



Process-based explanations of AI systems demonstrate that you have followed good governance processes and best practices throughout your design and use.^[13] This entails demonstrating that considerations of sustainability, safety, fairness, and responsible data management were operative end-to-end in the project lifecycle.

- For example, if you are trying to explain the fairness and safety of a particular AI-assisted decision, one component of your explanation will involve establishing that you have taken adequate measures across the system's production and deployment to ensure that its outcome is fair and safe.

Considerations for Child-Centred AI

When considering AI systems that are developed using children's data, in order to explain that the system is fair and safe, one must first adhere to the [UK ICO's Age Appropriate Design Code](#). However, a process-based explanation in this setting expands well past legal compliance. How have adequate measures of reporting and oversight been put in place to ensure that children are not harmed in the process of the design, development, and deployment of AI technologies that use their data?

Maxims of AI Explainability

The following maxims provide a broad steer on what to think about when explaining AI/ML-assisted decisions to individuals.

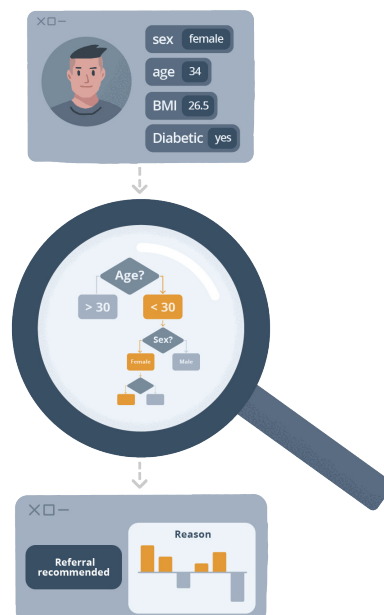
Maxim 1

Be Transparent

The maxim of being transparent is about making your use of AI/ML for decision-making obvious and appropriately explaining the decisions you make to individuals in a meaningful way. Transparency is addressed in Article 5(1) of the [UK General Data Protection Regulation \(GDPR\)](#), which says that personal data shall be:

“processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness, transparency').”^[14]

Project teams should satisfy all three aspects of the maxim which includes being honest about who you are alongside how and why you are using personal data.



Key Aspects of Being Transparent

- 1. Disclose AI use.** Proactively make people aware of a specific AI-enabled decision concerning them, in advance of making the decision. Be open and candid about:
 - your use of AI-enabled decisions;
 - when you use them; and
 - why you choose to do this.
- 2. Meaningfully explain decisions.** Provide the stakeholders with a coherent explanation which is:
 - truthful and meaningful;
 - written or presented appropriately; and
 - delivered at the right time.



Considerations for Child-Centred AI

When considering children's rights as they relate to AI systems, several child-centric guidance documents mention transparency. For example the [UNICEF Policy guidance on AI for children](#) mentions the principle of 'Providing transparency, explainability, and accountability for children' (p. 38).^[15] This involves ensuring children understand how AI systems impact them. UNICEF's guidance also calls for explicitly addressing children when promoting the explainability and transparency of AI systems as well as utilising age-appropriate language. When referencing transparency, UNICEF explicates the importance of informing children when they are interacting with an AI system rather than a human. Transparency is also a principle found in the [UK ICO Age Appropriate Design Code](#) and calls for actions such as providing clear privacy information, delivering 'bite-sized' explanations to the user when personal data is used for training, posting clear policies, community standards, and terms of use, and using child-friendly depictions of information that are tailored to specific ages.^[16]

Be Accountable

The maxim of being accountable is about ensuring appropriate oversight of your AI/ML-assisted decision systems and being answerable to others in your organisation, external bodies such as regulators, and the individuals and the individuals impacted by AI/ML-assisted decisions.

The UK GDPR includes accountability as a principle which involves taking responsibility for complying with the other data protection principles and being able to demonstrate that compliance. It also mandates implementing appropriate technical and organisational measures, and data protection by design and default.^[17] Being accountable for explaining AI/ML-assisted decisions concentrates these dual requirements on the processes and actions you carry out when designing (or procuring/outsourcing) and deploying AI/ML models.



More details about GDPR can be found in [Workbook 5: Responsible Data Stewardship in Practice](#).

Key Aspects of Being Accountable

1. Assign responsibility

- Identify those within your organisation who manage and oversee the 'explainability' requirements of an AI decision-support system and assign ultimate responsibility for this.
- Ensure you have a designated and capable human point-of-contact for individuals to clarify or contest a decision.

2. Justify and evidence

- Actively consider and make justified choices about how to design and deploy AI/ML models that are appropriately explainable to individuals.
- Take steps to prove and document that you made these considerations, and that they are present in the design and deployment of the models themselves.
- Show that you provided explanations to individuals.



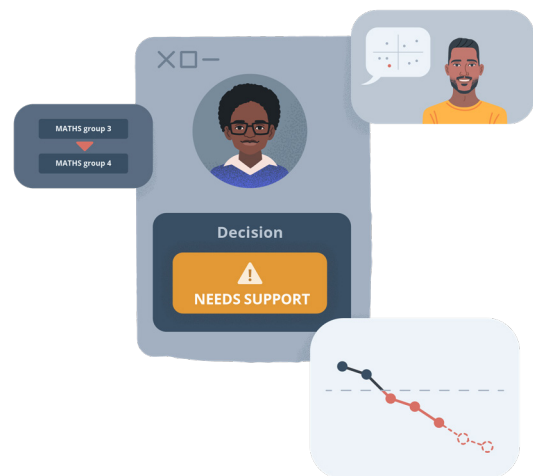
Considerations for Child-Centred AI

Accountability is also mentioned in the [UNICEF Policy guidance on AI for children](#) under the requirement of 'Providing transparency, explainability, and accountability for children' (pg. 38).^[21] It is critical for roles and responsibilities to be established within an organisation to ensure accountability for decisions made about children's data. Additionally, UNICEF states it is imperative that AI systems are developed so that they protect and empower child users according to legal and policy frameworks, **regardless of children's understanding of the system**. They state, 'the development of AI systems cannot ignore or exploit any child's lack of understanding or vulnerability' (pg. 39). Accountability should also be accompanied by AI oversight bodies with a specific focus on child rights and the inclusion of child rights experts.

Maxim 3

Consider Context^[22]

There is no one-size-fits-all approach to explaining AI/ML-assisted decisions. Considerations of context involve paying attention to several different, but interrelated, elements that can have an effect on explaining AI/ML-assisted decisions and managing the overall process.^[18] This is not a one-off consideration. It should be considered at all stages of the process, from concept to deployment and presentation of the explanation to the decision recipient.



Key Aspects of Considering Context

- 1. Choose appropriate models and explanation.** If you plan to use AI/ML to help make decisions about people, you should consider:
 - the setting;
 - the potential impact;
 - what an individual should know about a decision, so you can choose an appropriately explainable AI model; and
 - prioritising delivery of the relevant explanation types. Find out more about explanation types on [page 28](#).

2. Tailor governance and explanation. Your governance of the explainability of AI models should be:

- robust and reflective of best practice; and
- tailored to your organisation and the particular circumstances and needs of each stakeholder.

3. Identify the audience of the explanation. The audience of your explanation has an effect on what type of explanations are meaningful or useful for them. Explanations may be addressed to senior responsible owners, analysts, oversight bodies, individuals impacted by decisions, or others. You should consider:

- For end-users or implementers:
 - the depth and level of explanation that is appropriate to assist them in carrying out evidence-based reasoning in a way that is context-sensitive and aware of the model's limitations.
- For auditors:
 - The level and depth of explanation that is fit for the purpose of the relevant review.
- For individuals impacted by decisions:
 - the level of expertise they have about the decision;
 - the range of people subject to decisions made (to account for the range of knowledge or expertise);
 - whether the individuals require any reasonable adjustments in how they receive the explanation; and
 - how to accommodate the explanation needs of the most vulnerable individuals.^{[19] [20]}



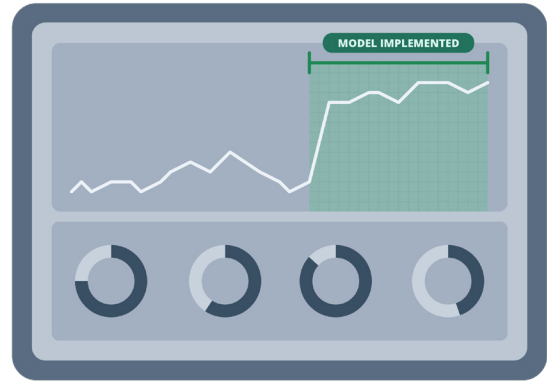
Considerations for Child-Centred AI

In addition to the nine requirements outlined in the [UNICEF Policy guidance on AI for children](#), there are overarching recommendations that apply in all contexts. This includes adapting the AI/ML system to the national or local context while keeping children in mind from design to deployment. These considerations should be taken into account from the design stage to avoid algorithmic bias resulting from contextual blindness. Additionally, the requirement of 'Ensure inclusion of and for children' (pg. 33) recommends active participation of children across all stages of the project lifecycle to ensure children are considered in the context of the system's intended use.^[23] When considering potential impacts, specific focus should be given to 'actively support the most marginalized children' (pg. 34) including girls, minority groups, children with disabilities, and those in refugee contexts to ensure that they may benefit from, and not be disadvantaged by AI systems.^[24]

Reflect on Impacts

In making decisions and performing tasks that have previously required the thinking and reasoning of responsible humans, AI/ML systems are increasingly serving as trustees of human decision-making. However, individuals cannot hold these systems directly accountable for the consequences of their outcomes and behaviours.^[25]

The value of reflecting on the impacts of your AI/ML system helps you explain to individuals affected by its decisions that the use of algorithmic techniques will not harm or impair their wellbeing.^[26] This means asking and answering questions about the ethical purposes and objectives of your AI/ML project at the initial stages.^[27]



You should then revisit and reflect on the impacts identified in the initial stages of the AI/ML project throughout the development and implementation stages.^[28] If any new impacts are identified, you should document them, alongside any implemented mitigation measures where relevant.^[29] This will help you explain to individuals impacted what impacts you have identified and how you have reduced any potentially harmful effects as best as possible.

Key Aspects of Reflecting on Impacts

1. **Ensure individual wellbeing.**^[30] Think about how to build and implement your AI/ML system in a way that;
 - fosters the physical, emotional, and mental integrity of impacted individuals;
 - ensures their abilities to make free and informed decisions about their own lives;
 - safeguards their autonomy and their power to express themselves;
 - supports their abilities to flourish, to fully develop themselves, and to pursue their interests according to their own freely determined life plans;
 - preserves their ability to maintain a private life independent from the transformative effects of technology; and
 - secures their capacities to make well-considered, positive, and independent contributions to their social groups and to the shared life of the community, more generally.

2. Ensure societal wellbeing. Think about how to build and implement your AI/ML system in a way that:

- safeguards meaningful human connection and social cohesion;
- prioritises diversity, participation, and inclusion;
- encourages all voices to be heard and all opinions to be weighed seriously and sincerely;
- treats all individuals equally and protects social equity;
- uses AI technologies as an essential support for the protection of fair and equal treatment under the law;
- utilises innovation to empower and to advance the interests and well-being of as many individuals as possible; and
- anticipates the wider impacts of the AI/ML technologies you are developing by thinking about their ramifications for others around the globe, for the biosphere as a whole, and for future generations.



Considerations for Child-Centred AI

Reflecting on impacts overlaps with the [UNICEF Policy guidance on AI for children](#) requirements of 'Prioritise fairness and non-discrimination for children' (pg. 34) and 'Protect children's data and privacy' (pg. 35). These requirements call for actively supporting marginalised children to ensure benefits from AI systems.^[31] This entails making certain that datasets include a diversity of children's data and implementing responsible data approaches to ensure that children's data is handled with care and sensitivity. The [Age Appropriate Design Code](#) contains the principle of 'Detrimental use of data' which states that children's data should not be used in any way that could negatively affect their well-being or go against industry standards, regulatory provisions, or government advice.^[32]

High-Level Considerations for Building Appropriately Explainable AI Systems

This section highlights four high-level considerations for project teams to consider so as to achieve higher degrees of explainability of the model and improved interpretability of outputs to wide and diverse audiences.

KEY CONCEPT Interpretability

In the context of AI/ML systems, interpretability is the degree to which a human can access and comprehend how and why a model performed the way it did in a specific context and therefore understand the rationale behind its output or behaviour. Interpretability is a concept that is adjacent to explainability, but it is more centred on the ability of human interpreters to grasp the interworking and underlying logic of an AI system.^[33]

Consideration 1: Context, Potential Impact, and Domain-Specific Needs



Look first to context, potential impact, and domain-specific needs when determining the interpretability requirements of your project. This includes considerations about:

- type of application;
- domain specific expectations, norms, and rules; and
- existing technology.

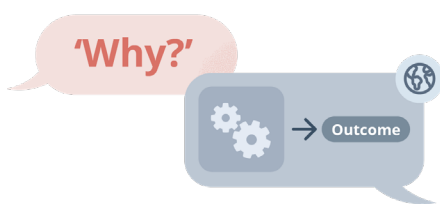
Consideration 1 in Depth

1. Type of application: Start by assessing both the kind of tool you are building and the environment in which it will apply. Understanding your AI system's purpose and context of application will give you a better idea of the stakes involved in its use and hence also a good starting point to think about the scope of its interpretability needs.

2. Domain specificity: By acquiring solid domain knowledge of the environment in which your AI system will operate, you will gain better insight into any potential sector-specific standards of explanation or benchmarks of justification which should inform your approach to interpretability.

3. Existing technology: If one of the purposes of your AI project is to replace an existing algorithmic technology that may not offer the same sort of expressive power or performance level as the more advanced AI techniques that you are planning to deploy, you should carry out an assessment of the performance and interpretability levels of the existing technology. This will provide you with an important reference point when you are considering possible trade-offs between performance and interpretability that may occur in your own prospective system.

Consideration 2: Draw on Standard Interpretable Techniques



Draw on standard interpretable techniques when possible. Find the right fit between:

- domain-specific risks and needs;
- available data resources and domain knowledge; and
- task appropriate AI/ML techniques.

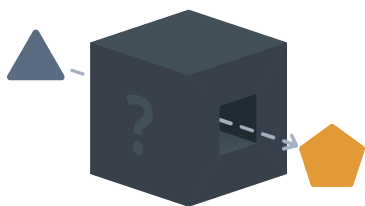
Consideration 2 in Depth

1. Domain specific risks and needs: High-impact, safety-critical, or other potentially sensitive environments heighten demands for thoroughgoing accountability and transparency. In some of these instances, such demands may make choosing standard but sophisticated non-opaque AI/ML techniques an overriding priority.

2. Available data resources and domain knowledge: Where data resources lend to well-structured, meaningful representations and domain expertise can be incorporated into model design, interpretable techniques may often be more desirable than opaque ones. Careful data pre-processing and iterative model development can, in these cases, hone the accuracy of such interpretable systems in ways that may make the advantages gained by the combination of their performance and transparency outweigh the benefits of more opaque approaches.

3. Task appropriate AI/ML techniques: For use cases where AI/ML-based predictive modelling or classification involves tabular or structured data, interpretable techniques may hold advantages, but for tasks in areas like computer vision, natural language processing, and speech recognition, where unstructured and high-dimension data is required, drawing on standard interpretable techniques will not be possible.

Consideration 3: Considerations in Using 'Black Box' AI Systems



When considering the use of a 'black box' AI system, you should proceed with diligence and:

- thoroughly weigh up the potential impacts and risk of using an opaque model;
- consider options for supplemental interpretability tools; and
- formulate an action plan to optimise explainability.

Consideration 3 in Depth

1. Thoroughly weigh up impacts and risks:

As a general policy, you and your team should utilise 'black box' models only where their potential impacts and risks have been thoroughly considered in advance, and you have determined that your use case and domain specific needs support the responsible design and implementations of these systems.

2. Consider supplemental interpretability tools:

Consider what sort of explanatory resources the interpretability tool will provide users and implementers in order (1) to enable them to exercise better-informed evidence-based judgments and (2) to assist them in offering plausible, sound, and reasonable accounts of the logic behind algorithmically generated output to affected individuals and concerned parties.

3. Formulate an action plan to optimise explainability:

This should include a clear articulation of the explanatory strategies your team intends to use, a detailed plan that indicates the stages in the project workflow when the design of these strategies will need to take place, and a succinct formulation of your explanation priorities and delivery strategy. Further actions are detailed in the Six Tasks for Explainability Assurance Management that are described in Part Two of this guidance.

KEY CONCEPT

Black Box Model

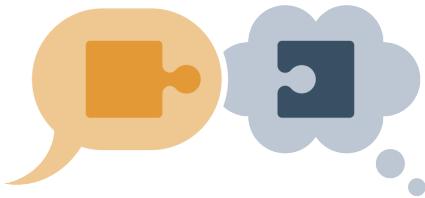
We define a 'black box' model as any AI system whose inner workings and rationale are opaque or inaccessible to human understanding. These systems may include:

- neural networks, including recurrent and convolutional neural networks (models consisting of interconnected nodes that make predictions based on correlations from input data);
- ensemble methods (such as the random forest technique that strengthens an overall prediction by combining and aggregating the results of several or many different base models); or
- support vector machines (a classifier that uses a special type of mapping function to build a divider between two sets of features in a high dimensional space).

For example of model types, see [Appendix A](#).

KEY CONCEPT**Explainable AI**

Explainable AI, in the strictly technical sense, (or XAI) refers to a set of processes and methods designed to help individuals comprehend the results and outputs of opaque AI/ML systems. [Appendix B](#) provides a sample of the supplementary techniques and tools of explainable AI that have been developed to assist in providing access to the underlying logic of 'black box' models.^[34]

Consideration 4: Interpretability and Human Understanding

Think about interpretability in terms of the capacities of human understanding and its limitations.

Consideration 4 in Depth

When considering the interpretability needs of your AI project, it is important to start by thinking about the capacities and limitations of human cognition. Unlike complex AI/ML systems that map inputs to outputs by connecting hundreds, thousands, millions, or even billions of variables at once, human understanding functions by combining and connecting a relatively small number of variables at a time. Consequently, simplicity, or informational parsimony, is crucial for ensuring interpretable AI. Seeing interpretability as a continuum of comprehensibility that is dependent on the capacities and limits of the individual human interpreter should key you into what is needed in order to deliver an interpretable AI system. Such limits to consider should include not only cognitive boundaries but also varying levels of access to relevant vocabularies of explanation and varying levels of expertise and technical literacy.

Types of Explanation

Context determines what information is required, useful, or accessible to explain decisions involving AI and, therefore, what types of explanations are the most appropriate. In this section, we introduce six explanation types designed to help your AI project team build concise and clear explanations. Each explanation type addresses explanations of the outcomes and processes that relate to a SSAFE-D (Sustainability, Safety, Accountability, Fairness, Explainability, and Data Stewardship) principle, highlighting the intertwined relation between these SSAFE-D principles. These explanation types are further divided into the subcategories of **Process-Based** and **Outcome-Based** explanations.



More details about the SSAFE-D Principles can be found in [Workbook 1: AI Ethics and Governance in Practice: An Introduction](#).



Rationale Explanation

Helps people understand the reasons that led to a decision outcome.



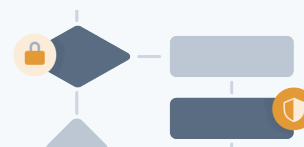
Fairness Explanation

Helps people understand the steps taken to ensure AI decisions are generally unbiased and equitable, and whether or not they have been treated equitably themselves.



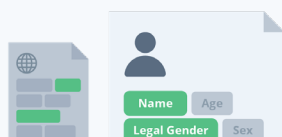
Responsibility Explanation

Helps people understand who is involved in the development and management of the AI model, and who to contact for a human review of a decision.



Safety Explanation

Helps people understand the measures that are in place and the steps taken to maximise the performance, reliability, security, and robustness of the AI outcomes, as well as what is the justification for the chosen type of AI system.



Data Explanation

Helps people understand what data about them, and what other sources of data, were used in a particular AI decision, as well as the data used to train and test the AI model.



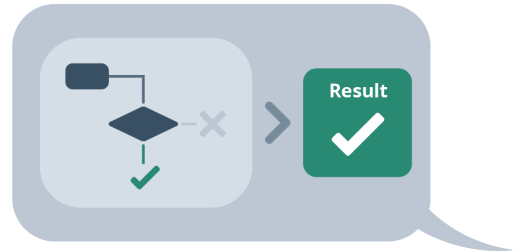
Impact Explanation

Helps people understand the considerations taken about the effects that the AI decision-support system may have on an individual and society.

Rationale Explanation

What Does This Explanation Type Help People to Understand?

It is about the 'why?' of an AI decision. It helps people understand the reasons that led to a decision outcome, in an accessible way.



What You May Need to Show

- How the system performed and behaved to get to that decision outcome.
- How the different components in the AI system led it to transform inputs into outputs in a particular way, so you can communicate which features, interactions, and parameters were most significant.
- How these technical components of the logic underlying the result can provide supporting evidence for the decision reached.
- How this underlying logic can be conveyed as easily understandable reasons to decision recipients.
- How you have thought about how the system's results apply to the concrete context and life situation of the affected individual.

Rationale Explanations Might Answer

- Have we selected an algorithmic model, or set of models, that will provide a degree of interpretability that corresponds with its impact on affected individuals?
- Are the supplementary explanation tools capable of providing meaningful and accurate information about our complex system's underlying logic?

Process-Based Explanations clarify...

- How the procedures you have set up help you provide meaningful explanations of the underlying logic of your AI model's results.
- How these procedures are suitable given the model's particular domain context and its possible impacts on the affected individuals and wider society.
- How you have set up your system's design and deployment workflow so that it is appropriately interpretable and explainable, including its data collection and preprocessing, model selection, explanation extraction, and explanation delivery procedures.

Outcome-Based Explanations provide...

- The formal and logical rationale of the AI system. How the system is verified against its formal specifications, so you can verify that the AI system will operate reliably and behave in accordance with its intended functionality.
- The technical rationale of the system's output. How the model's components (its variables and rules) transform inputs into outputs, so you know what role these components play in producing that output. By understanding the roles and functions of the individual components, it is possible to identify the features and parameters that significantly influence an output.
- Translation of the system's workings - its input and output variables, parameters and so on – into accessible everyday language, so you can clarify, in plain and understandable terms, what role these factors play in reasoning about the real-world problem that the model is trying to address or solve.
- Clarification on how a statistical result is applied to the individual concerned. A decision from the AI system will usually have a probability associated with it corresponding to the confidence of the AI model of that decision. You can specify how that probability influenced the final decision made, the confidence threshold at which decisions were accepted, and the reasonings behind choosing that threshold.



Considerations for Child-Centred AI

Rationale explanation begins with explaining how the system operated the way it did. Children should be able to fully understand how their data was mapped throughout the AI system to determine a specific output. Rationale explanation often involves more technical concepts than some of the other explanation types; thus it is critical that these technical explanations of model choice, the system's inner workings, and statistical results are delivered in an age-appropriate manner. One way to assist with rationale explanation is by having children involved from the design stages of the AI system so that they are informed upfront of the types of models being used as well as being appraised of model decisions made along the way.^[35]

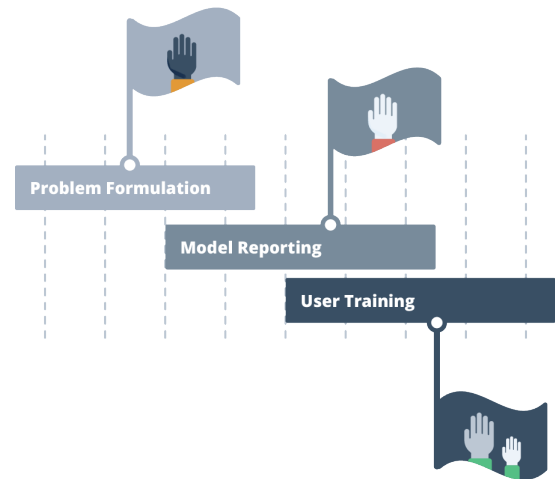


A rationale explanation seeks to provide information about the outcomes and processes involved in implementing the principle of Explainability, as detailed within this workbook.

Responsibility Explanation

What Does This Explanation Type Help People to Understand?

It helps people understand 'who' is involved in the development and management of the AI model, and 'who' to contact for a human review of a decision.



What You May Need to Show

- Who is accountable at each stage of the AI system's design and deployment, from defining outcomes for the system at its initial phase of design, through to providing the explanation to the affected individual at the end.
- Definitions of the mechanisms by which people will be held accountable, as well as how you have made the design and implementation processes of your AI system traceable and auditable.

Process-Based Explanations clarify...

- The roles and functions across your organisation that are involved in the various stages of developing and implementing your AI system, including any human involvement in the decision-making. If your system, or parts of it, are procured, you should include information about the providers or developers involved.
- Broadly, what the roles do, why they are important, and where overall responsibility lies for management of the AI model – who is ultimately accountable.
- Who is responsible at each step from the design of an AI system through to its implementation to make sure that there is effective accountability throughout.

Outcome-Based Explanations

Because a responsibility explanation largely has to do with the governance of the design and implementation of AI systems, it is, in a strict sense, entirely process-based. Even so, there is important information about post-decision procedures that you should be able to provide:

- Cover information on how to request a human review of an AI-enabled decision or object to the use of AI, including details on who to contact, and what the next steps will be (e.g. how long it will take, what the human reviewer will take into account, how they will present their own decision and explanation).
- Give individuals a way to directly contact the role or team responsible for the review. You do not need to identify a specific person in your organisation. One person involved in this should have implemented the decision and used the statistical results of a decision-support system to come to a determination about an individual.



Considerations for Child-Centred AI

Responsible explanation follows from responsible AI that ensures systems are verifiably ethical, beneficial, legal, and robust, and that organisations that deploy or use such systems are held accountable. The [UNICEF Policy guidance on AI for children](#) guidance further expands on this by stating everyone in the AI ecosystem needs a clear understanding which includes who designed an AI system and for what purpose. Responsible explanation also aligns with UNICEF's principle of 'Provide transparency, explainability, and accountability for children', which emphasises any form of explanation should strive to explicitly address children and parents/caregivers.^[36] Importantly, a child who directly or indirectly interacts with an AI system (e.g. a toy, chatbot or online system) has the 'right for explanation at an age-appropriate level and inclusive manner'. It is also the responsibility of the developer and deployer of a system to set up mechanisms for redress and encourage the reporting of potentially harmful features. This should be accessible by children and parents/caregivers who can easily navigate and understand who is responsible for the system and how to contact them. The [UK ICO Age-Appropriate Design Code](#) states online tools should include mechanisms for children and parents/caregivers to communicate with system providers and track their complaints or requests.^[37] UNICEF's guidance also calls for the establishment of AI oversight bodies consisting of a multifaceted and interdisciplinary range of stakeholders responsible for auditing systems and receiving and addressing user appeals.

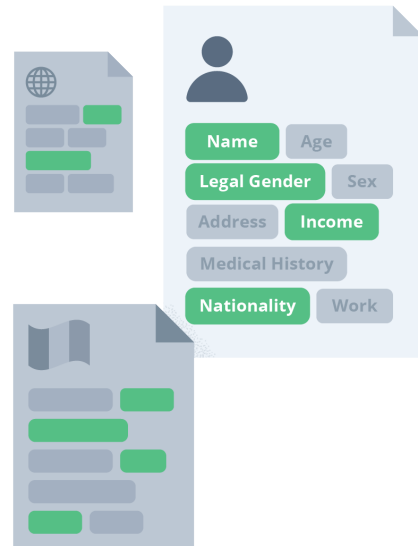


A responsibility explanation seeks to provide information about the outcomes and processes involved in implementing the principle of Accountability. For a comprehensive understanding or refresher on what this principle encompasses, please consult the [AI Accountability in Practice](#) workbook.

Data Explanation

What Does This Explanation Type Help People to Understand?

Data explanations are about the 'what' of AI-assisted decisions. They help people understand what data is held about them, and what other sources of data, were used in a particular AI decision. Generally, they also help individuals understand more about the data used to train and test the AI model, how this data is stored and, if no longer used for training, how it was deleted. You could provide some of this information within the fair processing notice you are required to provide under Articles 13 and 14 of the GDPR.



What You May Need to Show

- How the data used to train, test, and validate your AI model was managed and utilised from collection through to processing and monitoring (and deletion if applicable).
- What data you used in a particular decision and how.

Process-Based Explanations include...

- What training/testing/validating data was collected, the sources of that data, and the methods that were used to collect it.
- Who took part in choosing the data to be collected or procured and who was involved in its recording or acquisition. This should include how procured or third-party provided data was vetted.
- How data quality was assessed and the steps that were taken to address any quality issues discovered, such as completing or removing data.
- What the training/testing/validating split was and how it was determined.
- How data pre-processing, labelling, and augmentation supported the interpretability and explainability of the model.
- What measures were taken to ensure the data used to train, test, and validate the system was representative, relevant, accurately measured, and generalisable.
- How you ensured that any potential bias and discrimination in the dataset have been mitigated.

Outcome-Based Explanations

- Clarify the input data used for a specific decision, and the sources of that data. This is outcome-based because it refers to your AI system's result for a particular stakeholder.
- In some cases, the output data may also require an explanation, particularly where a user has been placed in a category which may not be clear to them. For example, in the case of anomaly detection for financial fraud identification, the output might be a distance measure (i.e. a distance calculated using various statistical or ML techniques that serves as a classification or scoring mechanism) which places them at a certain distance away from other people based on their transaction history.^[38] Such a classification may require an explanation.

Considerations for Child-Centred AI

When considering data explanation, it is critical that children's data agency is promoted and kept at the forefront of all decisions made along the way. The [UNICEF Policy guidance on AI for children](#) recommends that a privacy-by-design approach is taken when designing and implementing AI systems that use children's data. As data explanation tends to utilise more technical-based language — as seen in the Rational Explanation section above — than the other remaining explanation types, it is extremely important to reflect upon how these technical terms and systems can be explained to children in age-accessible language. Showing what data was used in a particular decision will assist with contributing towards transparency and building trust amongst children and organisations using their data to design, develop, and deploy AI systems. It is imperative that data equity is placed at the forefront to ensure that a diverse range of children's data are included and that transparent reporting was in place to demonstrate that this was achieved. The [UK ICO Age Appropriate Design Code](#) contains principles of data minimisation—collecting only the minimum amount of personal data necessary to carry out the system—and data sharing—considering the best interests of the child when contemplating sharing data, both of which should be implemented accordingly.

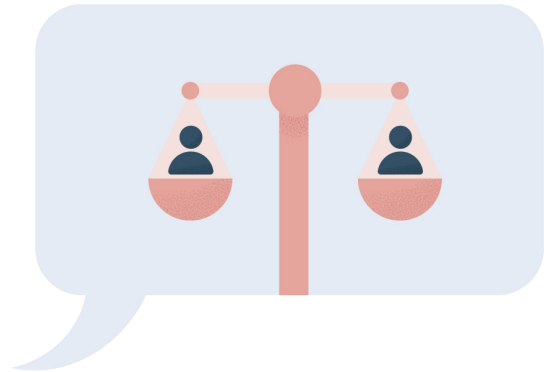


A data explanation type seeks to provide information about the outcomes and processes involved in implementing the principle of Data Stewardship. For a comprehensive understanding or refresher on what this principle encompasses, please consult the [Responsible Data Stewardship in Practice](#) workbook.

Fairness Explanation

What Does This Explanation Type Help People to Understand?

The fairness explanation is about helping people understand the steps you took (and continue to take) to ensure your AI decisions are generally unbiased and equitable. It also gives people an understanding of whether or not they have been treated equitably themselves.



What You May Need to Show

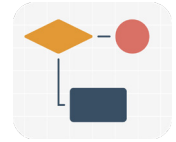
An explanation of fairness can relate to several stages of the design, development, and deployment of AI systems:

1. Data Fairness. The system is trained and tested on properly representative, relevant, accurately measured, and generalisable datasets (note that this dataset fairness component will overlap with data explanation). This may include showing that you have made sure your data is:



- as representative as possible of all those affected;
- sufficient in terms of its quantity and quality, so it represents the underlying population and the phenomenon you are modelling;
- assessed and recorded through suitable, reliable and impartial sources of measurement and has been sourced through sound collection methods;
- up-to-date and accurately reflects the characteristics of individuals, populations and the phenomena you are trying to model; and
- relevant by calling on domain experts to help you understand, assess and use the most appropriate sources and types of data to serve your objectives.

2. Model Design and Development Fairness. It has model architectures that do not include target variables, features, processes, or analytical structures (correlations, interactions, and inferences) which are unreasonable or unjustifiable. This may include showing that you have done the following:



- Attempted to identify any underlying structural biases that may play a role in translating your objectives into target variables and measurable proxies. When defining the problem at the start of the AI project, these biases could influence what system designers expect target variables to measure and what they statistically represent.
- Mitigated bias in the data pre-processing phase by taking into account the sector or organisational context in which you are operating. When this process is automated or outsourced, show that you have reviewed what has been done, and maintained oversight. You should also attach information on the context to your metadata, so that those coming to the pre-processed data later on have access to the relevant properties when they undertake bias mitigation.
- Mitigated bias when the feature space was determined (i.e. when relevant features were selected as input variables for your model). Choices made about grouping or separating and including or excluding features, as well as more general judgements about the comprehensiveness or coarseness of the total set of features, may have consequences for protected groups of people.
- Mitigated bias when tuning parameters and setting metrics at the modelling, testing, and evaluation stages (i.e. into the trained model). Your AI development team should iterate the training of the model and peer review it to help ensure that how they choose to adjust the parameters and metrics of the model are in line with your objectives of mitigating bias.
- Mitigated bias by watching for hidden proxies for discriminatory features in your trained model, as these may act as influences on your model's output. Designers should also look into whether the significant correlations and inferences determined by the model's learning mechanisms are justifiable.

3. Metric-Based Fairness. It does not have discriminatory or inequitable impacts on the lives of the people it affects. This may include showing that:



- you have been explicit about the formal definition(s) of fairness you have chosen and why. Data scientists can apply different formalised fairness criteria to choose how specific groups in a selected set will receive benefits in comparison to others in the same set, or how the accuracy or precision of the model will be distributed among subgroups; and
- the method you have applied in operationalising your formalised fairness criteria (for example) adjust for outcome preferences by reweighting model parameters, embedding trade-offs in a classification procedure, or re-tooling algorithmic results.

4. System Implementation Fairness. It is deployed by users sufficiently trained to implement it responsibly and without bias. This may include showing that you have appropriately prepared and trained the implementers of your system to:



- Avoid automation bias (over-relying on the outputs of AI systems) or automation-distrust bias (under-relying on AI system outputs because of a lack of trust in them).
- Use its results with an active awareness of the specific context in which they are being applied. They should understand the particular circumstances of the individual to which that output is being applied; and understand the limitations of the system. This includes understanding the statistical uncertainty associated with the result as well as the relevant error rates and performance metrics.
- Be up-to-date and accurately reflect the characteristics of individuals, populations and the phenomena you are trying to model.
- Be relevant by calling on domain experts to help you understand, assess and use the most appropriate sources and types of data to serve your objectives.

Process-Based Explanations include...

- Your chosen measures to mitigate risks of bias and discrimination at the data collection, preparation, model design and testing stages.
- How these measures were chosen and how you have managed informational barriers to bias-aware design such as limited access to data about protected or sensitive traits of concern.
- How the measures chosen impacts other performance and fairness metrics, and this influence on the final decisions.
- The results of your initial (and ongoing) fairness testing, self-assessment, and external validation – showing that your chosen fairness measures are deliberately and effectively being integrated into model design. You could do this by showing that different groups of people receive similar outcomes, or that protected characteristics have not played a factor in the results.

Outcome-Based Explanations include...

- Details about how your formal fairness criteria were implemented in the case of a particular decision or output.
- Presentation of the relevant fairness metrics and performance measurements in the delivery interface of your model. This should be geared to a non-technical audience and done in an easily understandable way.
- Explanations of how others similar to the individual were treated (i.e. whether they received the same decision outcome as the individual). For example, you could use information generated from counterfactual scenarios to show whether or not someone with similar characteristics, but of a different ethnicity or gender, would receive the same decision outcome as the individual.



Considerations for Child-Centred AI

Fairness explanation contains facets that overlap with the [UNICEF Policy guidance on AI for children](#) principle of 'Prioritizing fairness and non-discrimination'. This principle includes providing active support for the most marginalised children so that they can receive benefits from AI systems. As stated in the UNICEF guidance, this requires attention to 'the differences in cultural, social, and regional contexts of AI-related policies and activities', which should include considerations that expand past ensuring access to these technologies—although this still remains a key barrier to accessing the benefits that AI systems may provide. Several other key points outlined by UNICEF include ensuring a diversity of children's data in new datasets that are being developed and removing bias against children or certain groups of children. In addition to ensuring data representativeness and completeness, it is critical that teams consider the trade-off of various fairness metrics and how these could affect children differently. How will these decisions be reported in a way that is transparent and accessible for children?

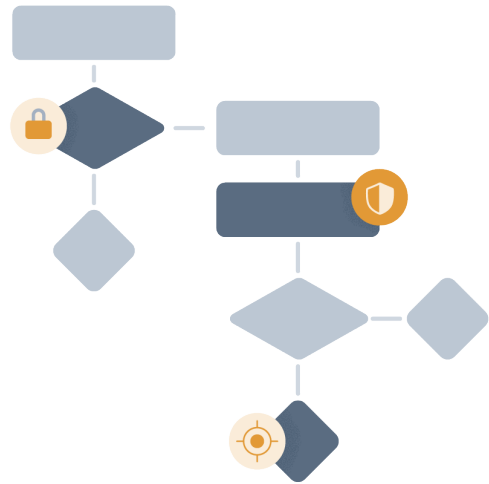


A fairness explanation type seeks to provide information about the outcomes and processes involved in implementing the principle of Fairness. For a comprehensive understanding or refresher on what this principle encompasses, please consult the [AI Fairness in Practice](#) workbook.

Safety Explanation

What Does This Explanation Type Help People to Understand?

The safety explanation helps people understand the measures you have put in place, and the steps you have taken (and continue to take) to maximise the performance, reliability, security, and robustness of AI-assisted decisions. It can also be used to justify the type of AI system you have chosen to use, such as comparisons to other systems or human decision makers.



What You May Need to Show

- The proportion of examples for which your model generates a correct output. This component may also include other related performance measures such as precision, sensitivity (true positives), and specificity (true negatives). Individuals may want to understand how accurate, precise, and sensitive the output was in their particular case.
- How dependably the AI system does what it was intended to do. If it did not do what it was programmed to carry out, individuals may want to know why, and whether this happened in the process of producing the decision that affected them.
- The system is able to protect its architecture from unauthorised modification or damage of any of its component parts. The system remains continuously functional and accessible to its authorised users and keeps confidential and private information secure, even under hostile or adversarial conditions.
- The system functions reliably and accurately in practice. Individuals may want to know how well the system works if things go wrong, how this has been anticipated and tested, and how the system has been immunised from adversarial attacks.

Process-Based Explanations Include...



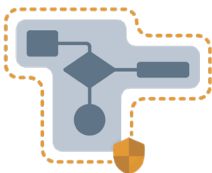
1. Performance Safety Objective

- How you measure it (e.g. maximising precision to reduce the risk of false negatives).
- Why you chose those measures, and how you went about assuring it.
- What you did at the data collection stage to ensure your training data was up-to-date and reflective of the characteristics of the people to whom the results apply.
- What kinds of external validation you have undertaken to test and confirm your model's 'ground truth'.
- What the overall accuracy rate of the system was at testing stage.
- What you do to monitor this (e.g. measuring for concept drift over time).



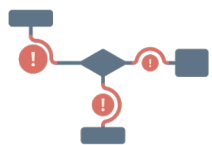
2. Reliability Safety Objective

- How you measure it and how you went about assuring it.
- Results of the formal verification of the system's programming specifications, i.e. how encoded requirements have been mathematically verified.



3. Security Safety Objective

- How you measure it and how you went about assuring it (e.g. how limitations have been set on who is able to access the system, when, and how).
- How you manage the security of confidential and private information that is processed in the model.



4. Robustness Safety Objective

- How you measure it.
- Why you chose those measures.
- How you went about assuring it, e.g. how you've stress-tested the system to understand how it responds to adversarial intervention, implementer error, or skewed goal-execution by an automated learner (in reinforcement learning applications).

Outcome-Based Explanations Include...

While you may not be able to guarantee accuracy at an individual level, you should be able to provide assurance that, at run-time, your AI system operated reliably, securely, and robustly for a specific decision. In the case of accuracy and the other performance metrics, however, you should include in your model's delivery interface the results of your cross-validation (training/testing splits) and any external validation carried out.

You may also include relevant information related to your system's confusion matrix (the table that provides the range of performance metrics) and ROC curve (receiver operating characteristics)/AUC (area under the curve). Include guidance for users and affected individuals that makes the meaning of these measurement methods, and specifically the ones you have chosen to use, easily accessible and understandable. This should also include a clear representation of the uncertainty of the results (e.g. confidence intervals and error bars).



Considerations for Child-Centred AI

The safety of an AI system is particularly important when considering children as they have unique needs and considerations. The [UNICEF Policy guidance on AI for children](#) child-centric principle of 'Ensuring safety for children', draws attention to various considerations that should be in place. The first of these is a mechanism for continuous monitoring and assessment of the impact of AI systems on children as well as continuous monitoring of these impacts throughout the entire lifecycle. UNICEF also calls for the testing of AI systems using children's data for safety, security, and robustness.



A safety explanation type seeks to provide information about the outcomes and processes involved in implementing the principle of Safety. For a comprehensive understanding or refresher on what this principle encompasses, please consult the [AI Safety in Practice](#) workbook.

Impact Explanation

What Does This Explanation Type Help People to Understand?

An impact explanation helps people understand how you have considered the effects that your AI decision-support system may have on an individual, i.e. what the outcome of the decision means for them. It is also about helping individuals to understand the broader societal effects that the use of your system may have. This may help reassure people that the use of AI will be of benefit. Impact explanations are therefore often well suited to delivery before an AI-assisted decision has been made.



What You May Need to Show

Demonstrate that you have thought about how your AI system will potentially affect individuals and wider society. Clearly show affected individuals the process you have gone through to determine these possible impacts.

Process-Based Explanations include...

- Showing the considerations you gave to your AI system's potential effects, how you undertook these considerations, and the measures and steps you took to mitigate possible negative impacts on society, and to amplify the positive effects.
- Information about how you plan to monitor and re-assess impacts while your system is deployed.

Outcome-Based Explanations

Although the impact explanation is mainly about demonstrating that you have put appropriate forethought into the potential 'big picture' effects, you should also consider how to help decision recipients understand the impact of the AI-assisted decisions that specifically affect them. For instance, you might explain the consequences for the individual of the different possible decision outcomes and how, in some cases, changes in their behaviour would have brought about a different outcome with more positive impacts. This use of counterfactual assessment would help affected individuals make changes that could lead to a different outcome in the future or allow them to challenge the decision.



Considerations for Child-Centred AI

It is critical that the potential impacts of an AI system that uses children's data are fully considered and weighed. This is especially relevant for systems that are not intended for children to use but that children may have access to, such as smart devices in the household. In order to fully understand possible impacts, it is imperative that organisations engage with children to understand how possible impacts will differ from other audiences due to children's specific contexts and needs. Negative impacts of AI systems if not considered properly before deployment could have long-term effects on children's mental health and well-being, future pathways, safety and security, amongst many others. One way to go about considering all of the potential impacts is to engage in a meaningful way with children through the entire AI project lifecycle to effectively investigate impacts that may not have been thought of.

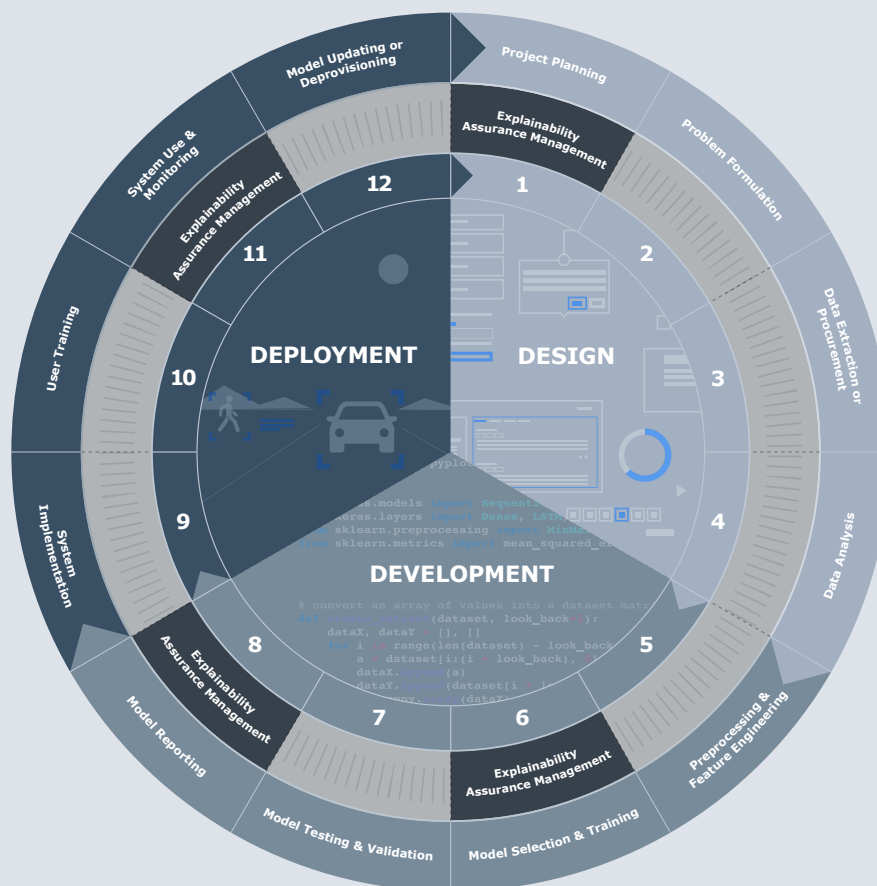


An impact explanation type seeks to provide information about the outcomes and processes involved in implementing the principle of Sustainability. For a comprehensive understanding or refresher on what this principle encompasses, please consult the [AI Sustainability in Practice Part One](#) and [AI Sustainability in Practice Part Two](#) workbooks.

Part Two: Putting the Principle of Explainability Into Practice

There are a number of tasks both to help you design and deploy appropriately transparent and explainable AI systems and to assist you in providing clarification of the results these systems produce to a range of impacted stakeholders (from operators, implementers, and auditors to decision recipients). These tasks make up Explainability Assurance Management for AI projects, offering a systematic approach to:

- designing, developing, and deploying AI projects in a transparent and explanation-aware fashion; and
- selecting, extracting and delivering explanations that are differentiated according to the needs and skills of the different audiences they are directed at.



Task 1

Select Priority Explanations by Considering the Domain, Use Case, and Impact on the Individual



Related AI Lifecycle Stage:

Design Phase

1. Getting to know the different types of explanation will help you identify the dimensions of an explanation that decision recipients will find useful.
2. In most cases, explaining AI-assisted decisions involves identifying what is happening in your AI system and who is responsible. That means you should prioritise the rationale and responsibility explanation types.
3. The setting and sector you are working in is important in figuring out what kinds of explanation you should be able to provide. You should therefore consider domain, context, and use case.
4. In addition, consider the potential impacts of your use of AI to determine which other types of explanation you should provide. This will also help you think about how much information is required, and how comprehensive it should be.
5. Choosing what to prioritise is not an exact science, and while your choices may reflect what the majority of the people you make decisions about want to know, it is likely that other individuals will still benefit from the explanations you have not prioritised. These will probably also be useful for your own accountability or auditing purposes.



Considerations for Child-Centred AI

At the Project Planning stage, when considering children's rights or projects that involve children, additional care needs to be made where children's data is to be included as part of AI systems. In addition to explanations related to the system itself and who is responsible, increased transparency on the use and processing of data should be provided for parents, guardians, and children that help explain their participation in simple language that is easy to understand. More details should also be offered regarding the risks of not abiding by child-centric requirements as part of the Project Planning stage, where justifications for children's involvement should be clearly outlined.

Task 2

Collect and Pre-Process Your Data in an Explanation-Aware Manner



Related AI Lifecycle Stages:

Data Extraction or Procurement

Data Analysis

1. The data that you collect and pre-process before inputting it into your system has an important role to play in the ability to derive each explanation type.
2. Careful labelling and selection of input data, as discussed in the [Responsible Data Stewardship in Practice](#) workbook, can help provide information for your rationale explanation.
3. To be more transparent you may wish to provide details about who is responsible at each stage of Data Collection and Preprocessing. You could draw from your Workflow Governance Map (provided in the [AI Accountability in Practice](#) workbook) and provide this as part of your responsibility explanation.
4. To aid your data explanation, you could draw from your Data Factsheet (provided in the [Responsible Data Stewardship in Practice](#) workbook) to include details on:
 - the source of the training data;
 - how it was collected;
 - assessments about its quality; and
 - steps taken to address quality issues, such as completing or removing data.
5. You should check the data used within your model to ensure it is sufficiently representative of those you are making decisions about. You should also consider whether pre-processing techniques, such as re-weighting, are required. These decisions should be documented in your Bias Self-Assessments (provided in the [AI Fairness in Practice](#) workbook), and will help your fairness explanation.
6. You should ensure that the Modelling, Testing, and Monitoring stages of your system development lead to accurate results. These results should be documented in your Safety Self-Assessments to aid your safety explanation.
7. Documenting your Stakeholder Impact Assessment, and steps taken throughout the model design to implement the results of these assessments, will aid in your impact explanation.



Considerations for Child-Centred AI

When considering data extraction, procurement, and analysis of children's data, it is important to ensure that the [UNICEF Policy guidance on AI for children](#) and other normative tools regarding the responsible use of children's data, such as the GDPR and the [UK ICO Age Appropriate Design Code](#) is applied. We would suggest that any data-related to children is pseudonymised or anonymised to limit potential harms. Under current data protection regulations, children under the age of 13 are unable to consent to the use of their personal data, so the lawful basis for processing such personal data must be clearly communicated, in consultation with children as well as their parents or guardians should personal data be used and what the potential impact may be.

Task 3

Build Your System to Ensure You Are Able to Extract Relevant Information for a Range of Explanation Types



Related AI Lifecycle Stage:

Model Selection & Training

1. Deriving the rationale explanation is key to understanding your AI system and helps you comply with parts of the GDPR. It requires looking 'under the hood' and helps you gather the information you need for some of the other explanations, such as safety and fairness. However, this is a complex task that requires you to know when to use more and less interpretable models and how to understand their outputs.
2. To choose the right AI model for your explanation needs, you should think about the domain you are working in, and the potential impact of the system.
3. When you are processing social or demographic data, which can contain unfair biases or lurking discriminatory proxies, you need to choose a more interpretable model or have safeguards in place to sufficiently mitigate and manage these biases.
4. When selecting a model for your project, you should consider whether:
 - There are costs and benefits of using a newer and potentially less explainable AI model;
 - the data you use requires a more or less explainable system;
 - your use case and domain context encourage choosing an inherently interpretable system;
 - your processing needs lead you to select a 'black box' model; and
 - the supplementary interpretability tools that help you to explain a 'black box' model (if chosen) are appropriate in your context.
5. To extract explanations from inherently interpretable models, look at the logic of the model's mapping function by exploring it and its results directly.
6. To extract explanations from 'black box' systems, there are many techniques you can use. Make sure that they provide a reliable and accurate representation of the system's behaviour.



Considerations for Child-Centred AI

The Turing-UNICEF Pilot Project on [Understanding AI Ethics and Safety for Children](#) established the difficulties that adults and children face in understanding the inner workings of complex systems and models. As a result, identifying and mitigating risks is key to ensuring that the selected AI model is justified in its use. Data risk management frameworks (tools and methodologies that aim to establish clarity on the benefits and risks of data and datasets) may be useful for this process. Given that it may be impossible to transparently communicate black box AI systems to children and their parents or guardians, the focus should turn to documenting model selection processes and ensuring that supplemental interpretability tools are used where appropriate.

Interpretable Algorithms

When possible and application-appropriate, draw on standard and algorithmic techniques that are as interpretable as possible.

In high impact, safety-critical, or other potentially sensitive environments, you are likely to need an AI system that maximises accountability and transparency. In some cases, this will mean you prioritise choosing standard but sophisticated non-opaque techniques. These techniques (some of which are outlined in the table in [Appendix A](#)) may include decision trees/rule lists, linear regression and its extensions like generalised additive models, case-based reasoning, or logistic regression. In many cases, reaching for the 'black box' model first may not be appropriate and may even lead to inefficiencies in project development. This is because more interpretable models are also available, which perform very well but do not require supplemental tools and techniques for facilitating interpretable outcomes.

Careful data pre-processing and iterative model development can hone the accuracy of interpretable systems. As a result, the advantages gained by the combination of their improved performance and their transparency may outweigh those of less transparent approaches.

'Black Box' AI Systems

When you consider using opaque algorithmic techniques, make sure that the supplementary interpretability tools that you will use to explain the model are appropriate to meet the domain-specific risks and explanatory needs that may arise from deploying it.

For certain data processing activities, it may not be feasible to use straightforwardly interpretable AI systems. For example, the most effective machine learning approaches are likely to be opaque when you are using AI applications to classify images, recognise speech, or detect anomalies in video footage. The feature spaces of these kinds of AI systems grow exponentially to hundreds of thousands or even millions of dimensions. At this scale of complexity, conventional methods of interpretation no longer apply.

You should only use 'black box' models if you have thoroughly considered their potential impacts and risks in advance. The members of your team should also have determined that your use case and your organisational capacities/resources support the responsible design and implementation of these systems.

Likewise, you should only use them if supplemental interpretability tools provide your system with a domain-appropriate level of explainability. This needs to be reasonably sufficient to mitigate the potential risks of the system and provide decision recipients with meaningful information about the rationale of any given outcome. A range of supplementary techniques and tools that assist in providing some access to the underlying logic of 'black box' models is explored below and in [Appendix B](#).

Task 4

Translate the Rationale of Your System's Results Into Useable and Easily Understandable Reasons



Related AI Lifecycle Stage:

Model Reporting

1. Once you have extracted the rationale of the underlying logic of your AI model, you will need to take the statistical output and incorporate it into your wider decision-making process.
2. Implementers of the outputs from your AI system will need to recognise the factors that they see as legitimate determinants of the outcome they are considering.
3. For the most part, the AI systems we consider in this workbook will produce statistical outputs that are based on correlation rather than causation. You therefore need to check whether the correlations that the AI model produces make sense in the case you are considering.
4. Decision recipients should be able to easily understand how the statistical result has been applied to their particular case.



Considerations for Child-Centred AI

Considerations for Child Centred AI: In conjunction with the previous tasks, model reporting on projects related to children's rights and data should be explained in simple language to ensure that children and their parents or guardians understand the impact of the model's results. While it may not be possible to state statistical outputs in non-technical terms, your team should endeavour to outline what different generalised results or outcomes from the model mean and how that translates to informing real-world decision-making processes. This includes explaining the different inputs into the model and why those specific bits of information are used.

Task 5

Prepare Implementers to Deploy Your AI System



Related AI Lifecycle Stage:

User Training

1. In cases where decisions are not fully automated, implementers need to be meaningfully involved. This means that they need to be appropriately trained to use the model's results responsibly and fairly.
2. Their training should cover:
 - the basics of how machine learning works;
 - the limitations of AI and automated decision-support technologies;
 - the benefits and risks of deploying these systems to assist decision-making, particularly how they help humans come to judgements rather than replacing that judgement; and
 - how to manage cognitive biases, including both decision-automation bias and automation-distrust bias.



Considerations for Child-Centred AI

Those who are selected to implement an AI system must understand child-centred design^[39] if they are to engage with children's data as part of the system's deployment. Having knowledge of child-centred design will help implementers understand why the management of children's data or a system that deals with children's data may require more stringent ethical considerations. If implementers are to directly speak to and work with children, they should go through background checks, such as the [Disclosure and Barring Service \(DBS\)](#), and training on engaging with children to ensure that they are sensitive to their needs and perspectives.

Task 6

Consider How to Build and Present Your Explanation



Related AI Lifecycle Stage:

Model Reporting

1. To build an explanation, you should start by gathering together the information gained when implementing Tasks 1-4. You should review the information and determine how this provides an evidence base for the process-based or outcome-based explanations.
2. You should then revisit the contextual factors to establish which explanation types should be prioritised.
3. How you present your explanation depends on the way you make AI-assisted decisions, and on how people might expect you to deliver explanations you make without using AI.
4. You can 'layer' your explanation by proactively providing individuals first with the explanations you have prioritised and making additional explanations available in further layers. This helps to avoid information (or explanation) overload.
5. You should think of delivering your explanation as a conversation, rather than a one-way process. People should be able to discuss a decision with a competent human being.
6. Providing your explanation at the right time is also important.
7. To increase trust and awareness of your use of AI, you can proactively engage with your stakeholders by making information available about how you use AI systems to help you make decisions.



Considerations for Child-Centred AI

With regard to children's rights and data, based on the information provided in the previous tasks, a short summary should be written to explain your AI-assisted decisions. Graphics, videos, and interactive resources could be made available to support multiple ways of delivering material for developing an understanding of the project and model. Additionally, as much effort as possible should be made to ensure that such explanations are accessible. Clear communication on the project, the model, the information and data, how potential risks have been mitigated, as well as the benefits of the system to an appropriate audience, in this case, children and their parents or guardians, can help limit unexpected outcomes. Where possible, references to children-related policies should be added throughout the AI explanation process to pinpoint where such considerations have been applied.

The Explainability Assurance Management template will help you and your team accomplish the six tasks illustrated previously.



Explainability Assurance Management Template for Project Name

Date completed: Team members involved:

.....

Task 1

Select Priority Explanations by Considering the Domain, Use Case, and Impact on the Individual

Related AI Lifecycle Stage: Design Phase

1. Essential Prioritised Explanations: **Rationale** and **Responsibility**

- | | |
|--|--|
| <p>a. Considering the project domain, use case, and potential impacts outlined in the Stakeholder Engagement Process (SEP) report, what other explanation types will you prioritise?</p> <ul style="list-style-type: none">• The SEP report can be found in the AI Sustainability in Practice Part One workbook. <p>.....</p> | <p>b. Considering the project domain, use case, and potential impacts outlined in the SEP report, what information will explanations require and how comprehensive will this information be?</p> <p>.....</p> <p>c. What other explanation types will be considered for this project?</p> <p>.....</p> |
|--|--|

Checklist for Task 1^[40]

- ☐ We have prioritised rationale and responsibility explanations. We have therefore put in place and documented processes that optimise the end-to-end transparency and accountability of our AI model.
- ☐ We have considered the setting and sector in which our AI model will be used, and how this affects the types of explanation we provide.
- ☐ We have considered the potential impacts of our system, and how these affect the scope and depth of the explanation we provide.

Task 2

Collect and Preprocess Your Data in an Explanation-Aware Manner

Related AI Lifecycle Stages: Data Extraction or Procurement, Data Analysis

1. Consider **Rationale Explanation**

Drawn from the Data Factsheet Template found in the [Responsible Data Stewardship in Practice](#) workbook.

- a. How was data selected and labelled?

.....

2. Consider **Responsibility Explanation**

Drawn from the Governance Workflow Map found in the [AI Accountability in Practice](#) workbook.

- a. Who is responsible at each stage of Data Collection and Preprocessing?

.....

3. Consider **Data Explanation**

Drawn from the Data Factsheet Template found in the [Responsible Data Stewardship in Practice](#) workbook.

- | | |
|--|---|
| <p>a. What is the source of the training data?</p> <p>.....</p> | <p>c. What are the results of assessments about data integrity, quality, and protection and privacy?</p> <p>.....</p> |
| <p>b. How was the data collected?</p> <p>.....</p> | <p>d. What steps were taken to address integrity issues, quality issues, and/or data protection and privacy issues?</p> <p>.....</p> |

4. Consider **Fairness Explanation**

Drawn from the Bias Self-Assessment Template found in the [AI Fairness in Practice](#) workbook.

- | | |
|---|--|
| <p>a. Is the data used in the model representative of those you are making decisions about?</p> <p>.....</p> | <p>b. Are pre-processing techniques, such as re-weighting, required?</p> <p>.....</p> |
|---|--|

5. Consider **Safety Explanation**

Drawn from the Safety Self-Assessment Template found in the [AI Safety in Practice](#) workbook.

- | | |
|---|---|
| <p>a. How have you established reasonable safety objectives?</p> <p>.....</p> | <p>c. How have you ensured that the modelling, testing, and monitoring stages of the system development lead to accurate results?</p> <p>.....</p> |
| <p>b. What are the results of your modelling, testing, and monitoring stages?</p> <p>.....</p> | |

6. Consider **Impact Explanation**

Drawn from the Stakeholder Impact Assessment Template found in [AI Sustainability in Practice Part Two](#) workbook.

- a. How have you implemented the results of your Stakeholder Impact Assessment?
-

Checklist for Task 2^[41]

- ☐ Our data are representative of those we make decisions about, and are reliable, relevant and up-to-date.
- ☐ We have checked with a domain expert to ensure that the data we are using is appropriate and adequate.
- ☐ We know where the data has come from, the purpose it was originally collected for, and how it was collected.
- ☐ Where we are using synthetic data, we know how it was created and what properties it has.
- ☐ We know what the risks are of using the data we have chosen, as well as the risks to data subjects of having their data included.
- ☐ We have labelled the data we are using in our AI system with information including what it is, where it is from, and the reasons why we have included it.
- ☐ Where we are using unstructured or high-dimensional data, we are clear about why we are doing this and the impact of this on explainability.
- ☐ We have ensured as far as possible that the data does not reflect past discrimination, whether based explicitly on protected characteristics or possible proxies.
- ☐ We have mitigated possible bias through pre-processing techniques such as re-weighting, up-weighting, masking, or excluding features and their proxies.
- ☐ It is clear who within our organisation is responsible for data collection and pre-processing.

Task 3

Build Your System to Ensure You Are Able to Extract Relevant Information for a Range of Explanation Types

Related AI Lifecycle Stage: Model Selection & Training

1. Consider **Rationale Explanation**

- a. What are the explanation needs for this project, considering its domain, use case, and potential impacts?
.....
- b. Do the domain, use case, or potential risks encourage using an inherently explainable system?
.....
- c. What are the costs and benefits of using new and potentially less explainable models?
.....
- d. Does the data being used require a more or less explainable system?
.....
- e. Do your data preprocessing needs lead to you to a black box model?
.....
 - If so, are supplementary interpretability tools appropriate in this context?
.....

2. Considering the Above:

- a. What model are you selecting for this project?
.....
- b. If selecting a black box model: What supplementary interpretability tools are you using to help you provide explanations?
.....

c. Does this selection enable you to provide:

- Rationale Explanations?

.....

- Responsibility Explanations?

.....

- Data Explanations? Does the model selected constrain the requirement that any model must be interpretable to ensure individuals' right to be informed?

- The SEP report can be found in the [AI Sustainability in Practice Part One](#) workbook.

.....

- Fairness Explanations?

.....

- Safety Explanations?

.....

- Impact Explanations?

.....

Checklist for Task 3^[42]

We recommend this is filled out by a data scientist

Selecting an Appropriately Explainable model:

- | | |
|--|---|
| <input type="checkbox"/> We know what the interpretability/transparency expectations and requirements are in our sector or domain. | <input type="checkbox"/> Where we are using social or demographic data, we have considered the need to choose a more interpretable model. |
| <input type="checkbox"/> In choosing our AI model, we have taken into account the specific type of application and the impact of the model on decision recipients. | <input type="checkbox"/> Where we are using biophysical data, for example in a healthcare setting, we have weighed the benefits and risks of using opaque or less interpretable models. |
| <input type="checkbox"/> We have considered the costs and benefits of replacing the existing technology we use with an AI system. | |

- ☐ Where we are using a 'black box' system, we have considered the risks and potential impacts of using it.
- ☐ Where we are using a 'black box' system we have also determined that the case we will use it for and our organisational capacity both support the responsible design and implementation of these systems.
- ☐ Where we are using a 'black box' system we have considered which supplementary interpretability tools are appropriate for our use case.
- ☐ Where we are using 'challenger' models^[43] alongside more interpretable models, we have established that we are using them lawfully and responsibly, and we have justified why we are using them.
- ☐ We have considered how to measure the performance of the model and how best to communicate those measures to implementers and decision recipients.
- ☐ We have mitigated any bias we have found in the model and documented these mitigation processes.
- ☐ We have made it clear how the model has been tested, including which parts of the data have been used to train the model, which have been used to test it, and which have formed the holdout data (i.e. test data that is intentionally excluded from the dataset).
- ☐ We have a record of each time the model is updated, how each version has changed, and how this affects the model's outputs.
- ☐ It is clear who within our organisation is responsible for validating the explainability of our AI system.

All the explanation extraction tools we use:

- ☐ Convey the model's results reliably and clearly.
- ☐ Offer affected individuals plausible, accurate, and easily understandable accounts of the logic behind the model's output.
- ☐ Help implementers of AI-assisted decisions to exercise better-informed judgements.

For interpretable AI models:

- ☐ We are confident in our ability to extract easily understandable explanations from models such as regression-based and decision/rule-based systems, Naïve Bayes, and K nearest neighbour.

For supplementary explanation tools to interpret 'black box' AI models:

- ☐ We are confident that they are suitable for our application.
- ☐ In selecting the supplementary tool, we have prioritised the need for it to provide a reliable, accurate and close approximation of the logic behind our AI system's behaviour, for both local and global explanations.
- ☐ We recognise that they will not give us a full picture of the opaque model and have made sure to clearly convey this limitation to implementers and decision recipients.

Combining supplementary explanation tools to produce meaningful information about your AI system's results:

- ☐ We have included a visualisation of how the model works.
- ☐ We have included counterfactual tools to explore alternative possibilities and actionable recourse for individual cases.
- ☐ We have included an explanation of variable importance and interaction effects, both global and local.

Task 4

Translate the Rationale of Your System's Results Into Useable and Easily Understandable Reasons

Related AI Lifecycle Stage: Model Reporting

1. Consider **Rationale Explanation**

- a. Do correlations that the AI model produces make sense in the case you are considering?
.....
 - b. Can recipients of your explanation easily understand how the statistical result has been applied to their particular case?
.....
- Is the explanation accessible and non-technical?
.....
 - Does it avoid jargon?
.....
 - Does it provide a glossary of terms to remove any assumptions regarding definitions?
.....

Checklist for Task 4^[44]

- ☐ We have taken the technical explanation delivered by our AI system and translated this into reasons that can be easily understood by the decision recipient.
- ☐ We have used tools such as text, visual media, graphical representations, summary tables, or a combination, to present information about the logic of the AI system's output.
- ☐ We have justified how we have incorporated the statistical inferences from the AI system into our final decision and rationale explanation.

Task 5

Prepare Implementers to Deploy Your AI System

Related AI Lifecycle Stage: User Training

1. Consider **Rationale Explanation** and **Fairness Explanation**

- a. Have implementers been appropriately trained to use the model's results responsibly and fairly?

.....

- b. Has this training covered:

- The basics of how machine learning works?

.....

- The limitations of AI and automated decision-support technologies?

.....

- The benefits and risks of deploying these systems to assist decision-making?

.....

- How to manage cognitive biases, including both decision-automation bias and automation-distrust bias?

.....

Checklist for Task 5^[45]

Where there is a 'human-in-the-loop' we have trained our implementers to:

- ☐ Understand the associations and correlations that link the input data to the model's prediction or classification.
- ☐ Interpret which correlations are consequential for providing a meaningful explanation by drawing on their domain knowledge or the decision recipient's specific circumstances.
- ☐ Combine the chosen correlations and outcome determinants with what they know of the individual affected to come to their conclusion.
- ☐ Apply the AI model's results to the individual case at hand, rather than uniformly across decision recipients.
- ☐ Recognise situations where decision-automation bias and automation-distrust bias can occur, and mitigate against this.
- ☐ Understand the strengths and limitations of the system.

Task 6

Consider How to Build and Present Your Explanation

Related AI Lifecycle Stage: Model Reporting

1. Initial Considerations

- a. How do you make AI-assisted decisions?
.....
- b. How does the information gathered when implementing Tasks 1-4 provide an evidence base for the process-based or outcome-based explanations?
.....
- c. Considering the contextual factors, which explanation types should be prioritised?
.....
- d. How might people expect you to deliver explanations?
.....

2. Considering the questions above, complete the questions for your prioritised explanations

Explanation type	Will you provide layers of explanations? What explanations have you prioritised and what explanations will be made available in further layers?	Is this explanation made available as a conversation or one-way process?	When will you provide this explanation?
Rationale			
Responsibility			
Data			
Fairness			
Safety			
Impact			

Checklist for Task 6^[46]

- ☐ We have gathered the information collected in Tasks 1-4 and reviewed how these fit within the process-based and outcome-based explanations.
- ☐ We have considered the contextual factors and how this will impact the order in which we deliver the explanation types, and how this will affect our delivery method.
- ☐ We have presented our explanation in a layered way, giving the most relevant explanation type(s) upfront, and providing the other types in additional layers.
- ☐ We have made it clear how decision recipients can contact us if they would like to discuss the AI-assisted decision with a human being.
- ☐ We have provided the decision recipient with the process-based and relevant outcome-based explanation for each explanation type, in advance of making a decision.
- ☐ We have proactively made information about our use of AI available in order to build trust with our customers and stakeholders.

AI Explainability in Practice

Activities



68 [Activities Overview](#)

70 [Interactive Case Study: AI in Children's Social Care](#)

72 [Details About the AI System Under Consideration](#)

73 [Details About the Database](#)

76 [Hypothetical Scenario: The Smith Family](#)

80 [Content Review and Discussion](#)


82 [Information Gathering](#)

84 [Evaluating Explanations](#)

Activities Overview

In the previous sections of this workbook, we have presented an introduction to the core concepts of explainability. In this section we provide concrete tools for applying these concepts in practice. Activities will help participants work towards explaining and justifying AI project processes and AI-supported outcomes to ensure their AI project is sustainable, fair, safe, accountable, and maintain data quality, integrity, and protection.

We offer a collaborative workshop format for team learning and discussion about the concepts and activities presented in the workbook. To run this workshop with your team, you will need to access the resources provided in the link below. This includes a digital board and printable PDFs with case studies and activities to work through.

 [Workshop resources for AI Explainability in Practice](#)

A Note on Activity Case Studies

Case studies within the Activities sections of the AI Ethics and Governance in Practice workbook series offer only basic information to guide reflective and deliberative activities. If activity participants find that they do not have sufficient information to address an issue that arises during deliberation, they should try to come up with something reasonable that fits the context of their case study.

Note for Facilitators

In this section, you will find the participant and facilitator instructions required for delivering activities corresponding to this workbook. Where appropriate, we have included considerations to help you navigate some of the more challenging activities.

Activities presented in this workbook can be combined to put together a capacity-building workshop or serve as stand-alone resources. Each activity corresponds to a section within the Key Concepts in this workbook. Some activities have prerequisites, which are detailed on the following page.

We sometimes provide ideas of how a **co-facilitator** can help manage large groups.



Content Review and Discussion

Review the case study for this workshop.

Corresponding Sections

- [Part One: Introduction to AI Explainability \(page 10\)](#)



Information Gathering

Practise gathering relevant information for building explanations of AI systems.

Corresponding Sections

- [Types of Explanation \(page 28\)](#)
- [Part Two: Putting the Principle of Explainability Into Practice \(page 44\)](#)

Prerequisites

- ↗ [Activity: Content Review and Discussion \(page 80\)](#)



Evaluating Explanations

Practise evaluating the extent to which AI explanations meet their purpose and align with the Maxims of AI Explainability.

Corresponding Sections

- [Part One: Introduction to AI Explainability \(page 10\)](#)
- [Part Two: Putting the Principle of Explainability Into Practice \(page 44\)](#)

Prerequisites

- ↗ [Activity: Content Review and Discussion \(page 80\)](#)
- ↗ [Activity: Information Gathering \(page 82\)](#)

Interactive Case Study: AI in Children's Social Care



Your team is a local authority social care team, which has a statutory duty to safeguard and promote the welfare of children in need within your borough by providing care services. When your team receives a referral and has reasons to be concerned that a child may be suffering, or likely to suffer, significant harm, you are required to undertake an investigation into the child's circumstances.



Following these, some children receive support from your local authority while remaining at home with their families. Others are referred to be placed in the care or supervision of your local authority. Most of these children are moved to foster placements.

As far as is reasonably consistent with your safeguarding duties, however, your team has the duty to promote the upbringing of children by their families by providing an appropriate range and level of services.

The Children's Social Care (CSC) system across England, however, has faced an increase in demand for its services alongside austerity measures, which have limited the resources available to local authorities. **Your team is no exception to this challenge and is considering the use of an AI system that could aid care workers when conducting investigations.**

75%

A stack of three gold coins is shown. A small orange line points from the top coin to a yellow warning triangle icon.

75% of councils are overspending for children's services.^[47]

89%

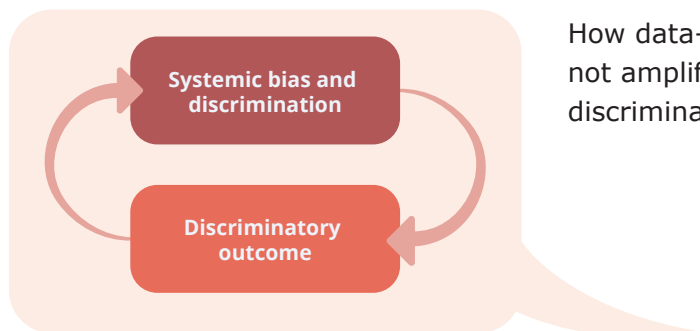
A stack of four orange books is shown. The top book has a small photo of a child on its cover. A small orange line points from the top book to a yellow warning triangle icon.

89% of directors of CSC services reporting in 2016-2017 that they found it increasingly challenging to fulfil their statutory duties to provide support to children in need due to the limited available resources at their disposal!^[48]

The system would predict children's' likelihood of being at risk. It would be used alongside current methods such as conducting interviews with children and families to determine if a child should be taken into care.

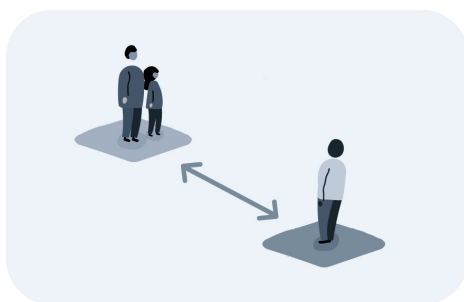
Community Concerns

Although this system could offer your organisation evidence-based insights to ensure that children receive care at the proper time, this high impact context calls for appropriate attention to potential impacts on affected individuals, families, and communities. The community impacted by CSC services has already expressed various concerns about the use of ML in this setting:



How data-driven ML systems are merely reinforcing, if not amplifying, historical patterns of systemic bias and discrimination.

How the mixed results of existing AI systems are signalling widespread conditions of poor data quality and questionable data collection and recording practices.



How the depersonalising and de-socialising effects of trends toward the automation of CSC are harming the care environment and the relationship between social care workers and families.

Details About the AI System Under Consideration

Status of the system:

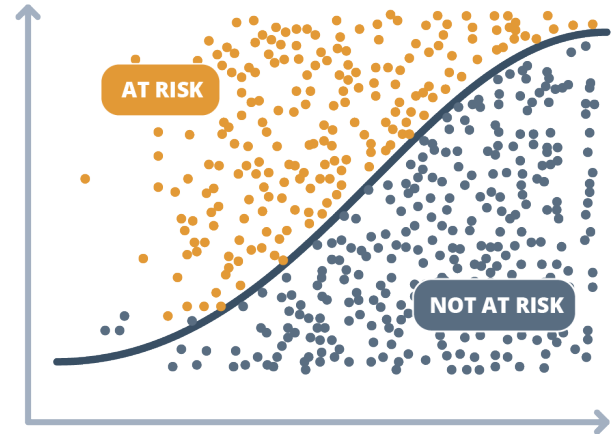
Being trialled

Type of model:

Binary logistic regression

Target variable indicating risk:

The child being taken into care within six months of a referral



What Is a Logistic Regression Model?

It is a supervised ML model that assists with finding a relationship between features and the probability of a particular outcome. It provides a value between 0 and 1 and converts this value into a classification.

The logistic regression model used in this system is binary, meaning that it predicts one of two mutually exclusive classes, in this case, “at risk” (positive class) or “not at risk” (negative class).

This type of model is considered to be highly interpretable, because it is linear and feature importance can be easily isolated—though the transformation of input variables makes the relationship between them and the output a little more challenging to grasp.

Details About the Database

Your team oversees the care of over 100,000 households with dependent children in your borough, and collects data including:

demographic data income levels past referrals employment of parents/guardians
history of past domestic abuse criminal records educational information
past drug treatment current access to public benefits

However:

- There is **no intended purpose for the types of data collected**.
- Data is recorded in **separate, non-centralised databases, and/or spreadsheets**.
- The data available when developing the model was not standardised and there were dispersed amounts of missing data, making a large proportion of the data unfit for use, resulting in **large amounts of unstructured and missing data**.

Data Pre-processing & Feature Engineering

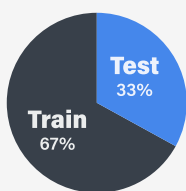
In the Data Pre-processing and Feature Engineering stage, your team replaced missing values with the mean of non-missing cases for that variable. You also transformed the data into a format appropriate for modelling, such as by creating dummy variables for each categorical variable (or feature).

Available variables that are used to predict children at risk in your system trial

Variable	Variable Name	Type
Income	Income	Numeric
Number of Dependents	num_dependents	Numeric
Age PG* 1	age_pg_1	Numeric
Age PG 2	age_pg_2	Numeric
Number of referrals to CSC	num_referrals	Numeric
Past use of CSC services	past_csc_Yes, past_csc_No	Binary
Education PG 1	education_pg_1_Masters, education_pg_1_PhD, education_pg_1_University Degree, education_pg_1_GCSE	Categorical
Education PG 2	education_pg_2_Masters, education_pg_2_PhD, education_pg_2_University Degree, education_pg_2_GCSE	Categorical
Employment PG 1	employment_pg_1_Full-time, employment_pg_1_Part-time, employment_pg_1_Unemployed	Categorical
Employment PG 2	employment_pg_2_Full-time, employment_pg_2_Part-time, employment_pg_2_Unemployed	Categorical
Past domestic abuse	past_domestic_abuse_Yes, past_domestic_abuse_No	Binary
Criminal record either parent	criminal_record_either_parent_Yes, criminal_record_either_parent_No	Binary
Past drug abuse treatment	drug_abuse_treatment_Yes, drug_abuse_treatment_No	Binary
Public benefits	public_benefits_Yes, public_benefits_No	Binary
Child at risk**	child_at_risk	Binary

*PG is parent/guardian – PG1 is the primary caregiver, PG2 is the secondary caregiver (if applicable)

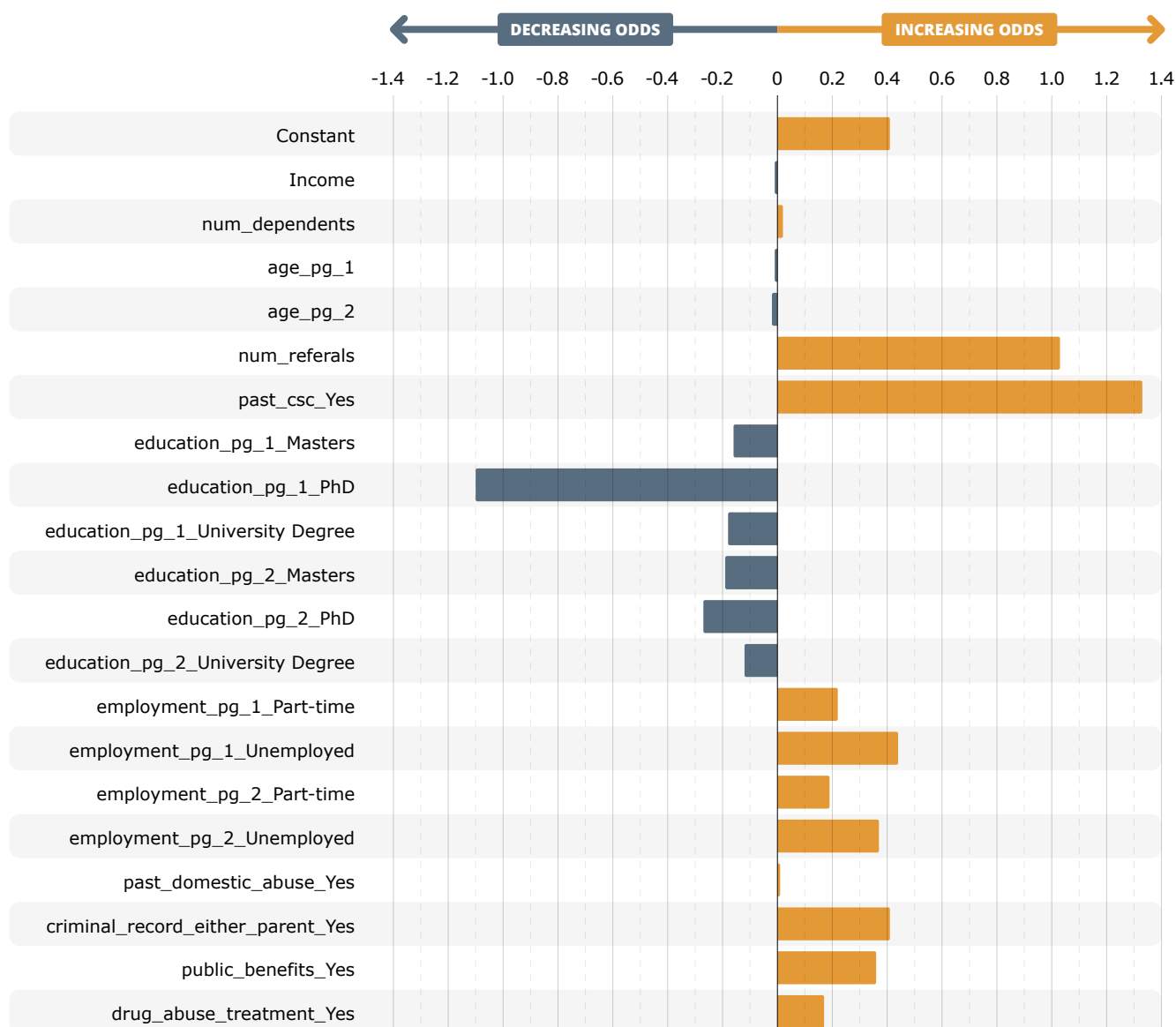
** *Child at risk* is the target variable



67% of the data was used to train the model, **33%** of the data was used to test the model.

The relative feature importance of the trained model, which indicates how each variable contributes to increasing or decreasing the of odds of the model categorising a child as “at risk” is as follows:

Relative Feature Importance of Trained Model



Hypothetical Scenario: The Smith Family

To test the system's end-to-end explainability, your team has gathered the information from a real example in the dataset together with the detailed notes of the care worker that was in charge of the case. To carry out the test, **you are to run a hypothetical scenario of providing an explanation of the outcome to the impacted family.**



The Smith Family

Your team received a call from a neighbour regarding their concern about the safety of the child of the Smith Family. A care worker arrived at their house soon after to assess the situation. The care worker used your model to assist their decision, **which has determined that the Smith child is at risk and should be taken into care.**



The information about the family requesting an explanation of the outcome is as follows:

- It consists of **two parents**, ages 41 and 28, respectively, and one child.
- The highest education level of both parents is GCSE examinations.
- One parent **works part-time**, while the other is **unemployed**.
- Their current **income** level is £11,953.
- The Smith family is receiving **public benefits** at this time, and this is reflected as increasing the probability of risk in the model's prediction.
- There is **no history of past domestic abuse or drug abuse** treatment.
- There is a **criminal record** for one of the parents. This resulted from their participation two decades ago in a political demonstration.
- There are **two past referrals** to CSC. From the case notes and records, it is evident that the parents claim that referrals came from an angry neighbour with whom they had a personal dispute and that their prior involvement with CSC services was not necessary.
- The Smith family was contacted and worked with the CSC on one occasion 5 years ago.



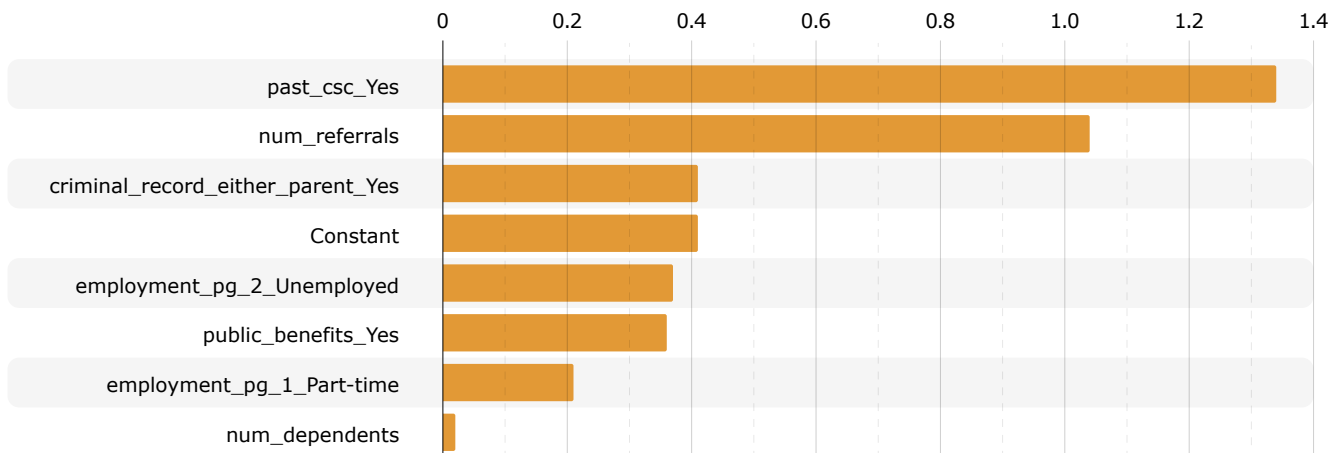
The predicted probability for the Smith child being at risk is 90.36%. With a threshold* of .5, your model has determined that this child is at risk and should be taken into care.

* *The value established to separate the positive class of "at risk" from the negative class of "not at risk".*

The breakdown of inputs and coefficients for the Smith family is as follows:

Variable	Log odds	Odds	Input
Constant	0.4088	1.50501069	-
Income	-0.0002	0.99980002	11,953
num_dependents	0.0240	1.02429032	1
Age_pg_1	-0.0004	0.99960008	41
Age_pg_2	-0.0196	0.98059083	28
num_referrals	1.0389	2.82610659	2
past_csc_Yes	1.3304	3.78255611	1
education_pg_1_Masters	-0.1559	0.85564475	0
education_pg_1_PhD	-1.1049	0.33124400	0
education_pg_1_University Degree	-0.1783	0.83669138	0
education_pg_2_Masters	-0.1914	0.82580220	0
education_pg_2_PhD	-0.2714	0.76231151	0
education_pg_2_University Degree	-0.1245	0.88293826	0
employment_pg_1_Part-time	0.2169	1.24221987	1
employment_pg_1_Unemployed	0.4433	1.55783962	0
employment_pg_2_Part-time	0.1917	1.21130707	0
employment_pg_2_Unemployed	0.3675	1.44411980	1
past_domestic_abuse_Yes	0.0017	1.00170145	0
criminal_record_either_parent_Yes	0.4105	1.50757132	1
public_benefits_Yes	0.3582	1.43075174	1
drug_abuse_treatment_Yes	0.1658	1.18033701	0

The 8 most important features for the decision made about the Smith family are as follows:



The primary model metrics can be found below:

Precision

86%

Number of true positives divided by the number of all cases classified as 'at risk' (**true positives** and **false positives**).

Test Accuracy

86%

Number of correct classifications divided by total number of classifications made.

Recall

84%

Number of true positives divided by the number of all actual cases 'at risk' (**true positives** and **false negatives**).

True Positive

model correctly classifies child as **at risk**

True Negative

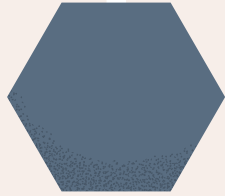
model correctly classifies child as **not at risk**

False Positive

model incorrectly classifies child as **at risk**

False Negative

model incorrectly classifies child as **not at risk**



Your Task

Using the information you have about the context of the use of this model, data available, model choice, results, and metrics used to assess performance, your team must now think how the local authority social care team would provide different types of explanations to the Smith family.





60 mins

Participant Instructions

Content Review and Discussion

Objective

In this activity, your workshop facilitator will give a short presentation about Key Concepts from this workshop. Your team will then review the case study for this workshop.

Team Instructions

1. While your facilitator presents, feel free to read along with the **Key Concepts** section of the board, and make note of any questions you have. After the presentation, feel free to ask any questions.
2. The team will be split into groups. Once in your groups, take a few minutes to read over the case study.
3. Have a group discussion, considering the following questions:
 - What concerns may this AI system raise for AI Explainability?
 - Which explanation types do you think are most important for this case study?
4. After the discussion, reconvene as a team having one volunteer from each group report back with a summary of their group discussion.



60 mins

Facilitator Instructions

Content Review and Discussion

1. Give the team a moment to read over the activity instructions, answering any questions.
2. Using the **Key Concepts** section of the board and the Considerations section of this activity, provide a quick presentation of the Four Principles of AI Explainability and the Explanation Types.
3. Following the presentation, ask the team if they have any questions, using the considerations section of this activity to answer.
4. Next, split the team into two groups.
 - **Facilitators** and **co-facilitators** are to each join one group, using the considerations section of this activity to provide support.
5. In the groups, give participants a few minutes to read over the case study for this workshop, letting them know that they can ask questions at any time.
6. When a few minutes have passed, lead a group discussion, considering the questions on the participants' instructions.
 - As the groups discuss, **facilitators** and **co-facilitators** are to take notes in the Notes section on the board.
7. When the time is up, ask the team to reconvene, having one volunteer from each team share a summary of their group discussion.

Facilitator Considerations

Presentation Preparation

In preparation for your presentation, read over the workbook sections on the [Maxims of AI Explainability](#), as well as the [Types of Explanation](#).

- The subsections of each maxim which cover key aspects of each can be used as reference for the section of the presentation focused on the four maxims.
- The process and outcome-based explanations for each explanation type can be used as reference for the section of the presentation covering the six explanation types.



60 mins

Participant Instructions

Information Gathering

Objective

Practise gathering relevant information for building explanations of AI systems.

Team Instructions

1. In this activity, your team will be split into groups, each focused on gathering relevant information for building an explanation for the Smith family.
 - Each group will be assigned an explanation type.
2. In your groups, go over the **Checklist** on the board pertaining to your assigned explanation.
3. Review each of the checklist items, discussing how these apply to your AI system and how they are relevant to preparing an explanation for the Smith family.
 - Tick the checkbox if your group considers the item to be relevant for the explanation.
 - Then, detail why you think this item is important alongside any information you identify from the case study that is useful to support this item in the **Supporting Information** column.
4. Groups will have 30 minutes to gather information.
5. A volunteer note-taker should write team answers under the **Supporting Information** column, they will also report back to the team in the next activity.
6. When 30 minutes have passed, consider the following questions as a team:
 - Based on the information gathered so far, what could be the contents of a conversation where the care worker is explaining the decision to the Smith family?
 - What information is missing from the checklist items which could be essential to include in an explanation to the Smith family?
7. Reconvene as a team.

Data Explanation	
Checklist Item	Supporting Information <small>Why is it important? What information can support this?</small>
What You May Need to Show	
<input type="checkbox"/> How the data used to train, test, and validate your AI model was managed and utilised from collection through processing and monitoring.	
<input type="checkbox"/> What data you used in a particular decision and how.	
What May Go Into the Explanation	
Process-Based Explanations	

Checklist



60 mins

Facilitator Instructions

Information Gathering

1. Give the team a moment to read over the activity instructions, answering any questions.
2. Next, split the team into groups, each assigned an explanation type.
 - **Group 1:** Rationale Explanation
 - **Group 2:** Data Explanation
 - **Group 3:** Fairness Explanation
3. Let the groups know that they will have 40 minutes to gather relevant information about the system required for building an explanation for the Smith family.
 - Groups are to talk through each of the requirements in their Explanation Checklist on the board, discussing how these apply to the AI system.
 - They are to write their answers in notes placed on the space provided within the checklist.
4. **Facilitators** and **co-facilitators** are to join groups, using the considerations section of this activity to provide support where needed.
5. When 30 minutes have passed, let the groups know that they have 10 minutes to finish gathering their information.
 - At this point groups are to begin discussing how they will summarise their findings to the team.
6. When all 40 minutes have passed, ask the team to reconvene.



45 mins

Participant Instructions

Evaluating Explanations

Objective

Practise evaluating the extent to which AI explanations meet their purpose and align with the Maxims of AI Explainability.

Team Instructions

1. In this activity, volunteer note-takers are to share each group's results from the previous activity.
2. After each volunteer has shared, have a group discussion, considering the following questions:
 - Has enough information been gathered to achieve the purpose of this explanation type? Consider the descriptions on the Key Concepts section of the board.
 - Are there any challenges or concerns that come up when envisioning how this scenario might play out? Consider the extent to which the scenario matches the Maxims of AI Explainability.
3. Next, revisit the questions from the previous activity:
 - What concerns may the AI system in the case study raise for AI Explainability?
 - Which explanation types do you think are most important for this AI system?
4. Your workshop co-facilitator will take notes about your discussion on the board.



45 mins

Facilitator Instructions

Evaluating Explanations

1. Give participants a moment to read over the activity instructions, answering any questions.
2. Next, give volunteer note-takers from the previous activity 5 minutes to report back group findings to the team.
3. After each volunteer shares, lead a 5 minute group discussion about how the explanation applies to the hypothetical scenario of providing explanations to the Smith family, considering the questions on the participant instructions:
 - Has enough information been gathered to achieve the purpose of this explanation type? Consider the descriptions on the Key Concepts section of the board.
 - Are there any challenges or concerns that come up when envisioning how this scenario might play out? Consider the extent to which the scenario matches the [Maxims of AI Explainability](#).
4. When all groups have shared, give the team 10 minutes to revisit the questions on Activity 2:
 - What concerns may the AI system in the case study raise for AI Explainability?
 - Which explanation types do you think are most important for this AI system?
5. As the groups discuss, **facilitators** and **co-facilitators** are to take notes in the Notes section of this activity on the board.

Appendix A: Algorithmic Techniques^[49]

To help you get a better picture of the spectrum of algorithmic techniques, this Appendix lays out some of the basic properties, potential uses, and interpretability characteristics of the most widely used algorithms at present. These techniques are also listed in the table below. We recommend that you work with a data scientist or related expert in considering or applying these techniques.

The 11 techniques listed in the left column are considered to be largely interpretable, although for some of them, like the regression-based and tree-based algorithms, this depends on the number of input features that are being processed. The four techniques in the right column are more or less considered to be 'black box' algorithms.

Broadly Interpretable Systems	Broadly "Black Box" Systems
<ul style="list-style-type: none">• Linear regression (LR)• Logistic regression• Generalised linear model (GLM)• Generalised additive model (GAM)• Regularised regression (LASSO and Ridge)• Rule/decision lists and sets• Decision tree (DT)• Supersparse linear integer model (SLIM)• K-nearest neighbour (KNN)• Naïve Bayes• Case-based reasoning (CBR)/ Prototype and criticism	<ul style="list-style-type: none">• Ensemble methods• Random Forest• Support vector machines (SVM)• Artificial neural net (ANN)

Basic Description	Possible Uses	Interpretability
Linear regression (LR)		
Makes predictions about a target variable by summing weighted input/predictor variables.	Advantageous in highly regulated sectors like finance (e.g. credit scoring) and healthcare (predict disease risk given eg lifestyle and existing health conditions) because it is simpler to calculate and have oversight over.	High level of interpretability because of linearity and monotonicity. Can become less interpretable with increased number of features (i.e. high dimensionality).
Logistic regression		
Extends linear regression to classification problems by using a logistic function to transform outputs to a probability between 0 and 1.	Like linear regression, advantageous in highly regulated and safety-critical sectors, but in use cases that are based in classification problems such as yes/no decisions on risks, credit, or disease.	Good level of interpretability but less so than LR because features are transformed through a logistic function and related to the probabilistic result logarithmically rather than as sums.
Regularised regression (LASSO and Ridge)		
Extends linear regression to classification problems by using a logistic function to transform outputs to a probability between 0 and 1.	Like linear regression, advantageous in highly regulated and safety-critical sectors, but in use cases that are based in classification problems such as yes/no decisions on risks, credit, or disease.	Good level of interpretability but less so than LR because features are transformed through a logistic function and related to the probabilistic result logarithmically rather than as sums.
Generalised linear model (GLM)		
To model relationships between features and target variables that do not follow normal (Gaussian) distributions, a GLM introduces a link function that allows for the extension of LR to non-normal distributions.	This extension of LR is applicable to use cases where target variables have constraints that require the exponential family set of distributions (for instance, if a target variable involves number of people, units of time or probabilities of outcome, the result has to have a non-negative value).	Good level of interpretability that tracks the advantages of LR while also introducing more flexibility. Because of the link function, determining feature importance may be less straightforward than with the additive character of simple LR, and a degree of transparency may be lost.

Basic description	Possible uses	Interpretability
Generalised additive model (GAM)		
To model non-linear relationships between features and target variables (not captured by LR), a GAM sums non-parametric functions of predictor variables (like splines or tree-based fitting) rather than simple weighted features.	This extension of LR is applicable to use cases where the relationship between predictor and response variables is not linear (i.e. where the input-output relationship changes at different rates at different times) but optimal interpretability is desired.	Good level of interpretability because, even in the presence of non-linear relationships, the GAM allows for clear graphical representation of the effects of predictor variables on response variables.
Decision tree (DT)		
A model that uses inductive branching methods to split data into interrelated decision nodes which terminate in classifications or predictions. DT's moves from starting 'root' nodes to terminal 'leaf' nodes, following a logical decision path that is determined by Boolean-like 'if-then' operators that are weighted through training.	Because the step-by-step logic that produces DT outcomes is easily understandable to non-technical users (depending on number of nodes/features), this method may be used in high-stakes and safety-critical decision-support situations that require transparency as well as many other use cases where volume of relevant features is reasonably low.	High level of interpretability if the DT is kept manageably small, so that the logic can be followed end-to-end. The advantage of DT's over LR is that the former can accommodate non-linearity and variable interaction while remaining interpretable.
Rule/decision lists and sets		
Closely related to DT's, rule/decision lists and sets apply series of if-then statements to input features in order to generate predictions. Whereas decision lists are ordered and narrow down the logic behind an output by applying 'else' rules, decision sets keep individual if-then statements unordered and largely independent, while weighting them so that rule voting can occur in generating predictions.	Because the step-by-step logic that produces DT outcomes is easily understandable to non-technical users (depending on number of nodes/ features), this method may be used in high-stakes and safety-critical decision-support situations that require transparency as well as many other use cases where volume of relevant features is reasonably low.	Rule lists and sets have one of the highest degrees of interpretability of all optimally performing and non-opaque algorithmic techniques. However, they also share with DT's the same possibility that degrees of understandability are lost as the rule lists get longer or the rule sets get larger.

Basic description	Possible uses	Interpretability
Case-based reasoning (CBR)/ Prototype and criticism		
Using exemplars drawn from prior human knowledge, CBR predicts cluster labels by learning prototypes and organising input features into subspaces that are representative of the clusters of relevance. This method can be extended to use maximum mean discrepancy (MMD) to identify 'criticisms' or slices of the input space where a model most misrepresents the data. A combination of prototypes and criticisms can then be used to create optimally interpretable models.	CBR is applicable in any domain where experience-based reasoning is used for decision-making. For instance, in medicine, treatments are recommended on a CBR basis when prior successes in like cases point the decision maker towards suggesting that treatment. The extension of CBR to methods of prototype and criticism has meant a better facilitation of understanding of complex data distributions, and an increase in insight, actionability, and interpretability in data mining.	CBR is interpretable-by-design. It uses examples drawn from human knowledge in order to syphon input features into human recognisable representations. It preserves the explainability of the model through both sparse features and familiar prototypes.
Supersparse linear integer model (SLIM)		
SLIM utilises data-driven learning to generate a simple scoring system that only requires users to add, subtract, and multiply a few numbers in order to make a prediction. Because SLIM produces such a sparse and accessible model, it can be implemented quickly and efficiently by non-technical users, who need no special training to deploy the system.	SLIM has been used in medical applications that require quick and streamlined but optimally accurate clinical decision-making. A version called Risk-Calibrated SLIM (RiskSLIM) has been applied to the criminal justice sector to show that its sparse linear methods are as effective for recidivism prediction as some opaque models that are in use.	Because of its sparse and easily understandable character, SLIM offers optimal interpretability for human-centred decision-support. As a manually completed scoring system, it also ensures the active engagement of the interpreter-user, who implements it.

Basic description	Possible uses	Interpretability
Naïve Bayes		
Uses Bayes rule to estimate the probability that a feature belongs to a given class, assuming that features are independent of each other. To classify a feature, the Naïve Bayes classifier computes the posterior probability for the class membership of that feature by multiplying the prior probability of the class with the class conditional probability of the feature.	While this technique is called naïve for reason of the unrealistic assumption of the independence of features, it is known to be very effective. Its quick calculation time and scalability make it good for applications with high dimensional feature spaces. Common applications include spam filtering, recommender systems, and sentiment analysis.	Naïve Bayes classifiers are highly interpretable, because the class membership probability of each feature is computed independently. The assumption that the conditional probabilities of the independent variables are statistically independent, however, is also a weakness, because feature interactions are not considered.
K-nearest neighbour (KNN)		
Used to group data into clusters for purposes of either classification or prediction, this technique identifies a neighbourhood of nearest neighbours around a data point of concern and either finds the mean outcome of them for prediction or the most common class among them for classification.	KNN is a simple, intuitive, versatile technique that has wide applications but works best with smaller datasets. Because it is non-parametric (makes no assumptions about the underlying data distribution), it is effective for non-linear data without losing interpretability. Common applications include recommender systems, image recognition, and customer rating and sorting.	KNN works off the assumption that classes or outcomes can be predicted by looking at the proximity of the data points upon which they depend to data points that yielded similar classes and outcomes. This intuition about the importance of nearness/proximity is the explanation of all KNN results. Such an explanation is more convincing when the feature space remains small, so that similarity between instances remains accessible.
Support vector machines (SVM)		
Uses a special type of mapping function to build a divider between two sets of features in a high dimensional feature space. An SVM therefore sorts two classes by maximising the margin of the decision boundary between them.	SVM's are extremely versatile for complex sorting tasks. They can be used to detect the presence of objects in images (face/no face; cat/no cat), to classify text types (sports article/arts article), and to identify genes of interest in bioinformatics.	Low level of interpretability that depends on the dimensionality of the feature space. In context-determined cases, the use of SVM's should be supplemented by secondary explanation tools.

Basic description	Possible uses	Interpretability
Artificial neural net (ANN)		
Family of non-linear statistical techniques (including recurrent, convolutional, and deep neural nets) that build complex mapping functions to predict or classify data by employing the feedforward—and sometimes feedback—of input variables through trained networks of interconnected and multi-layered operations.	ANN's are best suited to complete a wide range of classification and prediction tasks for high dimensional feature spaces—ie cases where there are very large input vectors. Their uses may range from computer vision, image recognition, sales and weather forecasting, pharmaceutical discovery, and stock prediction to machine translation, disease diagnosis, and fraud detection.	The tendencies towards curviness (extreme non-linearity) and high-dimensionality of input variables produce very low-levels of interpretability in ANN's. They are considered to be the epitome of 'black box' techniques. Where appropriate, the use of ANN's should be supplemented by secondary explanation tools.
Random forest		
Builds a predictive model by combining and averaging the results from multiple (sometimes thousands) of decision trees that are trained on random subsets of shared features and training data.	Random forests are often used to effectively boost the performance of individual decisions trees, to improve their error rates, and to mitigate overfitting. They are very popular in high-dimensional problem areas like genomic medicine and have also been used extensively in computational linguistics, econometrics, and predictive risk modelling.	Very low levels of interpretability may result from the method of training these ensembles of decision trees on bagged data and randomised features, the number of trees in a given forest, and the possibility that individual trees may have hundreds or even thousands of nodes.
Ensemble methods		
As their name suggests, ensemble methods are a diverse class of meta-techniques that combines different 'learner' models (of the same or different type) into one bigger model (predictive or classificatory) in order to decrease the statistical bias, lessen the variance, or improve the performance of any one of the sub-models taken separately.	Ensemble methods have a wide range of applications that tracks the potential uses of their constituent learner models (these may include DT's, KNN's, Random Forests, Naïve Bayes, etc.).	The interpretability of Ensemble Methods varies depending upon what kinds of methods are used. For instance, the rationale of a model that uses bagging techniques, which average together multiple estimates from learners trained on random subsets of data, may be difficult to explain. Explanation needs of these kinds of techniques should be thought through on a case-by-case basis.

Appendix B: Supplementary Models^[50]

In this Appendix, we will provide some useful information about Supplementary AI Explainability Models. Before going into detail about these models, we will provide you with a couple of commonly used distinctions made in the field of explainable AI that will help you and your team to think about what is possible and desirable for an AI explanation. We will also provide you with some technical strategies for explaining 'black box' AI models through supplementary explanation tools.

Local vs global explanation

The distinction between the explanation of single instances of a model's results and an explanation of how it works across all of its outputs is often characterised as the difference between local explanation and global explanation. Both types of explanation offer potentially helpful support for providing significant information about the rationale behind an AI system's output.

A local explanation aims to interpret individual predictions or classifications. This may involve identifying the specific input variables or regions in the input space that had the most influence in generating a particular prediction or classification.

Providing a global explanation entails offering a wide-angled view that captures the inner-workings and logic of that model's behaviour as a whole and across predictions or classifications. This kind of explanation can capture the overall significance of features and variable interactions for model outputs and significant changes in the relationship of predictor and response variables across instances. It can also provide insights into dataset-level and population-level patterns, which are crucial for both big picture and case-focused decision-making.

Internal/ model intrinsic vs. external/ post-hoc explanation

Providing an internal or model intrinsic explanation of an AI model involves making intelligible the way its components and relationships function. It is therefore closely related to, and overlaps to some degree with, global explanation - but it is not the same. An internal explanation makes insights available about the parts and operations of an AI system from the inside. These insights can help your team understand why the trained model does what it does, and how to improve it.

Similarly, when this type of internal explanation is applied to a 'black box model', it can shed light on that opaque model's operation by breaking it down into more understandable, analysable, and digestible parts. For example, in the case of an artificial neural network (ANN), it can break it down into interpretable characteristics of its vectors, features, interactions, layers, parameters etc. This is often referred to as 'peeking into the black box'.

Whereas you can draw internal explanations from both interpretable and opaque AI systems, external or post-hoc explanations are more applicable to 'black box' systems where it is not possible to fully access the internal underlying rationale due to the model's complexity and high dimensionality.

Post-hoc explanations attempt to capture essential attributes of the observable behaviour of a 'black box' system by subjecting it to a number of different techniques that reverse-engineer explanatory insights. Post-hoc approaches can do a number of different things:

- test the sensitivity of the outputs of an opaque model to perturbations in its inputs;
- allow for the interactive probing of its behavioural characteristics; or
- build proxy-based models that utilise simplified interpretable techniques to gain a better understanding of particular instances of its predictions and classifications, or of system behaviour as a whole.

Technical Strategies for Explaining 'Black Box' AI Models Through Supplementary Explanation Tools

If, after considering domain, impact, and technical factors, you have chosen to use a 'black box' AI system, your next step is to incorporate appropriate supplementary explanation tools into building your model.

There is no comprehensive or one-size-fits-all technical solution for making opaque algorithms interpretable. The supplementary explanation strategies available to support interpretability may shed light on significant aspects of a model's global processes and components of its local results.

However, often these strategies operate as imperfect approximations or as simpler surrogate models, which do not fully capture the complexities of the original opaque system. This means that it may be misleading to overly rely on supplementary tools.

With this in mind, 'fidelity' may be a suitable primary goal for your technical 'black box' explanation strategy. In order for your supplementary tool to achieve a high level of fidelity, it should provide a reliable and accurate approximation of the system's behaviour.

For practical purposes, you should think both locally and globally when choosing the supplementary explanation tools that will achieve fidelity.

Thinking locally is a priority, because the primary concern of AI explainability is to make the results of specific data processing activity clear and understandable to affected individuals.

Even so, it is just as important to provide supplementary global explanations of your AI system. Understanding the relationship between your system's component parts (its

features, parameters, and interactions) and its behaviour as a whole will often be a critical to setting up an accurate local explanation. It will also be essential to securing your AI system's fairness, safety and optimal performance. This will help you provide decision recipients with the fairness explanation and safety explanation.

This sort of global understanding may also provide crucial insights into your model's more general potential impacts on individuals and wider society, as well as allow your team to improve the model, so that you can properly address concerns raised by such global insights.

In the following pages we provide you with a table containing details of some of the more widely used supplementary explanation strategies and tools, and we highlight some of their strengths and weaknesses. Keep in mind, though, that this is a rapidly developing field, so remaining up to date with the latest tools will mean that you and technical members of your team need to move beyond the basic information we are offering there. The following pages cover the following supplementary explanation strategies:

Local Supplementary Explanation Strategies	Global Supplementary Explanation Strategies
<ul style="list-style-type: none"> • Individual Conditional Expectations Plot (ICE) • Sensitivity Analysis and Layer-Wise Relevance Propagation (LRP) • Local Interpretable Model-Agnostic Explanation (LIME) and anchors • Shapley Additive ExPlanations (SHAP) • Counterfactual Explanation • Surrogate models (SM) (Could also be used for global explanation) • Self-Explaining and Attention-Based Systems (Could also be used for global explanation) 	<ul style="list-style-type: none"> • Partial Dependence Plot (PDP) • Accumulated Local Effects Plot (ALE) • Global Variable Importance • Global Variable Interaction

Supplementary explanation strategy:
Surrogate models (SM)

What is it and what is it useful for?

SM's build a simpler interpretable model (often a decision tree or rule list) from the dataset and predictions of an opaque system. The purpose of the SM is to provide an understandable proxy of the complex model that estimates that model well, while not having the same degree of opacity. They are good for assisting in processes of model diagnosis and improvement and can help to expose overfitting and bias. They can also represent some non-linearities and interactions that exist in the original model.

Limitations

As approximations, SM's often fail to capture the full extent of non-linear relationships and high-dimensional interactions among features. There is a seemingly unavoidable trade-off between the need for the SM to be sufficiently simple so that it is understandable by humans, and the need for that model to be sufficiently complex so that it can represent the intricacies of how the mapping function of a 'black box' model works as a whole. That said, the R² measurement can provide a good quantitative metric of the accuracy of the SM's approximation of the original complex model.

Global/local? Internal/post-hoc?

For the most part, SM's may be used both globally and locally. As simplified proxies, they are post-hoc.

Supplementary explanation strategy:
Partial Dependence Plot (PDP)

What is it and what is it useful for?

A PDP calculates and graphically represents the marginal effect of one or two input features on the output of an opaque model by probing the dependency relation between the input variable(s) of interest and the predicted outcome across the dataset, while averaging out the effect of all the other features in the model. This is a good visualisation tool, which allows a clear and intuitive representation of the nonlinear behaviour for complex functions (like random forests and SVM's). It is helpful, for instance, in showing that a given model of interest meets monotonicity constraints across the distribution it fits.

Limitations

While PDP's allow for valuable access to non-linear relationships between predictor and response variables, and therefore also for comparisons of model behaviour with domain-informed expectations of reasonable relationships between features and outcomes, they do not account for interactions between the input variables under consideration. They may, in this way, be misleading when certain features of interest are strongly correlated with other model features.

Because PDP's average out marginal effects, they may also be misleading if features have uneven effects on the response function across different subsets of the data—ie where they have different associations with the output at different points. The PDP may flatten out these heterogeneities to the mean.

Global/local? Internal/post-hoc?

PDP's are global post-hoc explainers that can also allow deeper causal understandings of the behaviour of an opaque model through visualisation. These insights are, however, very partial and incomplete both because PDP's are unable to represent feature interactions and heterogenous effects, and because they are unable to graphically represent more than a couple of features at a time (human spatial thinking is limited to a few dimensions, so only two variables in 3D space are easily graspable).

Supplementary explanation strategy:

Individual Conditional Expectations Plot (ICE)

What is it and what is it useful for?

Refining and extending PDP's, ICE plots graph the functional relationship between a single feature and the predicted response for an individual instance. Holding all features constant except the feature of interest, ICE plots represent how, for each observation, a given prediction changes as the values of that feature vary. Significantly, ICE plots therefore disaggregate or break down the averaging of partial feature effects generated in a PDP by showing changes in the feature-output relationship for each specific instance, ie observation-by-observation. This means that it can both detect interactions and account for uneven associations of predictor and response variables.

Limitations

When used in combination with PDP's, ICE plots can provide local information about feature behaviour that enhances the coarser global explanations offered by PDP's. Most importantly, ICE plots are able to detect the interaction effects and heterogeneity in features that remain hidden from PDP's in virtue of the way they compute the partial dependence of outputs on features of interest by averaging out the effect of the other predictor variables. Still, although ICE plots can identify interactions, they are also liable to missing significant correlations between features and become misleading in some instances.

Constructing ICE plots can also become challenging when datasets are very large. In these cases, time-saving approximation techniques such as sampling observation or binning variables can be employed (but, depending on adjustments and size of the dataset, with an unavoidable impact on explanation accuracy).

Global/local? Internal/post-hoc?

ICE plots offer a local and post-hoc form of supplementary explanation.

Supplementary explanation strategy:
Accumulated Local Effects Plots (ALE)

What is it and what is it useful for?

As an alternative approach to PDP's, ALE plots provide a visualisation of the influence of individual features on the predictions of a 'black box' model by averaging the sum of prediction differences for instances of features of interest in localised intervals and then integrating these averaged effects across all of the intervals. By doing this, they are able to graph the accumulated local effects of the features on the response function as a whole. Because ALE plots use local differences in prediction when computing the averaged influence of the feature (instead of its marginal effect as do PDP's), it is able to better account for feature interactions and avoid statistical bias. This ability to estimate and represent feature influence in a correlation-aware manner is an advantage of ALE plots.

ALE plots are also more computationally tractable than PDP's because they are able to use techniques to compute effects in smaller intervals and chunks of observations.

Limitations

A notable limitation of ALE plots has to do with the way that they carve up the data distribution into intervals that are largely chosen by the explanation designer. If there are too many intervals, the prediction differences may become too small and less stably estimate influences. If the intervals are widened too much, the graph will cease to sufficiently represent the complexity of the underlying model.

While ALE plots are good for providing global explanations that account for feature correlations, the strengths of using PDP's in combination with ICE plots should also be considered (especially when there are less interaction effects in the model being explained). All three visualisation techniques shed light on different dimensions of interest in explaining opaque systems, so the appropriateness of employing them should be weighed case-by-case.

Global/local? Internal/post-hoc?

ALE plots are a global and post-hoc form of supplementary explanation.

Supplementary explanation strategy:
Global Variable Importance

What is it and what is it useful for?

The global variable importance strategy calculates the contribution of each input feature to model output across the dataset by permuting the feature of interest and measuring changes in the prediction error: if changing the value of the permuted feature increases the model error, then that feature is considered to be important. Utilising global variable importance to understand the relative influence of features on the performance of the model can provide significant insight into the logic underlying the model's behaviour. This method also provides valuable understanding about non-linearities in the complex model that is being explained.

Limitations

While permuting variables to measure their relative importance, to some extent, accounts for interaction effects, there is still a high degree of imprecision in the method with regard to which variables are interacting and how much these interactions are impacting the performance of the model.

A bigger picture limitation of global variable importance comes from what is known as the 'Rashomon effect'. This refers to the variety of different models that may fit the same data distribution equally well. These models may have very different sets of significant features. Because the permutation-based technique can only provide explanatory insight with regard to a single model's performance, it is unable to address this wider problem of the variety of effective explanation schemes.

Global/local? Internal/post-hoc?

Global variable importance is a form of global and post-hoc explanation.

Supplementary explanation strategy: **Global Variable Interaction**

What is it and what is it useful for?

The global variable interaction strategy computes the importance of variable interactions across the dataset by measuring the variance in the model's prediction when potentially interacting variables are assumed to be independent. This is primarily done by calculating an 'H-statistic' where a no-interaction partial dependence function is subtracted from an observed partial dependence function in order to compute the variance in the prediction. This is a versatile explanation strategy, which has been employed to calculate interaction effects in many types of complex models including ANN's and Random Forests. It can be used to calculate interactions between two or more variables and also between variables and the response function as a whole. It has been effectively used, for example, in biological research to identify interaction effects among genes.

Limitations

While the basic capacity to identify interaction effects in complex models is a positive contribution of global variable interaction as a supplementary explanatory strategy, there are a couple of potential drawbacks to which you may want to pay attention.

First, there is no established metric in this method to determine the quantitative threshold across which measured interactions become significant. The relative significance of interactions is useful information as such, but there is no way to know at which point interactions are strong enough to exercise effects.

Second, the computational burden of this explanation strategy is very high, because interaction effects are being calculated combinatorially across all the data points. This means that as the number of data points increase, the number of necessary computations increase exponentially.

Global/local? Internal/post-hoc?

Global variable interaction is a form of global and post-hoc explanation.

Supplementary explanation strategy:

Sensitivity Analysis and Layer-Wise Relevance Propagation (LRP)

What is it and what is it useful for?

Sensitivity analysis and LRP are supplementary explanation tools used for artificial neural networks. Sensitivity analysis identifies the most relevant features of an input vector by calculating local gradients to determine how a data point has to be moved to change the output label. Here, an output's sensitivity to such changes in input values identifies the most relevant features. LRP is another method to identify feature relevance that is downstream from sensitivity analysis. It uses a strategy of moving backward through the layers of a neural net graph to map patterns of high activation in the nodes and ultimately generates interpretable groupings of salient input variables that can be visually represented in a heat or pixel attribution map.

Limitations

Both sensitivity analysis and LRP identify important variables in the vastly large feature spaces of neural nets. These explanatory techniques find visually informative patterns by mathematically piecing together the values of individual nodes in the network. As a consequence of this piecemeal approach, they offer very little by way of an account of the reasoning or logic behind the results of an ANNs' data processing.

Recently, more and more research has focused on attention-based methods of identifying the higher-order representations that are guiding the mapping functions of these kinds of models as well as on interpretable CBR methods that are integrated into ANN architectures and that analyse images by identifying prototypical parts and combining them into a representational wholes. These newer techniques are showing that some significant progress is being made in uncovering the underlying logic of some ANN's.

Global/local? Internal/post-hoc?

Sensitivity analysis and salience mapping are forms of local and post-hoc explanation, although the recent incorporation of CBR techniques is moving neural net explanations toward a more internal basis of interpretation.

Supplementary explanation strategy:

Local Interpretable Model-Agnostic Explanation (LIME) and anchors

What is it and what is it useful for?

LIME works by fitting an interpretable model to a specific prediction or classification produced by an opaque system. It does this by sampling data points at random around the target prediction or classification and then using them to build a local approximation of the decision boundary that can account for the features which figure prominently in the specific prediction or classification under scrutiny.

LIME does this by generating a simple linear regression model by weighting the values of the data points, which were produced by randomly perturbing the opaque model, according to their proximity to the original prediction or classification. The closest of these values to the instance being explained are weighted the heaviest, so that the supplemental model can produce an explanation of feature importance that is locally faithful to that instance. Note that other interpretive models like decision trees may be used as well.

Limitations

While LIME appears to be a step in the right direction, in its versatility and in the availability of many iterations in very useable software, a host of issues that present challenges to the approach remains unresolved.

For instance, the crucial aspect of how to properly define the proximity measure for the 'neighbourhood' or 'local region' where the explanation applies remains unclear, and small changes in the scale of the chosen measure can lead to greatly diverging explanations. Likewise, the explanation produced by the supplemental linear model can quickly become unreliable, even with small and virtually unnoticeable perturbations of the system it is attempting to approximate. This challenges the basic assumption that there is always some simplified interpretable model that successfully approximates the underlying model reasonably well near any given data point.

LIME's creators have largely acknowledged these shortcomings and have recently offered a new explanatory approach that they call 'anchors'. These 'high precision rules' incorporate into their formal structures 'reasonable patterns' that are operating within the underlying model (such as the implicit linguistic conventions that are at work in a sentiment prediction model), so that they can establish suitable and faithful boundaries of their explanatory coverage of its predictions or classifications.

Global/local? Internal/post-hoc?

LIME offers a local and post-hoc form of supplementary explanation.

Supplementary explanation strategy:
Shapley Additive ExPlanations (SHAP)

What is it and what is it useful for?

SHAP uses concepts from cooperative game theory to define a 'Shapley value' for a feature of concern that provides a measurement of its influence on the underlying model's prediction.

Broadly, this value is calculated by averaging the feature's marginal contribution to every possible prediction for the instance under consideration. The way SHAP computes marginal contributions is by constructing two instances: the first instance includes the feature being measured, while the second leaves it out by substituting a randomly selected stand-in variable for it. After calculating the prediction for each of these instances by plugging their values into the original model, the result of the second is subtracted from that of the first to determine the marginal contribution of the feature. This procedure is then repeated for all possible combinations of features so that the weighted average of all of the marginal contributions of the feature of concern can be computed.

This method then allows SHAP, by extension, to estimate the Shapley values for all input features in the set to produce the complete distribution of the prediction for the instance. While computationally intensive, this means that for the calculation of the specific instance, SHAP can axiomatically guarantee the consistency and accuracy of its reckoning of the marginal effect of the feature. This computational robustness has made SHAP attractive as an explainer for a wide variety of complex models, because it can provide a more comprehensive picture of relative feature influence for a given instance than any other post-hoc explanation tool.

Limitations

Of the several drawbacks of SHAP, the most practical one is that such a procedure is computationally burdensome and becomes intractable beyond a certain threshold.

Note, though, some later SHAP versions do offer methods of approximation such as Kernel SHAP and Shapley Sampling Values to avoid this excessive computational expense. These methods do, however, affect the overall accuracy of the method.

Another significant limitation of SHAP is that its method of sampling values in order to measure marginal variable contributions assumes feature independence (ie that values sampled are not correlated in ways that might significantly affect the output for a particular calculation). As a consequence, the interaction effects engendered by and between the stand-in variables that are used as substitutes for left-out features are necessarily unaccounted for when conditional contributions are approximated. The result is the introduction of uncertainty into the explanation that is produced, because the complexity of multivariate interactions in the underlying model may not be sufficiently captured by the simplicity of this supplemental interpretability technique. This drawback in sampling (as

well as a certain degree of arbitrariness in domain definition) can cause SHAP to become unreliable even with minimal perturbations of the model it is approximating.

There are currently efforts being made to account for feature dependencies in the SHAP calculations. The original creators of the technique have introduced Tree SHAP to, at least partially, include feature interactions. Others have recently introduced extensions of Kernel SHAP.

Global/local? Internal/post-hoc?

SHAP offers a local and post-hoc form of supplementary explanation.

Supplementary explanation strategy: **Counterfactual Explanation**

What is it and what is it useful for?

Counterfactual explanations offer information about how specific factors that influenced an algorithmic decision can be changed so that better alternatives can be realised by the recipient of a particular decision or outcome.

Incorporating counterfactual explanations into a model at its point of delivery allows stakeholders to see what input variables of the model can be modified, so that the outcome could be altered to their benefit. For AI systems that assist decisions about changeable human actions (like loan decisions or credit scoring), incorporating counterfactual explanation into the development and testing phases of model development may allow the incorporation of actionable variables, ie input variables that will afford decision subjects with concise options for making practical changes that would improve their chances of obtaining the desired outcome.

In this way, counterfactual explanatory strategies can be used as way to incorporate reasonableness and the encouragement of agency into the design and implementation of AI systems.

Limitations

While counterfactual explanation offers a useful way to contrastively explore how feature importance may influence an outcome, it has limitations that originate in the variety of possible features that may be included when considering alternative outcomes. In certain cases, the sheer number of potentially significant features that could be at play in counterfactual explanations of a given result can make a clear and direct explanation difficult to obtain and selected sets of possible explanations seem potentially arbitrary.

Moreover, there are as yet limitations on the types of datasets and functions to which these kinds of explanations are applicable.

Finally, because this kind of explanation concedes the opacity of the algorithmic model outright, it is less able to address concerns about potentially harmful feature interactions and questionable covariate relationships that may be buried deep within the model's architecture. It is a good idea to use counterfactual explanations in concert with other supplementary explanation strategies—that is, as one component of a more comprehensive explanation portfolio.

Global/local? Internal/post-hoc?

Counterfactual explanations are a local and post-hoc form of supplementary explanation strategy.

Supplementary explanation strategy:

Self-Explaining and Attention-Based Systems

What is it and what is it useful for?

Self-explaining and attention-based systems actually integrate secondary explanation tools into the opaque systems so that they can offer runtime explanations of their own behaviours. For instance, an image recognition system could have a primary component, like a convolutional neural net, that extracts features from its inputs and classifies them while a secondary component, like a built-in recurrent neural net with an 'attention-directing' mechanism translates the extracted features into a natural language representation that produces a sentence-long explanation of the result to the user.

Research into integrating 'attention-based' interfaces is continuing to advance toward potentially making their implementations more sensitive to user needs, explanation-forward, and humanly understandable. Moreover, the incorporation of domain knowledge and logic-based or convention-based structures into the architectures of complex models are increasingly allowing for better and more user-friendly representations and prototypes to be built into them.

Limitations

Automating explanations through self-explaining systems is a promising approach for applications where users benefit from gaining real-time insights about the rationale of the complex systems they are operating. However, regardless of their practical utility, these kinds of secondary tools will only work as well as the explanatory infrastructure that is actually unpacking their underlying logics. This explanatory layer must remain accessible to human evaluators and be understandable to affected individuals. Self-explaining systems, in other words, should themselves remain optimally interpretable. The task of formulating a primary strategy of supplementary explanation is still part of the process of building out a system with self-explaining capacity.

Another potential pitfall to consider for self-explaining systems is their ability to mislead or to provide false reassurance to users, especially when humanlike qualities are incorporated into their delivery method. This can be avoided by not designing anthropomorphic qualities into their user interface and by making uncertainty and error metrics explicit in the explanation as it is delivered.

Global/local? Internal/post-hoc?

Because self-explaining and attention-based systems are secondary tools that can utilise many different methods of explanation, they may be global or local, internal or post-hoc, or a combination of any of them.

Endnotes

- 1 Burr, C., Fischer, C., and Rincon, C. (2023) Responsible Research and Innovation (Turing Commons Skills Track). Alan Turing Institute. 10.5281/zenodo.7755693.
- 2 Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. <https://doi.org/10.5281/ZENODO.3240529>
- 3 Burr, C., Fischer, C., and Rincon, C. (2023) Responsible Research and Innovation (Turing Commons Skills Track). Alan Turing Institute. 10.5281/zenodo.7755693.
- 4 Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. <https://doi.org/10.5281/ZENODO.3240529>
- 5 Burr, C., Fischer, C., and Rincon, C. (2023) Responsible Research and Innovation (Turing Commons Skills Track). Alan Turing Institute. 10.5281/zenodo.7755693.
- 6 Knack, A., Carter, R. J., and A. Babuta (2022, December). Human-Machine Teaming in Intelligence Analysis. Requirements for developing trust in machine learning systems. Centre for Emerging Technology and Security. https://cetas.turing.ac.uk/sites/default/files/2022-12/cetas_research_report_-_hmt_and_intelligence_analysis_vfinal.pdf
- 7 Burt, Andrew (2019, December 13). The AI Transparency Paradox. Harvard Business Review. <https://hbr.org/2019/12/the-ai-transparency-paradox>
- 8 Burt, Andrew (2019, December 13). The AI Transparency Paradox. Harvard Business Review. <https://hbr.org/2019/12/the-ai-transparency-paradox>
- 9 5Rights Foundation (2019). Demystifying the Age Appropriate Design Code. <https://5rightsfoundation.com/uploads/demystifying-the-age-appropriate-design-code.pdf>
- 10 ICO. (2020). Age Appropriate Design: A Code of Practice for Online Services. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/ageappropriate-design-a-code-of-practice-for-online-services/>
- 11 UNICEF. (2021). Policy guidance on AI for children 2.0. UNICEF. <https://www.unicef.org/innocenti/reports/policy-guidance-ai-children>
- 12 Leslie, D. (2020). Explaining Decisions Made with AI. The Alan Turing Institute and ICO. <http://dx.doi.org/10.2139/ssrn.4033308>
- 13 The Alan Turing Institute and ICO (2022). Explaining Decisions Made with AI. The Alan Turing Institute and ICO. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/>
- 14 Information Commissioner's Office. (2021). Guide to the General Data Protection Regulation (GDPR). <https://ico.org.uk/for-organisations/guide-to-data-protection>

- 15 UNICEF. (2021). Policy guidance on AI for children 2.0. UNICEF. <https://www.unicef.org/innocenti/reports/policy-guidance-ai-children>
- 16 Information Commissioner's Office. (2021). Age appropriate design: a code of practice for online services. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/>
- 17 Information Commissioner's Office. (2021). Guide to the General Data Protection Regulation (GDPR). <https://ico.org.uk/for-organisations/guide-to-data-protection/>
- 18 Leslie, D. (2020). Explaining Decisions Made with AI. The Alan Turing Institute and ICO. <http://dx.doi.org/10.2139/ssrn.4033308>
- 19 Equality Act 2010, c. 5. <https://www.legislation.gov.uk/ukpga/2010/15/part/2/chapter/2/crossheading/adjustments-for-disabled-persons>
- 20 Equality and Human Rights Commission. (2014). The Essential Guide to the Public Sector Equality Duty: England and Non-Devolved Public Authorities in Scotland and Wales. <https://www.equalityhumanrights.com/guidance/public-sector-equality-duty-psed>
- 21 UNICEF. (2021). Policy guidance on AI for children 2.0. UNICEF. <https://www.unicef.org/innocenti/reports/policy-guidance-ai-children>
- 22 Leslie, D. (2020). Explaining Decisions Made with AI. The Alan Turing Institute and ICO. <http://dx.doi.org/10.2139/ssrn.4033308>
- 23 UNICEF. (2021). Policy guidance on AI for children 2.0. UNICEF. <https://www.unicef.org/innocenti/reports/policy-guidance-ai-children>
- 24 UNICEF. (2021). Policy guidance on AI for children 2.0. UNICEF. <https://www.unicef.org/innocenti/reports/policy-guidance-ai-children>
- 25 Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 26, 2051–2068. <https://doi.org/10.1007/s11948-019-00146-8>
- 26 Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146, 102551. <https://doi.org/10.1016/j.ijhcs.2020.102551>
- 27 Burr, C., & Leslie, D. (2022). Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics*, 1-26. <https://doi.org/10.1007/s43681-022-00178-0>
- 28 Burr, C., & Leslie, D. (2022). Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics*, 1-26. <https://doi.org/10.1007/s43681-022-00178-0>
- 29 Königstorfer, F., & Thalmann, S. (2022). AI Documentation: A path to accountability. *Journal of Responsible Technology*, 11, 100043. <https://doi.org/10.1016/j.jrt.2022.100043>
- 30 Leslie, D., Rincón, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, S.J. Burr, C., Aitken, M., Katell, M., Fischer, C., Wong, J., and Kherroubi Garcia, I. (2023). *AI Sustainability in Practice. Part One: Foundations for Sustainable AI Projects*. The Alan Turing Institute.
- 31 UNICEF. (2021). Policy guidance on AI for children 2.0. UNICEF. <https://www.unicef.org/innocenti/reports/policy-guidance-ai-children>

- 32 Information Commissioner's Office. (2021). Age appropriate design: a code of practice for online services. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/>
- 33 Leslie, D., Burr, C., Aitken, M., Katell, M., Briggs, M., Rincon, C. (2021) Human rights, democracy, and the rule of law assurance framework: A proposal. The Alan Turing Institute. <https://doi.org/10.5281/zenodo.5981676>
- 34 IBM (n.d.). *What is explainable AI?* <https://www.ibm.com/topics/explainable-ai>
- 35 Information Commissioner's Office. (2021). Age appropriate design: a code of practice for online services. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/>
- 36 UNICEF. (2021). Policy guidance on AI for children 2.0. UNICEF. <https://www.unicef.org/innocenti/reports/policy-guidance-ai-children>
- 37 Information Commissioner's Office. (2021). Age appropriate design: a code of practice for online services. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/>
- 38 Benchaji, I., Douzi, S., El Ouahidi, B., & Jaafari, J. (2021). Enhanced credit card fraud detection based on attention mechanism and LSTM deep model. *Journal of Big Data*, 8, 1-21. <https://doi.org/10.1186/s40537-021-00541-8>
- 39 Save the Children Finland (2020). Child-centered design. <https://resourcecentre.savethechildren.net/document/child-centered-design/>
- 40 Leslie, D. (2020). Explaining Decisions Made with AI. The Alan Turing Institute and ICO. <http://dx.doi.org/10.2139/ssrn.4033308>
- 41 The Alan Turing Institute and ICO (2022, October). Explaining Decisions Made with AI. The Alan Turing Institute and ICO. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/>
- 42 Leslie, D. (2020). Explaining Decisions Made with AI. The Alan Turing Institute and ICO. <http://dx.doi.org/10.2139/ssrn.4033308>
- 43 A challenger model is a model developed or trained to benchmark the performance of an existing model or set of models, "explore alternative modelling assumptions and identify patterns that may not be captured by traditional models". Source: Deloitte (2023). The application of machine learning and challenger models in IRB Credit Risk modelling. Retrieved from <https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/risk/deloitte-nl-risk-challenger-models-model-estimation.pdf>
- 44 Leslie, D. (2020). Explaining Decisions Made with AI. The Alan Turing Institute and ICO. <http://dx.doi.org/10.2139/ssrn.4033308>
- 45 Leslie, D. (2020). Explaining Decisions Made with AI. The Alan Turing Institute and ICO. <http://dx.doi.org/10.2139/ssrn.4033308>
- 46 Leslie, D. (2020). Explaining Decisions Made with AI. The Alan Turing Institute and ICO. <http://dx.doi.org/10.2139/ssrn.4033308>

- 47 Local Government Association (2017). *LGA budget submission: Autumn 2017*. Retrieved from <https://www.local.gov.uk/parliament/briefings-and-responses/lga-autumn-budget-submission-2017>

- 48 All Party Parliamentary Group for Children (2017). *No good options: Report of the inquiry into children's social care in England*. Retrieved from <https://www.ncb.org.uk/sites/default/files/uploads/No%20Good%20Options%20Report%20final.pdf>

- 49 The Alan Turing Institute and ICO (2022). Explaining Decisions Made with AI. The Alan Turing Institute and ICO. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/>

- 50 The Alan Turing Institute and ICO (2022). Explaining Decisions Made with AI. The Alan Turing Institute and ICO. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/>

To find out more about the AI Ethics and Governance in Practice Programme please visit:

aiethics.turing.ac.uk

Version 1.2

This work is licensed under the terms of the Creative Commons Attribution License 4.0 which permits unrestricted use, provided the original author and source are credited. The license is available at:

<https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>

**The
Alan Turing
Institute**