Project explAIn

Interim report

ico.

Information Commissioner's Office

Contents

Executive summary

Project Expl*AI*n is a collaboration between the Information Commissioner's Office (ICO) and The Alan Turing Institute (The Turing) to create practical guidance to assist organisations with explaining artificial intelligence (AI) decisions to the individuals affected.

As part of this project, the ICO and The Turing conducted public and industry engagement research. This helped us understand different points of view on this complex topic.

This report sets out the methodology and findings of this research. Key findings are:

- the relevance of context for the importance, purpose and expectations of explanations;

- the need for improved education and awareness around the use of AI for decision-making; and

- challenges to deploying explainable AI such as cost and the pace of innovation.

The possible interpretations of these findings and their implications for the development of the guidance are discussed, including:

- the lack of a one-size-fits-all approach to explanations, including the potential for a list of explanation types to support organisations in making appropriate choices;

- the need for board-level buy-in on explaining AI decisions; and

- the value of a standardised approach to internal accountability to help assign responsibility for explainable AI decision-systems and foster an organisational culture of responsible innovation.

We acknowledge the limitations of the research, and a conclusion summarises the findings, setting out their value to the project and beyond.

The report ends with next steps for the project, including a summary of the planned guidance.

The ICO and The Turing gratefully acknowledge the support and input given to this project by Citizens' Juries c.i.c., the Jefferson Center, the Greater Manchester Patient Safety Translational Research Centre, techUK,

and all the industry representatives and members of the public that took part in our engagement research.

Introduction

Background

In October 2017, Professor Dame Wendy Hall and Jérôme Pesenti published their independent review on growing the AI industry in the UK[1]. The second of the report's recommendations to support uptake of AI was for the ICO and The Turing to:

> "...develop a framework for explaining processes, services and decisions delivered by AI, to improve transparency and accountability."[2]

In April 2018, the Government published its AI Sector Deal[3]. The deal tasked the ICO and The Turing to:

> "...work together to develop guidance to assist in explaining AI decisions."[4]

The independent report and the Sector Deal are part of ongoing efforts made by national and international regulators and governments to address the wider implications of transparency and fairness in AI decisions impacting individuals, organisations, and wider society.

Why is the ICO working on this?

As the UK regulator for the General Data Protection Regulation (GDPR), the ICO understands that, while innovative and data-driven technologies create enormous opportunities, they also present some of the biggest risks related to the use of personal data. We also recognise the need for effective guidance for organisations seeking to address data protection issues arising from the use of these technologies.

In particular, AI is a key priority area in the ICO's Technology Strategy.[5] Project Expl*AI*n is an opportunity to address this priority area and achieve the second goal of our Technology Strategy:

> "To provide effective guidance to organisations about how to address data protection risks arising from technology."

---

[1] Available <https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk>
[2] ibid Recommendation 2
[3] Available https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal>
[4] ibid
[5] https://ico.org.uk/media/about-the-ico/documents/2258299/ico-technology-strategy-2018-2021.pdf

Explaining AI decisions is one part of the ICO's work on exploring the data protection implications of AI.

In 2017, the ICO published a paper titled "Big data, artificial intelligence, machine learning and data protection"[6]. This discussed the data protection and societal impacts of these technologies and provided guidance on these issues.

Additionally, Dr Reuben Binns, the ICO's Postdoctoral Research Fellow in AI, is leading the development of an AI auditing framework, due to be finalised in 2020. The ICO will use the framework to assess the data protection compliance of organisations using AI. The framework will also inform guidance for organisations on the management of data protection risks arising from AI applications.

Why is The Alan Turing Institute working on this?

As the national institute for data science and AI, the mission of The Turing is to use these technologies to change the world for the better. The Turing's public policy programme, working in close collaboration with public authorities, supports this mission by bringing together technical, legal, and ethical expertise to ensure that data science and AI serve the public good.

To achieve this, AI systems must be designed to operate ethically, fairly, and safely. This requires great strides to be made in the interpretability of these systems. This is why the public policy programme has partnered with the ICO on Project Expl*AI*n.

This joint effort to assess how to effectively explain AI decisions couldn't be more timely or critical, given that AI systems are currently in use across the UK. This makes the need for the responsible design and implementation of explainable AI systems all the more urgent.

'Explainability' is an essential ingredient for the responsible development of AI and machine learning technologies. It allows the developers and implementers of these technologies to be better informed about their operations and outcomes. It also affords important safeguards to individuals subject to AI decisions, and it enables society as a whole to gain greater understanding about the benefits and drawbacks of AI systems.

As part of the project, The Turing has drawn on the multidisciplinary expertise of its academics from across the domains of digital ethics, public policy, data science and artificial intelligence. The goal of our contribution

---

[6] https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

to this valuable collaborative work is to deliver results that will encourage well-informed and ethical decisions to be made about the application of these powerful technologies in real world scenarios.

What is an AI decision?

AI is an umbrella term for a range of technologies and methodologies seeking to simulate human intelligence in attempting to solve complex tasks, such as decision-making.

AI decisions are often based on the outputs of machine learning models, trained on data to generate predictions, recommendations or classifications - for example, whether to grant a customer a loan or invite an applicant to an interview.

AI can be used for decision support (to inform the thinking of a human decision-maker) or it can be used to automate the generation and delivery of a decision without any human involvement.

What does the GDPR say about AI?

The GDPR is technology neutral, so it does not directly reference AI. However, it has a significant focus on large scale automated processing of personal data, specifically addressing the use of automated decision-making. As such, several provisions are highly relevant to the use of AI for decision-making:

- Principle 1. (a) requires personal data processing to be fair, lawful, and transparent.

- Articles 13-15 give individuals the right to be informed of the existence of solely automated decision-making, meaningful information about the logic involved, and the significance and envisaged consequences for the individual.

- Article 22 gives individuals the right not to be subject to a solely automated decision producing legal or similarly significant effects.

- Article 22(3) obliges organisations to adopt suitable measures to safeguard individuals when using solely automated decisions, including the right to obtain human intervention, to express his or her view, and to contest the decision.

- Recital 71 provides interpretative guidance of Article 22. It says individuals should have the right to obtain an explanation of a solely automated decision after it has been made.

- Article 35 requires organisations to carry out Data Protection Impact Assessments (DPIAs) when what they are doing with personal data, particularly when using new technologies, is likely to have high risks for individuals.

The objective of Project Expl*AI*n is to produce useful guidance that assists organisations with meeting the expectations of individuals when delivering explanations of AI decisions about them. This will support organisations to comply with the legal requirements above, but the guidance will go beyond this. It will promote best practice, helping organisations to foster individuals' trust, understanding, and confidence in AI decisions.

Methodology

Given the complex nature of AI decisions, and the sparsity of research on explanations in this context, it was necessary to carry out primary evidence-based research to help inform the development of the guidance.

We undertook two strands of research: public engagement and industry engagement. This allowed views to be gathered from a range of stakeholders with various (and sometimes competing) interests in explaining AI decisions.

The methodologies used for these research activities are detailed under the headings below.

Public engagement research

Choosing a research method

Understanding public opinion is vital to develop guidance that is effective at supporting organisations to meet the expectations of individuals when explaining AI decisions.

Public engagement activities, such as surveys and focus groups, are an efficient way to quickly gauge public opinion on a given topic. However, for this project, they were considered less useful due to the complex nature of AI decisions and the need for an in-depth exploration and discussion of the issues. As such, the 'citizens' jury' research methodology was chosen as an appropriate approach to public engagement that met these requirements.

Citizens' juries (juries) are the creation of the Jefferson Center's Ned Crosby. In 1971 he founded the concept with the belief that groups of everyday citizens can provide unique insight into tackling complex issues.[7] This is achieved by allowing 'jurors' time to learn about, discuss, and come to conclusions on complex social issues.

Setting up the juries

The ICO co-commissioned two juries with the National Institute for Health Research (NIHR) Greater Manchester Patient Safety Translational Research Centre (GM PSTRC), hosted by the University of Manchester. The GM PSTRC were conducting their own research into public perceptions on the use of AI in healthcare. The juries were an opportunity for collaborative public engagement that was mutually beneficial for the ICO's and the GM PSTRC's respective research projects.

---

[7] https://jefferson-center.org/about-us/how-we-work/

Citizens' Juries c.i.c., specialists in the design and facilitation of citizens' juries, organized the juries in collaboration with the Jefferson Center.[8]

Two juries were undertaken, one in Coventry and one in Manchester. Both juries followed the same design and process, the only difference being the location and the jurors. Jurors were made up of a cross-section of the population, representing the demographic breakdown of England as per the 2011 Census. Juror selection took into account gender, age, ethnicity, and educational attainment. In total, 36 individuals were selected to be jurors, 18 for Manchester, and 18 for Coventry.

Prior to selection, jurors were asked the following question, taken from a national survey commissioned by the Royal Society for Arts (RSA):

> **Q: How comfortable, if at all, are you with the following idea? As the accuracy and consistency of automated systems improve over time, more decisions can be fully automated without human intervention required.**
>
> a) Very comfortable
> b) Fairly comfortable
> c) Not very comfortable
> d) Not at all comfortable
> e) Don't know

Responses were compared against the results of the RSA survey to ensure jurors had a range of views on AI, matching those reported in the RSA survey.

Jury overview

Over five days in February 2019, each jury heard evidence from several expert witnesses. Jurors took part in exercises and discussions on AI decisions and explanations. Giving consideration to the importance of an explanation of an AI decision, when this is at the cost of the system's overall accuracy, jurors came to conclusions on what to prioritise and why.

Some experts argue that a trade-off between the 'explainability' and the accuracy of AI decisions is artificial. This was acknowledged in the design of the juries. However, it was necessary to use a dichotomy like this, giving jurors a clear choice between two competing priorities in order to understand how much importance they attached to explanations, and the reasons for this.

---

[8] https://citizensjuries.org

It was important to see if and how these priorities and reasons changed in different settings. To test this, the ICO and the GM PSTRC constructed hypothetical scenarios involving the use of AI decisions in various public and private sector contexts.

<u>Jury process</u>

On the first two days of each jury, jurors heard from four expert witnesses:

- Prof. Sonia Olhede talked about the trade-off between explanations and accuracy of AI decisions;

- Rhiannon Webster discussed the law concerning AI decisions, in particular data protection;

- Dr. Andre Freitas made the case for prioritising accuracy; and

- Prof. Alan Winfield made the case for prioritising explanations.

On the third and fourth days, jurors considered four scenarios involving the use of AI decisions in different contexts:

- Scenario 1 – Healthcare: stroke diagnosis in the NHS.

- Scenario 2 – Recruitment: candidate shortlisting in a private company.

- Scenario 3 – Healthcare: kidney transplant matching in the NHS.

- Scenario 4 – Criminal justice: offender selection for a rehabilitation programme (as an alternative to prosecution) in a UK police force.

The organisations in each scenario could choose between three hypothetical AI decision systems:

- System A – 75% accurate / full explanation.

- System B – 85% accurate / partial explanation.

- System C – 95% accurate / no explanation.

A fifth expert witness, Dr Allan Tucker, talked to jurors about the kind of explanation that the systems might give for each scenario. Jurors also watched video interviews with four scenario witnesses (people whose

work related to each scenario) discussing how much the accuracy of, and explanations of, decisions might matter in each case.

Jurors were asked the same three questions about each scenario, including which system should be chosen, and whether and why explanations were important. On the fifth day, jurors considered and answered three general questions about AI decisions, including whether human and AI decisions should require similar explanations, and what else can be done to build confidence in AI decisions.

Juror deliberations and responses to questions were captured on a digital survey tool and by jury facilitators.

Further details of the jury design and all jury materials (including expert witness presentations, full scenario descriptions and questions posed to jurors) are available on the GM PSTRC website[9].

Industry engagement research

Choosing a research method

Discussions with practitioners – the people and organisations developing, procuring and deploying AI decision systems – are a key part of the project. It is important that the planned guidance reflects actual applications of AI, addressing real challenges, for it to be of practical use for industry.

Roundtable discussions were chosen as the research method for engagement with industry. With well-selected participants and appropriate moderation, this approach provides an informal environment for honest, open and frank debate on complex or contentious topics.

While digital surveys or email consultations were considered as alternatives, it was determined that these approaches would not elicit the depth and quality of response that the roundtables could deliver.

Roundtables overview

Three roundtables were convened, each made up of a distinct set of people holding key organisational roles relating to AI decisions across the public, private and third sectors:

- Roundtable 1 – Data scientists and researchers.

- Roundtable 2 – Chief Data Officers (CDOs) and C-suite executives.

---

[9] http://www.patientsafety.manchester.ac.uk/research/themes/safety-informatics/citizens-juries/

- Roundtable 3 – Data Protection Officers (DPOs), lawyers, and consultants.

The roundtables were purposefully split according to role (technical, senior, and compliance) to allow insights unique to each discipline to emerge, reflecting the varying opinions and challenges of different industry stakeholders involved in the use of AI in decision-making.

The roundtables were held at The Alan Turing Institute in London over consecutive days in March 2019. Participants discussed their approach to explaining AI decisions, their reactions to the findings from the juries, and their thoughts on the planned guidance.

Roundtables process

Each roundtable was moderated by a chair with expertise relevant to its composition:

- David Leslie, Ethics Fellow at The Alan Turing Institute chaired roundtable 1.

- Sue Daley, Associate Director of Technology and Innovation at techUK chaired roundtable 2.

- Carl Wiper, Group Manager in the Innovation department at the ICO chaired roundtable 3.

Chairs used questions specific to each roundtable to instigate conversation, and where appropriate, to probe for further detail.

Discussion at each roundtable was structured around the same three topics:

- *AI in decision-making*

  The technical, organisational and compliance approaches (and barriers) to developing, procuring and deploying explainable AI decision-systems.

- *Citizens' juries*

  Reactions to juror insights, the gap (if at all) between juror expectations and technical / organisational feasibility, and steps to address this.

- *Planned guidance*

Reactions to a high-level outline of the planned guidance, including any useful or missing elements.

Participant discussions were captured by note-takers from the ICO.

Findings

Three key themes relating to explaining AI decisions emerged from the research:

1. the importance of context;

2. the need for education and awareness; and

3. the challenges in providing explanations.

The results informing these findings, from the public and industry engagement activities, are summarised under the headings below. For a full breakdown of juror responses, see the Citizens' Juries Report published by Citizens' Juries c.i.c. on the GM PSTRC website[10].

1. Context

Public engagement results

Importance of explanation

Quantitative results indicated that the context in which an AI decision is made affects the importance of receiving an explanation.

In the healthcare scenarios (1 and 3), most jurors felt it was less important to receive an explanation and opted for AI decision-system C (95% accurate / no explanation).

However, in the recruitment and criminal justice scenarios (2 and 4), jurors mostly thought explanations were important and tended to opt for AI decision-system A (75% accurate / full explanation) or B (85% accurate / partial explanation).

Juror responses to the general questions posed at the end of the juries supported the above. Most jurors felt that the relative importance of explanations and accuracy varied by context; not a single juror concluded that an explanation should be provided in all contexts.

Purpose of explanation

Qualitative results suggested context is also important in determining the reasons why an explanation is (or is not) important and the purpose for which it is used.

---

[10] http://www.patientsafety.manchester.ac.uk/research/themes/safety-informatics/citizens-juries/

In the healthcare scenarios (1 and 3), jurors tended to prioritise accuracy of decisions due to the need for a quick and precise diagnosis or match. They were concerned with "...fixing the problem..." and thought that "...alternative explanations..." could be provided later. Jurors also thought that an explanation may be of limited use because the scientific nature of the data used meant that "...the factors are not changeable".

In the recruitment and criminal justice scenarios (2 and 4), jurors instead prioritised explanations, referencing purposes they can be used for in these settings. Purposes included:

- challenging a decision – "...correct reason for rejection.";

- changing behaviour – "...feedback in order to allow them to improve."; and

- building trust / ensuring equity – "...to prove there is no bias."

Expectation of explanation

Quantitative results from the general jury questions indicated that context is also key for jurors as regards their expectations of receiving an explanation.

Most jurors placed less importance on explanations in contexts where they would not usually expect a human decision-maker to provide an explanation. Only one juror thought that AI decisions should be explained in contexts where a human would not normally explain their decision.

In contexts where humans would usually provide an explanation, most jurors indicated that explanations of AI decisions should be similar to human explanations. Jurors felt this was important to help build trust and to ensure explanations were understandable.

Industry engagement results

Importance of explanation

Participants from all three roundtables raised the relevance of context when considering how to approach explaining AI decisions. While this generally corresponds with findings from the juries, the roundtables tended to approach context from a different perspective, identifying other factors affecting explanations.

It was suggested that the contextual consideration of 'agent' (person) is key. Participants felt the importance of an explanation of an AI decision is likely to vary depending on the person it is given to. For instance, in a healthcare setting, it may be more important for a healthcare professional to receive an explanation of a decision, than for the patient, given their expertise and authority in this context.

Participants also discussed the importance of context in relation to the unique operations of the organisations deploying AI decision-systems. It was suggested that the planned guidance should be flexible enough to reflect differences in sectors, size of organisation, and business model (eg B2B and B2C).

Roundtable reactions to jury findings on the varying importance of explanations in different scenarios generally indicated that participants considered juror requirements as relatively modest and achievable, especially in healthcare settings.

Purpose of explanation

As with the jurors, roundtable participants recognised the value of explanations in allowing a decision to be challenged and enabling individuals to learn and change their behaviour in contexts where these purposes can be served (such as the recruitment and criminal justice scenarios presented to jurors). Participants from the data scientist roundtable noted that to support these purposes it is important for explanations to adequately and truthfully justify the decision (as opposed to explanations that merely list the causes for a decision or misrepresent the rationale to appease an individual).

Roundtable participants agreed that explanations can be used to build trust and detect bias but did not indicate that this explanation purpose was context-specific. Rather, it was highlighted that using explanations to identify and address underlying system bias was a key consideration across the board.

Further to the idea that explanations may vary by agent, some roundtable participants argued that there are contextual differences even for a single type of agent. For instance, explanation purposes may differ according to the level of expertise of an individual subject to, or interpreting and applying, the decision. It was proposed therefore that a hierarchy of explanations could allow individuals to choose the detail most useful to them.

Expectation of explanation

Some roundtable participants agreed with jurors that explanations of AI decisions should largely reflect the way in which human decision-makers provide explanations. It was suggested that there may be an unwarranted attitude of exceptionalism around expectations of explanations of AI decisions, ie that they should provide a greater level of detail than those explanations which are generally accepted in the case of human decisions.

Others however argued that AI decisions should indeed be held to higher standards. It was argued that humans are prone to delivering explanations that are socially beneficial for the explainer, and do not accurately represent the truth of the decision-making process. They thought it may therefore be preferable for the content of explanations, and the contexts they are provided in, to differ from current human practices.

2. Education and awareness

Public engagement results

Value of education and awareness

The juries considered what else can be done to build confidence in AI decisions. Jurors made a number of suggestions, 40% of which related to education or awareness-building activities.

Jurors ranked their suggestions, selecting ten that they felt would most effectively increase confidence. Six of these suggestions covered education and awareness. Excerpts from suggestions include:

- "Education: in schools/colleges…TV…radio…"

- "Public awareness and education."

- "Greater public awareness and involvement in…AI…"

Jurors therefore made a clear statement that, as a complement to the value of explanations (in certain contexts), a broader public understanding of the technology is desirable. It was proposed that such education and awareness could cover topics including:

- how AI decision-making systems work;

- key benefits; and

- common misconceptions.

Ownership of education and awareness

Generally, jurors did not specify who should be responsible for delivering education and awareness programmes (although the government was most commonly mentioned). Instead, jurors tended to focus on the method of delivery, including:

- social media;

- broadcast media; and

- the national curriculum.

<u>Industry engagement results</u>

Education and awareness also emerged as a key theme from the three roundtable discussions, although there were divergent views as to the value, and ownership, of such initiatives.

Value of education and awareness

Participants from roundtable 2 discussed education as a more narrow and directed activity than the jurors. Some participants felt a level of education for individuals is necessary so that specific explanations of AI decisions can be properly understood. Other participants at this roundtable saw the benefit of a broader educative piece, particularly in tackling misconceptions around AI, and what was perceived as its distorted portrayal in the media.

Participants at the other roundtables recognised the value of education and awareness to increase public awareness of rights related to AI decisions (which was currently perceived as low). However, a participant from roundtable 1 raised the possibility that this may incentivise behaviour (and increase the exercise of rights) that some organisations may not want due to additional burden on, and disruption to their services.

Additionally, discussion on roundtable 3 revealed a more sceptical perspective on the value of education and awareness. It was suggested that putting more information out about AI and decision-making may actually confuse matters. Participants felt there is no single (and simple) message to communicate.

Ownership of education and awareness

Participants gave a range of views on where ownership lies for education and awareness campaigns. Some felt that the organisations developing and using systems for AI decisions had a responsibility. Examples were

cited relating to the efforts made by some organisations to educate their customers about data protection in the run-up to the GDPR.

It was also raised, particularly by participants on roundtable 2 that organisations have a responsibility for internal education and awareness to ensure their staff understand the use AI decisions, its benefits and risks.

Others however thought organisations should not have a broad educative role. One participant noted that although organisations have a duty to explain their own products to customers, it is for government to educate the public on AI more generally. There was a perceived need to separate the technology from the range of AI-based products and services.

Some participants on roundtable 3 highlighted the role of trade bodies in educating the public on matters concerning their industry, or less directly, supporting their members by helping them to educate their customers.

3. Challenges

Public engagement results

Cost and resources

Qualitative results indicated that jurors' reasons for de-prioritising explanations often related to issues of cost and resource. Jurors remarked that by not providing individuals with explanations of AI decisions, organisations could:

- "...reduce costs...";

- "...save time, money ...";

- use "...less resources...";

- "...less manpower..."; and

- better "target resources...".

Explanation detail

The issue of the level of detail of the explanation was a less marked, but still observable, challenge identified by jurors. Some jurors felt, particularly in the healthcare and criminal justice scenarios (1, 3 and 4), that explanations of AI decisions may be too complex, or delivered at a time when individuals would not:

- "... understand [the] rational[e] ..."

- "...understand [the] explanation..."

- "...be in the best place to receive or understand the reasoning..."

Industry engagement results

Cost and resources

As with the juries, roundtable discussions highlighted the cost of compliance with transparency and explanation requirements as a potential challenge for the implementation of AI decisions systems.

Participants from roundtable 2 suggested that compliance is achievable, but felt some organisations used a perceived lack of technical feasibility as an excuse for not implementing explainable AI decision-systems. Participants thought that, in reality, cost and resource were more likely the overriding factors.

Explanation detail

Discussions at all three roundtables touched on level of detail as a possible challenge to the provision of explanations of AI decisions.

While jurors suggested individuals may not understand the detail provided in an explanation, roundtable participants noted a number of other risks associated with explanations that may be overly detailed:

- Distrust – giving individuals too much information about AI decisions may actually increase distrust or fear due to revealing the underlying complexities of the process.

- Commercial sensitivities – detailed explanations could disclose commercially sensitive material or infringe intellectual property.

- Third-party personal data – explanations may necessarily include the personal data of individuals other than the subject of the decision, potentially breaching data protection.

- Gaming – an explanation revealing too much about an AI decision may lead to gaming or otherwise exploiting the system.

Internal accountability

Accountability within organisations for implementing governance frameworks and ensuring the appropriate design (or procurement) and

deployment of explainable AI decision-systems emerged as a theme from discussions, particularly roundtable 2.

Participants indicated this can be challenging. With multiple staff across different business functions involved at various stages of the process, there are difficulties in assigning responsibility and ultimate ownership for AI decision-systems.

Some participants noted that traditional legal and compliance functions had ownership, some advised that responsibility was distributed across their organisation, while others had not yet assigned ownership for this. Participants also noted that issues of ownership and accountability are further complicated when buying off-the-shelf AI decision-systems, or outsourcing the development or deployment of these systems.

Regulation and guidance

Some participants suggested that the lack of a broad regulatory framework around AI is a challenge. They felt it is not clear whether there is a legal incentive to explain AI decisions to individuals, thereby causing a staggered and piecemeal attempt at implementing and developing explainable AI across different sectors. Participants recommended that the planned guidance distinguishes between the legal requirements and ethical, or good practice, considerations.

Additionally, participants observed there are limited guidance or tools to assist in the selection of appropriate AI models, and to assess the impact. Participants had mixed views as to the effectiveness of DPIAs, given their primary role as data protection compliance tools, and suggested they may be limited in their ability to test broader ethical considerations. While participants welcomed guidance that would help to address this gap, there was a general consensus that it should not be too prescriptive or unduly inhibitive, leaving space for organisations to innovate.

Innovation

The pace of technological innovation was discussed as a challenge to providing explanations of AI decisions. Some participants expressed frustration that, within their organisations, new innovative products were being developed so quickly and frequently that legal or compliance departments were not able to provide input. This hampered their ability to embed explanation capabilities as a core requirement of the products before they were deemed ready for market by management.

Discussion

1. Context

The strongest message emerging from the juries and roundtables was that context matters. Depending on context, the importance, purpose and expectations of explanations can change dramatically.

The unique characteristics of the four scenarios posed to jurors and the qualitative results from both strands of research indicate there may be several contextual factors contributing to this, including:

- The urgency of the decision – Is the decision time-sensitive? Or is there time for an individual to reflect on it?

- The impact of the decision – Is the decision safety-critical? Does it affect someone's legal status? Or are the consequences less severe?

- The ability to change the factors influencing the decision – Can an individual alter their behaviour for a future decision? Or are the factors fixed?

- The scope for bias in the decision – Is the decision non-controversial? Or might the decision be challenged on the basis of bias?

- The scope for interpretation in the decision-making process – Are the inferences made open to interpretation? Or is interpretation constrained due to safety and accuracy testing of the algorithm?

- The type of data used in the decision-making process – Does the decision use categories of data resulting from a scientific process? Or does it use data from social, or human processes?

- The recipient of the explanation of the decision – Does the person receiving the explanation have expertise in the domain the decision is made? Or do they have no specialist knowledge?

This suggests there is no one-size-fits-all approach for explanations of AI decisions. Rather, the content and delivery of explanations should be tailored to their audience based on a consideration of the relevant contextual factors for a particular decision.

This appears to align with the GDPR's requirements for the provision of meaningful information, and the adoption of suitable safeguards, when using solely automated decision-making. What information is 'meaningful' and what safeguards are 'suitable' is likely to differ depending on context.

However a flexible, case-by-case, approach such as this complicates matters for those looking to operationalise explainable AI decision-systems. It may therefore be useful to develop a list of explanation types to support organisations in identifying appropriate ways of explaining particular AI decisions, and delivering explanations, to the individuals they concern.

2. Education and awareness

Project Expl*AI*n is focused on explanations of AI decisions. But findings from the public and industry engagement research activities serve as a welcome reminder that explanations alone cannot address all the challenges associated with AI and its use in decision-making. Jurors in particular signalled the importance of education, awareness-raising, and involvement of the public in the development and application of AI.

This suggests that, as well as one-off engagement at the time an AI decision is made, there should be broader public engagement. This may help individuals gain a better understanding of the extent of AI decisions in everyday life, making them better equipped to anticipate its use and empowering them to be confident in interacting with such systems.

There are risks that awareness raising could simply serve to normalise the use of AI decisions, disproportionately emphasising its benefits so individuals are less likely to question its use and expect explanations. A campaign purely focused on the risks and potential negative consequences would be equally as harmful. Although no clear message emerged from the research around who should be responsible for a broad educative piece for the public, it is important that there are diverse voices behind this work to ensure a balanced message.

Consideration will be given to how the planned guidance, and broader ICO and Turing work on AI can support industry, government and other bodies to increase awareness and better engage the public on this complex topic.

3. Challenges

In the public engagement research, jurors primarily focused on when and why they did, or did not, prioritise explanations of AI decisions over accuracy. It is therefore interesting that a number of jurors considered the cost and resource burden on the organisation delivering the explanation. Although comments predominantly related to scenarios involving the use of public money, several jurors also acknowledged issues of cost and resource for the private sector organisation in the recruitment scenario. This suggests that, although not an excuse for failure to provide any explanation of an AI decision, individuals are

sensitive to the effort involved in, and the potential limitations to, the detail of an explanation in certain contexts.

The identification by some roundtable participants of cost and resource as the real challenge to delivering explanations, as opposed to technical feasibility, is important. It is reassuring that the organisations deploying these technologies are confident they can be explained. It also highlights that there is work to be done on raising the profile of explaining AI decisions at board-level within organisations, to ensure the necessary budget and personnel to address this issue. This suggests there is a space for the planned guidance to help with gaining board-level buy-in by clarifying the legal requirements for explaining AI decisions and emphasising the broader commercial and social benefits.

Where jurors identified issues with individuals being unable to understand an explanation due to the complexities of the decision, this may imply two things.

- First, a need to find ways to translate complex decision-making rationale into an appropriate form or language for a lay audience.

- Second, a need to identify the contexts in which individuals may not wish to shoulder the burden of understanding a decision and would prefer to delegate to another agent.

It does not necessarily follow that individuals are willing to forego any explanation at all in such circumstances. Instead, they may wish for other types of explanations satisfying their needs (such as information about the verification or safety of the AI decision-system). Further work on developing a list of explanation types may help here.

In light of concerns from roundtable participants around the provision of overly detailed explanations due to risks around distrust, commercial sensitivities, third-party data and gaming, there is a clear need for the planned guidance to acknowledge and balance these issues against requirements to provide appropriately detailed explanations to individuals.

Similarly, there was a lack of a broadly accepted or standardised approach for establishing internal accountability for explainable AI decision-systems across the organisations present at the roundtables. This suggests there is space for the planned guidance to support organisations in identifying the various internal and external stakeholders and assigning responsibility to ensure coherent governance of a complex multi-disciplinary area. More broadly, this may also help organisations to foster a culture supporting a multilateral, informed and responsible approach to innovation with technologies like AI.

Limitations of research

Public engagement research

We took care to design the juries in a way that minimised bias and provided balanced and impartial information for jurors, including the use of an oversight panel to review and feedback on these matters. Nonetheless, it is important to acknowledge the limitations of this research.

False dichotomy

As mentioned in the Methodology section, citizens' juries usually require jurors to make a clear choice on a particular issue. To achieve this, the trade-off between the 'explainability' and accuracy of AI decisions was presented to jurors. Some (including some roundtable participants) argue this is a false dichotomy because, although more challenging, there are still ways to explain highly accurate AI decision-making systems.

In addition, it is arguable that the accuracy of certain AI decision systems was overemphasised, and that information relating to their robustness (or lack of robustness) was omitted.

We chose to make these compromises in order to:

- simplify matters for jurors;

- present clear distinctions between each AI decision-system; and

- gauge public opinion on highly accurate, but opaque, AI decision-systems (which, although arguably rare now, may be more widespread in the near future).

However, we acknowledge that such choices may have led jurors to place more weight on, and trust in, the accuracy of AI decisions at the expense of giving more consideration to the potential value and utility of explanations.

Negative impact

Although the consequences of receiving an inaccurate AI decision were discussed, the framing of the scenarios and associated questions may also have influenced jurors to prioritise accuracy over 'explainability' in certain contexts. The phrasing of the questions may have encouraged jurors to consider the three AI decision-systems from the perspective of the majority for whom the decision is accurate, as opposed the minority receiving an inaccurate decision.

Some roundtable attendees questioned whether jurors would have been as willing to de-prioritise explanations had they been asked to consider the same question from the point of view of receiving an inaccurate decision (eg an incorrect diagnosis or match in the healthcare scenarios).

Encouraging greater consideration of this negative impact may have yielded further insights into expectations around explanations in contexts where they were not prioritised by jurors.

<u>Explanation format, content and delivery</u>

Jurors considered when and why explanations were (and were not) important in different scenarios, but were not asked to consider the format, content, or timing of delivery of the explanation (ie before or after a decision is made).

As such, findings from the juries do not directly translate into advice for organisations on what information to give to individuals, how to present it, and when to do so. Roundtable discussions highlighted this as an area organisations are keen to have a steer on.

While the reasoning for prioritising explanations in certain scenarios offers valuable insights into the purposes jurors want those explanations to serve in different settings, further work is required to map these purposes on to different explanation types.

Industry engagement research

<u>Representativeness</u>

Public, private and third sectors were represented at the roundtables as below:

Sector	Total
Public	9
Private	43
Third	5

We took steps to obtain a range of participants from across sectors, but the sample was not representative of the UK's sectoral landscape.

It is possible that findings from the roundtable overstate certain concerns (for instance from a well-represented private sector) and did not capture other input and insights from less represented sectors.

In addition, the make-up of participants within sectors was not intended to be representative. For instance, while SMEs form 99.3% of the UK's private sector[11], they only formed a small proportion of the organisations at the roundtables.

11

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/746599/OFFICIAL_SENSITIVE_-_BPE_2018_-_statistical_release_FINAL_FINAL.pdf

Conclusion

More and more organisations, across all sectors, are using AI to make, or to support, decisions about individuals. While there are undoubtedly benefits to this use of AI, there are also risks. Increasingly, governments and regulators are considering how to mitigate these risks. One such risk is the lack of transparency around how AI decisions are made. To help address this, the UK Government's AI Sector Deal tasked the ICO and The Turing to develop guidance assisting organisations with explaining AI decisions.

To inform this guidance, the ICO and The Turing carried out primary evidence-based research through public engagement (citizens' juries) and industry engagement (roundtables). Three key themes emerged from this research:

1. Context is key. The importance, purpose, and expectation of explanations of AI decisions depend on several interrelated contextual factors such as the impact of the decision, the ability to change it, and the data used to inform it.

2. While unclear where responsibility lies, there is a desire for a range of education and awareness raising activities to better engage and inform the public on the use, benefits and risks of AI in decision-making.

3. There are several challenges in explaining AI decisions, including cost, commercial sensitivities, and a lack of internal organisational accountability. However, technical feasibility was generally not considered an issue.

While there are some limitations to the research, the above findings remain incredibly valuable, giving key insights into a range of different stakeholder views on explaining AI decisions. The juries in particular provided a unique and informed public opinion on this complex issue.

As well as benefiting the ICO and The Turing on Project Expl*AI*n, it is hoped that others too can make use of these findings for their own thinking, research, or development of explainable AI decisions. All materials and reports generated for and by the jurors are freely available to access from the GM PSTRC website[12].

---

[12] http://www.patientsafety.manchester.ac.uk/research/themes/safety-informatics/citizens-juries/

Next steps

Although these strands of research are only one aspect of Project Expl*AI*n, the themes raised will inform further work to consider, assess and test the ICO and The Turing's understanding of the implications. The issues raised will also be addressed in the guidance currently under development.

Overleaf is an overview of the planned format and content for the guidance. This has been refined based on outputs from this research.

The current plan is a modular framework based around a set of overarching principles. The principles inform how organisations should approach the use of AI for making decisions about individuals. Guidance on organisational controls, technical controls, and explanation delivery support the implementation of the principles.

The planned guidance is subject to change based on further work on Project Expl*AI*n. A full draft will be put out for public consultation over the summer. Any and all interested parties are encouraged to comment and make suggestions during the consultation period. This will be signposted on the ICO website, in the ICO e-newsletter and through social media. Following consultation, the guidance will be published later in the autumn.

Overview

Legal framework – the legal requirements around AI decisions, including data protection and other relevant regimes.

Definitions – explaining what is meant by 'AI', 'AI decisions', and providing a list of explanation types.

Benefits and risks – highlighting the benefits of explaining AI decisions, and the risks of not doing so for organisations, individuals and wider society.

Principles

Transparency – the need to be open and engaged with customers and the wider public about the use of AI decisions.

Context – the need for explanations to reflect the unique context in which an AI decision is made.

Accountability – the need to have corporate governance measures in place to appropriately manage the whole AI decision process.

Organisational controls

Roles – guidance on mapping the roles involved in AI decisions, identifying reporting lines and assigning responsibility.

Policies and procedures – guidance on the necessary policies including training, risk assessment and monitoring.

Documentation – guidance on documenting AI decisions, including maintaining an audit trail.

Technical controls

Data collection – guidance on ensuring the integrity of the data used to train AI models.

Model selection – guidance on appropriate AI models for different contexts.

Explanation extraction – guidance on approaches to drawing out explanations of AI decisions.

Explanation delivery

Proactive engagement – guidance on engagement with individuals in advance of an AI decision.

Explanation selection – guidance on appropriate explanations types for different contexts.

Explanation timing – guidance on appropriate timing of delivery of explanations for different contexts.