Artificial intelligence (AI) & explainability
Citizens' Juries Report

A report of two citizens' juries designed to explore whether automated decisions that affect people should require an explanation, even if that impacts AI performance

May 2019

Commissioned by:

NIHR | Greater Manchester Patient Safety Translational Research Centre

Salford Royal NHS Foundation Trust University Teaching Trust

ico. Information Commissioner's Office

MANCHESTER 1824 The University of Manchester

Designed and delivered by:

citizens juries

Jefferson Center

Table of Contents

Foreword

Artificial intelligence (AI) is becoming increasingly widespread, and is becoming more effective in guiding decision-making in real-world domains. At the same time, the public's understanding of, and attitude to, AI decision making is not always clear. For example, several organisations are utilising big data to interpret scans, using pattern matching to spot anomalies that radiologists may miss, but it is unclear whether patients and the public will accept and trust such "artificial doctors".

One way to increase trust might be to explain what the algorithms are doing. Some argue that this is an ethical imperative and there should be an audit trail tracking the logic of all AI decisions or advice – like a flight data recorder – that can then be examined. However, the nature of machine learning may make it difficult or impossible to identify and spell out the steps that an evolving algorithm follows to reach a conclusion. This suggests that there may be a potential trade-off between AI transparency and AI performance, if the most accurate systems are also the most complex ones.

When our society is facing challenging questions like these, it is important to involve citizens in the discussion and give them a voice in policy making – in the end it is their lives that will be affected. We know a little about what the public thinks about AI through surveys, but there has been little research focusing on the particular trade-off between AI "explainability" and AI performance. For this reason the NIHR Greater Manchester Patient Safety Translational Research Centre (NIHR PSTRC) and the Information Commissioner's Office (ICO) commissioned Citizens Juries c.i.c. and the Jefferson Center to investigate this matter by organising citizens' juries in Coventry and Manchester.

The message emerging from this report is that, for those at the receiving end of public and private services, the context of the decision is key; accuracy of decision making outweighs explainability in healthcare, but not necessarily in other, non-medical situations. In general, whether (and how) an explanation should be given for an automated decision depends on whether an explanation would ordinarily be expected for a human decision; this suggests that moving from human decision-making to AI decision-making should not affect explainability.

One of the core aims of the NIHR PSTRC is to develop and test evidence-based digital and behavioural interventions to improve patient safety in primary care and at transitions between care-settings. Patient safety is not just about medical procedures or treatment options, it is about people; both those who deliver care and those who receive it or work in partnership together. The PSTRC has a strong involvement and engagement agenda working with healthcare professionals, the NHS, local authorities, industry and patients, carers and members of the public. The citizens' juries are integral to this agenda and play a key role in understanding public trust in digital health technologies.

As the UK regulator for data protection, the ICO believes that the benefits of AI can be realised without compromising people's data rights. The ICO is currently working with the Alan Turing Institute (the national institute for data science and AI) on Project ExplAIn, to produce guidance for organisations on explaining AI decisions to the people affected by them. The work of the citizens' juries is an essential part of our research and will help us in developing this guidance.

We would like to thank the 36 jury members for their time investment and important contributions to this report. We also thank Dr. Malcolm Oswald from Citizens Juries c.i.c. and Kyle Bozentko and Sarah Atwood from the Jefferson Center for their dedicated effort in designing and conducting the juries. Finally we thank the five expert witnesses (Prof. Sophia Olhede, Rhiannon Webster, Dr. André Freitas, Prof. Alan Winfield, and Dr. Allan Tucker); the four scenario witnesses; and the independent oversight panel (Reema Patel, Dr David Leslie, and Prof Søren Holm) for their contributions to the juries.

Professor Stephen Campbell
Director
NIHR GM Patient Safety Translational Research Centre

S. McD

Simon McDougall
Executive Director
ICO Technology, Policy & Innovation

Report on Two Citizens' Juries about Artificial Intelligence (AI) and Explainability

On 18 February 2019, 18 people gathered at the Welcome Centre in Coventry and began a five-day "citizens' jury". A week later, a different cross-section of 18 citizens met at the Bridgewater Hall in Manchester for five days and went through the same process. The task for these 36 citizens was to tackle a set of jury questions (sometimes referred to as the "jury charge") about how to balance artificial intelligence (AI) performance and the need for citizens to be provided with an explanation of automated decisions that affect them. Over five days, the citizens heard from, and asked questions of, expert witnesses, and carried out group exercises to explore the jury questions. They deliberated and reached conclusions together, and were polled on their individual views at the start and end of the jury.

The jury participants were drawn from around Coventry and Manchester respectively. Chosen from over 450 applicants, jury members were selected to broadly represent the demographic mix of England (according to the 2011 census) in terms of age, gender, ethnicity, educational attainment and employment status.

This report explains why the two juries were held, how they were designed, how the jurors were recruited, what they did, the juries' answers to the jury questions, and the results of the questionnaires completed at the start and end of the juries. It also summarises the key findings. Further information about the juries can be found at: www.bit.ly/GMPSTRCCitizensJuries

Why the citizens' juries were run

The General Data Protection Regulation (GDPR) has a requirement for some form of explanation to be provided to individuals where an automated decision (such as one generated by AI) is made about them. GDPR has applied in the UK since May 2018. To date, this provision has not yet been tested in the courts, and it is not yet entirely clear, for example, what form of explanation is required, or whether different types of explanation are required in different contexts (e.g. healthcare, criminal justice etc.).

One emerging use of automated decision-making is in healthcare. AI is expected to soon make its way into the NHS, for example, for making diagnoses. Some researchers have argued that "black boxes" are unacceptable, and any AI system requires transparency so that users can understand the basis for any advice or recommendations that are offered.[1] The NIHR Greater Manchester Patient Safety Translational Research Centre at the University of Manchester has a specific focus on potential safety issues associated with digital technologies, and therefore wanted to learn more about whether patients and the public felt it was important to receive an explanation when AI was used in their healthcare, even if that was at the expense of less accurate decision-making. This question is not hypothetical. Many AI researchers have warned that the most accurate AI, which uses "deep learning", is relatively opaque: even the designers of the software find it very challenging to explain how the AI reaches its predictions.

The Information Commissioner's Office (ICO) is aware that organisations are confronting this problem of providing explanations for AI decisions and a number of influential reports[1] have

[1] For example: Professor Dame Wendy Hall and Jérôme Pesenti *Growing the artificial intelligence industry in the*

highlighted the importance of explainability. As part of the UK's AI Sector Deal[2], the government has tasked the ICO to work with the Alan Turing Institute to produce guidance for developers and users of AI on this subject, and so the ICO was keen to co-commission two citizens' juries with the University of Manchester to explore some scenarios and questions with the public. A pair of juries was commissioned in order to validate that the results from one jury were broadly confirmed by a second group of 18 citizens, and not peculiar to a particular group of 18 jurors.

Planning and designing the citizens' juries

The two juries were planned, designed and refined over a period of approximately six months. There were many aspects to the jury design including:

- the jury questions;
- the jury demographics and recruitment approach;
- the brief and selection of individuals to act as expert witnesses;
- the brief and selection of individuals to act as members of the oversight panel;
- the programme of jury activities across the five days; and
- the design of the questionnaires completed at the start and end of the juries.

The design documentation is available at: www.bit.ly/GMPSTRCCitizensJuries

Bias, both conscious and unconscious, is a risk to consider in planning citizens' juries.[2] For example, it is very difficult to know what constitutes "impartial information" or balanced argument, and almost every design choice, even down to a bullet point on a presenter's slide, could be challenged on grounds that it might manipulate the citizens' jury towards one outcome or another.

Bias can be monitored and minimised but not eliminated. To monitor and minimise bias on this project, an oversight panel was appointed to review the jury design and materials, and report potential bias. Members of the panel each completed a bias evaluation form, published at www.bit.ly/GMPSTRCCitizensJuries. Every panel member was either "mostly satisfied" or "fully satisfied" that the two juries were designed with the aim of minimising bias, and were successful in minimising bias.

The end of jury questionnaires also asked about bias. Three out of 18 jurors on Coventry jury reported that there was some bias in favour of AI performance whereas all 18 jurors in Manchester stated that they felt they were presented with a fair balance of information.

Other design controls used to monitor and minimise bias included:

- The commissioners of the juries were involved in setting the jury questions but were independent from the design of the jury process and outcomes;
- The two juries worked with independent facilitators from the Jefferson Center to construct their own reports of their findings; and
- The detailed jury design and results documentation were published.

UK. October 2017 available at: www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

[2] https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal

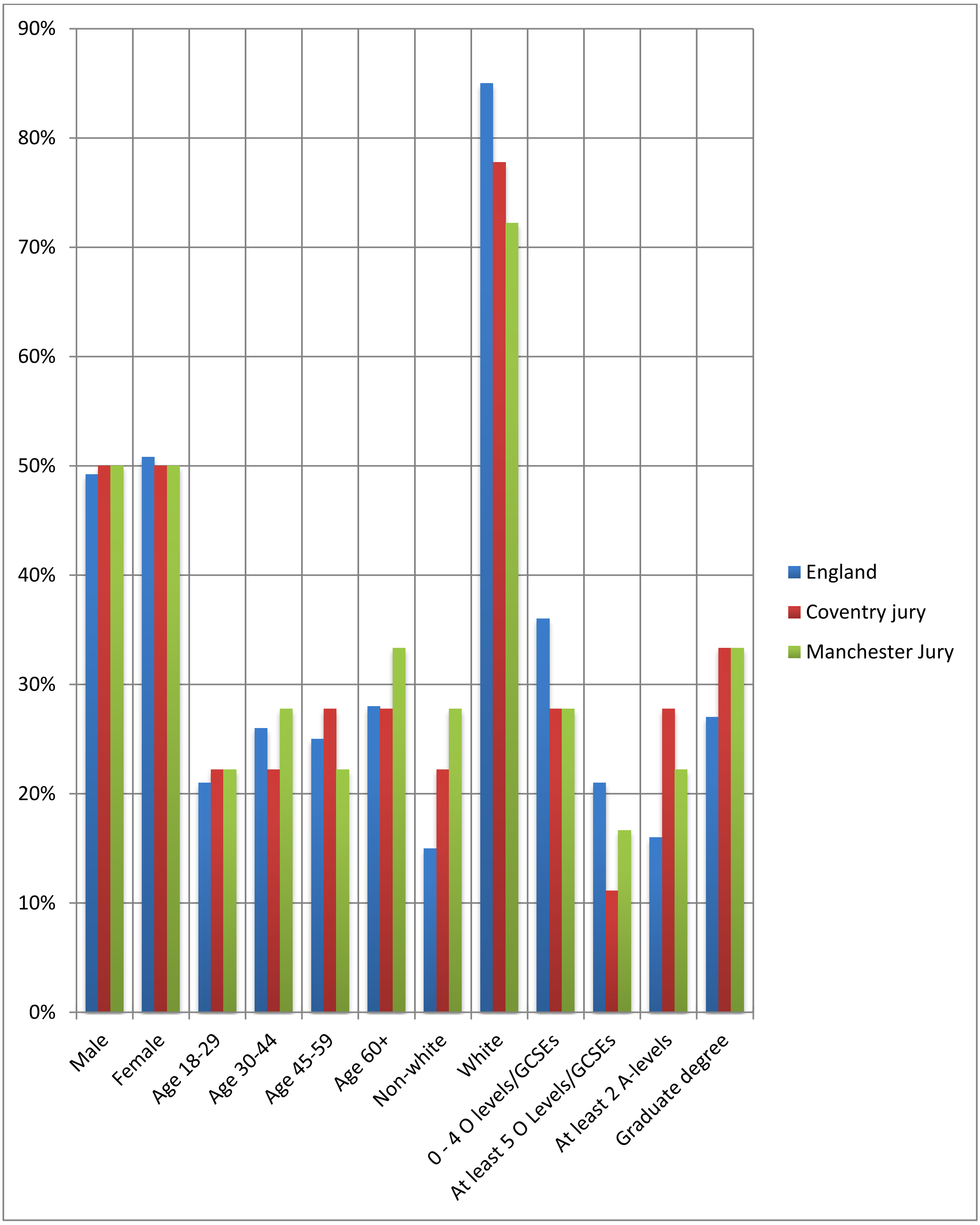
Jury recruitment

In total, 180 people applied to be part of the Coventry citizens' jury, and 271 applied for the Manchester jury. They applied by entering their personal details, including relevant demographics, into an on-line survey. Shortlisted candidates had a brief telephone interview so that any ineligible candidates (e.g. AI or data protection experts) could be identified and excluded. 18 people were recruited to each jury to provide a broadly representative sample of resident adults in England. Of the 36 jurors, 32 people were found through the "Indeed" jobs website, two through a voluntary action website, and two through word of mouth. In order to guard against any bias from using a jobs website, the sample was controlled for employment status (14 - employed; 7 - retired or students; 4 - unemployed; 8 - self-employed; 3 - "other").

Each juror was paid £500 for five days plus a £25 cash expense allowance. Reserve jurors were also recruited and paid to attend the morning of the first day. One reserve was needed to substitute for a jury member who did not attend the Coventry jury, and two reserves were needed in Manchester for reasons of illness.

The sample chosen was controlled for gender, age range, ethnicity (in terms of white/other), and educational attainment (see chart below). The samples were chosen to match (within target ranges) the demographics of 2011 UK Census Data for England from the Office for National Statistics.

Figure 1: Demographic make-up of Coventry and Manchester juries



In a sample of 18 people, it is possible by chance to recruit a disproportionately large number of people who are either very comfortable with AI making decisions or very uncomfortable with automated decisions. A skewed sample could affect the jury outcomes. For this reason, applicants also answered the following question taken from a RSA national survey[3] to test their prior views on

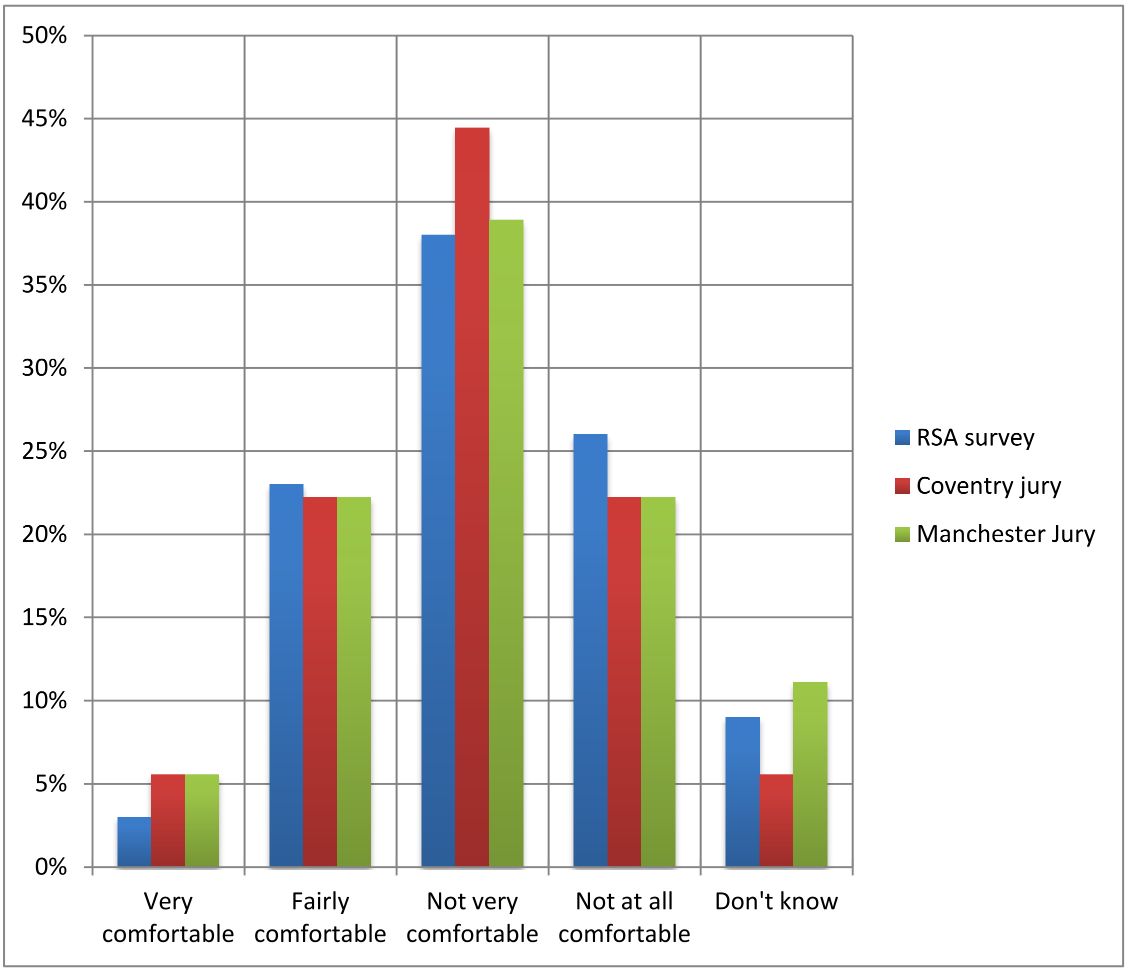
[3] Target sample percentages based on online survey of a representative sample of 2074 UK adults aged over 18 by YouGov plc in April 2018. See Figure 11 on page 30, and page 42 of the RSA report “Artificial Intelligence – Real Public Engagement”, available at https://www.thersa.org/discover/publications-and-articles/reports/artificial-intelligence-real-public-engagement

how comfortable they felt about AI:

> 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

The range of views in each jury on the above question matched closely those reported in the UK survey of 2074 adults commissioned by the RSA.

Figure 2: Prior view on RSA survey question of Coventry and Manchester jurors



The jury process and juries' reports

The two juries in Coventry and Manchester followed the same 5-day programme:

- Pre-jury questionnaire completed at the start of day one
- Two facilitators: Kyle Bozentko and Sarah Atwood of the Jefferson Center
- five expert witnesses;
- four scenario witnesses;
- Group exercises and deliberation; and
- End-of-jury questionnaire completed at the end of day five.

On days one and two of the juries, the jury watched introductory videos on AI, and heard evidence from, and asked questions of, four of the expert witnesses. Prof. Sonia Olhede explained the trade-off between AI performance and explainability. Rhiannon Webster provided an introduction to relevant data protection law. Dr. Andre Freitas made the case for prioritising AI performance above explainability, and Prof. Alan Winfield made the opposing case.

On days three and four, the juries considered, discussed and voted on four scenarios:

- Scenario 1: using AI in the NHS to diagnose stroke patients
- Scenario 2: using AI to shortlist job candidates in a private company
- Scenario 3: using AI to select patients to receive a kidney transplant
- Scenario 4: using AI to select people to be offered a rehabilitation programme rather than face trial for a minor offence for which they had been charged.

Each scenario involved three hypothetical systems: A, B and C. In each scenario:

- system A was 75% accurate and provided a full explanation as to how system A reached its conclusion;
- system B was 85% accurate and provided a partial explanation
- system C was 95% accurate but provided no explanation.

For each scenario, the jurors read and discussed the scenario text and watched a video of an interview with a scenario witness who answered a variety of questions in the interview to help the jury relate to the scenario. For example, for scenario 1, the interviewee filmed was a person who works with stroke patients, and she answered questions about how important an accurate stroke diagnosis is to a patient, and how important it is for a patient to hear how their diagnosis was reached. Dr. Allan Tucker then gave a presentation to the jury to explain the kind of explanation that systems A, B and C might be able to give for that specific scenario. Finally, the jury members individually answered questions about the scenario using an online survey tool.

On day five, the jurors did exercises and answered a set of general jury questions to try to generalise some of their thinking about balancing AI performance and explainability. Throughout the five-day process, and particularly on day five, considerable time was given to group work, enabling jury members to deliberate together, to raise questions and to articulate their reasoning. At the end of day five, the jurors viewed a report that had been built from the reasoning that had come from their group work and their voting. Kyle Bozentko, the facilitator of the two juries from the Jefferson Center, constructed the jurors' report with each jury using the votes and ranked reasons. The juries were led page-by-page through the jury report, which was displayed on the projector screen, to gain the jurors' acceptance that it fairly represented their views. After each jury, the reports were reviewed and sent to jury members so that any final changes could be made before the two jurors' reports were published.

Jury questions and answers

The juries tackled four scenarios, each with a set of questions, and three general questions. How the jurors voted on these is summarised below. The full text of the four scenarios and associated questions are provided in Appendix B. Each individual answered the questions individually using an online survey tool. The jurors also entered reasons for their answers, and examples of these are reproduced below. A spreadsheet containing the full set of jury questions and results is available at: www.bit.ly/GMPSTRCCitizensJuries

The four scenarios were constructed by the jury commissioners to provide clear choices rather than reflect real-world AI performance levels. In each case the juries had to decide whether it was better to choose system A, B or C.

Table 1 – Overview of Systems A, B and C used in the scenarios

	System A – Expert System	**System B – Random Forests**	**System C – Deep Learning**
Accuracy	75% (below human expert level)	85% (human expert level)	95% (beyond human level)
Transparency	Full explanation	Partial explanation	No explanation

Scenarios 1 and 3: stroke diagnosis and kidney transplant

The juries considered two healthcare scenarios: scenario 1 on using AI for stroke diagnosis, and scenario 3 on matching donated kidneys to kidney recipients. In both cases, the two juries favoured AI performance over explanation.

They were asked three questions about these two scenarios (Q1, Q2 and Q3 in the tables below). The questions and their answers are summarised below.

Table 2 - Scenarios 1 & 3, Q1
How important is it for a patient to receive an explanation of an automated decision?

Q1: How important it is for a patient to receive an explanation of an automated decision?	**Very important**	**Fairly important**	**Not very important**	**Not at all important**	**Don't know**
Stroke scenario - Coventry	1	3	10	4	0
Stroke scenario - Manchester	3	3	11	0	1
Kidney transplant scenario - Coventry	1	7	8	2	0
Kidney transplant scenario - Manchester	0	5	10	3	0

Table 3- Scenarios 1 & 3, Q2
If system C was chosen by the NHS, almost no explanation would be provided.
How much does this matter?

Q2: If system C was chosen by the NHS, almost no explanation would be provided. How much does this matter?	Very much	Quite a lot	Not very much	Not at all	Don't know
Stroke scenario - Coventry	2	1	12	3	0
Stroke scenario - Manchester	2	4	12	0	0
Kidney transplant scenario - Coventry	1	2	8	7	0
Kidney transplant scenario - Manchester	0	1	12	5	0

Individual jury members recorded their reasons for their answers to this question in an online survey. Reasons given for prioritising AI performance included:

For stroke scenario:

- "Because fixing the problem supersedes the need for each individual to understand how exactly it came about."
- "According to the witness, most stroke victims don't usually ask how the stroke was diagnosed."

For kidney transplant scenario:

- "Patient just needs to know they've got a match."
- "Life changing situation means it's more important to concentrate on transplant."

Table 4 - Scenarios 1 & 3, Q3
Which automated decision system do you think the NHS should choose?

Q3: Which automated decision system do you think the NHS should choose?	Juror Votes – System A	Juror Votes – System B	Juror Votes - System C
Stroke scenario - Coventry	0	2	16
Stroke scenario - Manchester	0	3	15
Kidney transplant scenario - Coventry	1	2	15
Kidney transplant scenario - Manchester	0	0	18

Jurors worked in groups to identify and then rank their reasons for and against each of the three systems. A full set of ranked reasons can be found in the jurors' reports for Coventry and Manchester.

The highest ranked reasons for preferring system C in the stroke scenario were similar for Coventry and Manchester:

- "It is more accurate – higher diagnostic success rate using the same symptoms presented by the patient in an initial consultation than with system A and B". (Coventry)
- "Initially it is accuracy of the diagnosis that is more important than the explanation"

(Manchester).

The highest ranked reasons for preferring system C in the kidney transplant scenario were:

- "It has a 95% accuracy rate and patients are less likely to ask how they are matched. When you win the lottery, you just need to know that you won, not which order the balls come up." (Coventry)
- "The greater degree of accuracy reduces the number of failures feeding back into system – ultimately reducing waiting lists." (Manchester)

Scenarios 2 and 4: Job recruitment and criminal justice

Scenarios 2 and 4 – on using AI for shortlisting job candidates, and for choosing which defendants are offered a rehabilitation programme rather than face criminal trial – yielded very different results to scenarios 1 and 3. Many jurors recognised the importance to affected individuals of receiving an explanation for decisions made about them. The two juries were asked the same three questions about scenarios 2 and 4 as for scenarios 1 and 3 (Q1, Q2 and Q3 in the tables below). The questions and their answers are summarised below.

Table 5 - Scenarios 2 & 4, Q1
How important it is for an individual to receive an explanation of an automated decision?

Q1: How important it is for an individual to receive an explanation of an automated decision?	Very important	Fairly important	Not very important	Not at all important	Don't know
Recruitment scenario - Coventry	4	9	5	0	0
Recruitment scenario - Manchester	4	7	6	1	0
Criminal justice scenario - Coventry	10	3	5	0	0
Criminal justice scenario - Manchester	7	6	3	2	0

Table 6 - Scenarios 2 & 4, Q2
If system C was chosen, almost no explanation would be provided.
How much does this matter?

Q2: If system C was chosen, almost no explanation would be provided. How much does this matter?	Very much	Quite a lot	Not very much	Not at all	Don't know
Recruitment scenario - Coventry	7	7	4	0	0
Recruitment scenario - Manchester	3	5	9	1	0
Criminal justice scenario - Coventry	11	1	5	1	0
Criminal justice scenario - Manchester	6	8	2	2	0

The tables above suggest that many jurors felt that explanations mattered considerably more in the recruitment and criminal justice scenario than in the two healthcare scenarios.

Reasons given for prioritising AI performance included:

- "The best candidate would be being selected which is the main focus of the business" (recruitment scenario);
- "Accuracy of the decision is paramount as it may determine the individual's future" (criminal justice scenario).

Reasons given for prioritising explanation included:

- "Candidates should be provided with feedback in order to allow them to improve" and "Company has no indication of how AI reached decision, what criteria is a successful candidate" (recruitment scenario).
- "The data is so subjective that there are bound to be so many spurious decisions" and "Transparency for offender, current victim and any potential future victim" (criminal justice scenario).

Table 7 - Scenarios 2 & 4, Q3
Which automated decision system do you think should be chosen?

	Juror Votes – System A	**Juror Votes – System B**	**Juror Votes - System C**
Recruitment scenario - Coventry	2	14	2
Recruitment scenario - Manchester	5	6	7
Criminal justice scenario - Coventry	8	7	3
Criminal justice scenario - Manchester	7	6	5

In the recruitment scenario, there was a clear preference amongst the Coventry jury for system B. Otherwise, unlike the two healthcare scenarios, the tables above suggest that jurors were divided as to which system they preferred, and more balanced when weighing AI performance against explainability.

In the recruitment scenario, system B had most support overall amongst jurors, and the highest ranking reasons for this preference were:

- "It has key points/features for detection. Applicant receives some feedback, also reduces time and resources" (Coventry)
- "It offers a high level of accuracy and gives some explanation/feedback" (Manchester).

In the criminal justice scenario, system A had slightly more support overall than systems B and C. The reasons ranked most highly by jurors for preferring system A were:

- "It is fully transparent, and bias can be clearly identified allowing human intervention/appeal of decision" (Coventry)
- "There is more interaction with officers. Feelings and remorse will be taken into account and a full explanation will be given for the end result" (Manchester).

General Jury Questions

On day five of the juries, having considered the four specific scenarios, the jury participants worked together to draw general conclusions about balancing AI performance and explainability. They answered three questions individually, and as a group ranked the most important reasons for their answers (each person "voting" by placing stickers against reasons on flip charts).

Table 8 – General Question 1
Should automated decisions be explained to individuals? Choose one of the following options.

	Some form of explanation should be offered to individuals in all contexts, even if there is a reduction in accuracy.	Some form of explanation should be offered to individuals in certain contexts, even if there is a reduction in accuracy in those contexts.	Automated decisions should be explained to individuals with as much detail as practically possible without reducing accuracy.
Coventry	0	11	7
Manchester	0	5	13

Table 9 – General Question 2
Should automated decisions be explained to individuals when used in contexts where a human decision-maker would not usually be expected to provide an explanation?
Choose one of the following options.

	Automated decisions should be explained to individuals in all contexts, even if human decisions would not usually be explained.	Automated decisions should be explained to individuals in certain contexts, even if human decisions would not usually be explained in those contexts.	Automated decisions should not be explained to individuals in contexts where human decisions would not usually be explained.
Coventry	0	8	10
Manchester	1	3	14

Table 10 – General Question 3
Should human and automated decisions require similar explanations?

	Yes	No	Don't know
Coventry	9	8	1
Manchester	12	4	2

The tables above indicate a clear conclusion amongst jury members that context matters, and explanations are not necessary in some contexts, and in some contexts they would not be desirable if that meant a reduction in AI performance. A majority of jurors felt that an explanation of how an automated decision was reached was not necessary in contexts where explanations would not normally be given for a non-automated decision.

Start and end of jury questionnaire results

All 18 individuals from each jury completed a questionnaire both at the start and end of the jury.

The questionnaire design and the full results are available at: www.bit.ly/GMPSTRCCitizensJuries

During jury recruitment, applicants were asked a question (taken from a YouGov poll of the public commissioned by the RSA[4]) to select a broadly representative sample in terms of how comfortable jurors felt about automated decision making. The spread of answers to this question amongst each jury closely matched the spread of answers given by the YouGov survey respondents. The same question was asked in the post-jury questionnaire on day five to gauge whether, and if so how, the views of jury members had changed by the end of the jury process.

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.

Table 11: Pre and post jury questionnaire results on automated decisions

How comfortable?	**Coventry**		**Manchester**	
	Pre	**Post**	**Pre**	**Post**
Very comfortable	1	4	1	7
Fairly comfortable	4	12	4	8
Not very comfortable	8	2	7	3
Not at all comfortable	4	0	4	0
Don’t know	1	0	2	0

The start-of-jury questionnaire contained just one question, and this same question was included in the end-of-jury questionnaire, leading to the following results:

[You will hear]/[The citizens’ jury heard] that decisions which affect people's lives can be automated using artificial intelligence. For example, in healthcare, the best-performing artificial intelligence could make more accurate diagnoses than human medical specialists. However, it may not be possible to explain to the people affected how the computer reached its decision. How important is it to give an explanation? Please select the answer that best describes your view.

Table 12: Pre and post jury questionnaire results on AI performance vs explainability

How important is it to give an explanation?	**Coventry**		**Manchester**	
	Pre	**Post**	**Pre**	**Post**
Making the most accurate automated decision is more important than providing an explanation	2	9	5	10
In some circumstances an explanation should be given even if that means the automated decision is less accurate	10	8	3	6
An explanation should always be given even if that means that automated decisions are less accurate	1	0	2	1

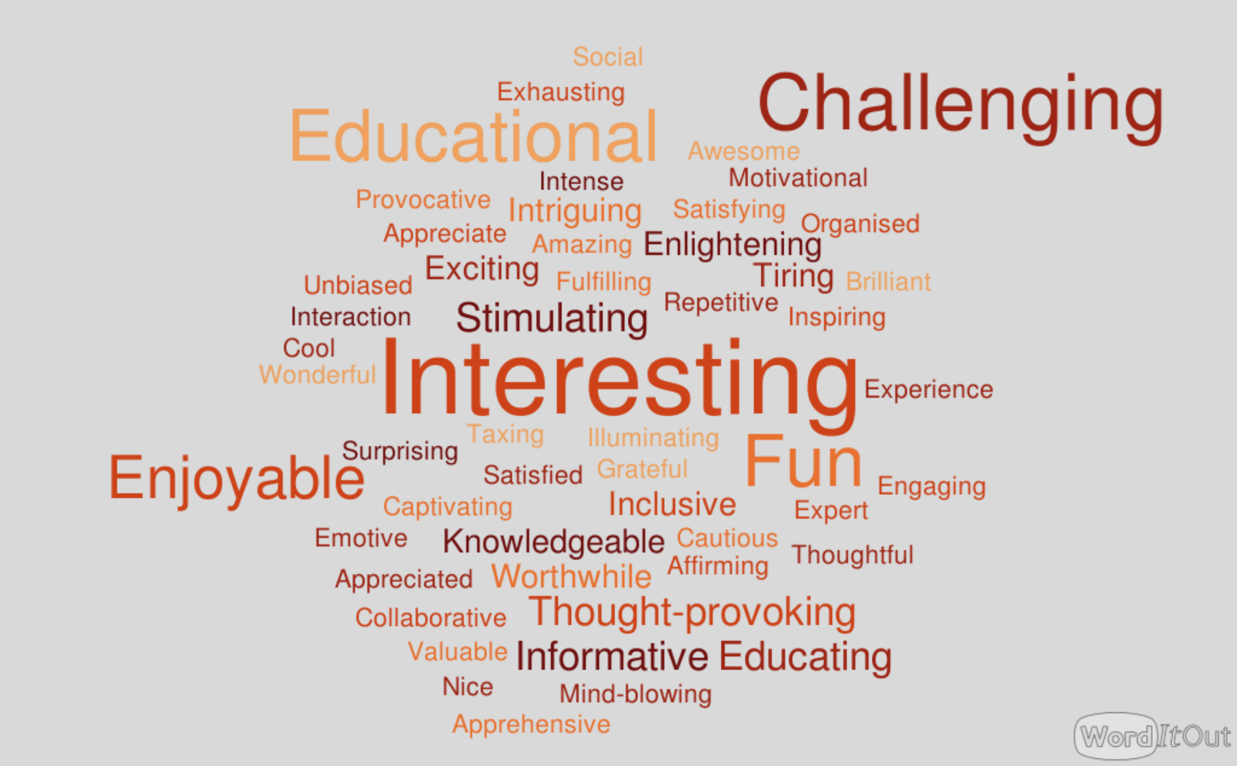
[4] Target sample percentages based on online survey of a representative sample of 2074 UK adults aged over 18 by YouGov plc in April 2018. See Figure 11 on page 30, and page 42 of the RSA report “Artificial Intelligence – Real Public Engagement”, available at https://www.thersa.org/discover/publications-and-articles/reports/artificial-intelligence-real-public-engagement

Humans, not computers, should always ultimately make the decisions and be able to explain them to the people affected	3	1	7	1
Don’t know	2	0	1	0

The above tables show a big swing by the end of the jury towards jurors feeling more comfortable with automated decision making, and in prioritising AI performance over explanations when decisions are automated. It is usual in citizens’ juries for many jurors to change their minds as they learn more about a subject. However, the above results should be treated with caution as the four scenarios presented to the jurors on days 3 and 4 showed AI making significantly more accurate decisions than human experts, something which has only happened in research settings so far.

In the end of jury questionnaire, each jury member was asked to provide three words to sum up their experience of the jury. The words of the 36 jury members are constructed in a “word cloud” below (large words were said more often).

Figure 3: “Word cloud” of jurors’ experience of the citizens’ jury



Key Findings

1. The importance placed on an explanation for automated decisions by the two juries varied according to the scenario they were being asked to consider.
2. In the two healthcare scenarios (stroke diagnosis and kidney transplant matching), the two juries had a clear preference for system C, where AI accuracy is highest and no explanation is available for the automated decision.
3. In the criminal justice scenario, support was spread across the three systems, but the greatest support was for system A (lowest AI accuracy, full explanation) and the fewest votes were for system C (high AI accuracy, no explanation).
4. In the job recruitment scenario, there was a clear preference for system B (average accuracy, partial explanation) in Coventry, with jurors fairly evenly split across the three systems in Manchester.
5. No juror concluded that an explanation should always be offered in all contexts (assuming that meant a reduction in AI accuracy), and a majority felt that explanations for AI decisions should be offered in situations where non-AI decisions come with an explanation.
6. Over the course of the five days, most jurors became more comfortable with, and supportive of, automated decision-making in principle, although this result should be treated with caution as the scenarios considered by the jury suggested that AI decision-making accuracy could significantly exceed that of human experts.
7. There were strong similarities between the conclusions reached by the Coventry and Manchester juries (with Manchester slightly more willing to prioritise AI accuracy over explainability), although some of their reasoning differed.
8. The independent oversight panel identified some bias in the jury materials which was then addressed, and three out of 36 jurors identified some bias in the information they received in favour of AI accuracy.

Appendix 1: further information about the juries

The Citizens' Jury Method

Like much public policy, balancing the benefits of AI performance and the need for explanations of AI is a complex area with a lot of information and many arguments to consider. Surveys and focus groups provide useful information about what the public thinks, but they are not mechanisms to inform people. A citizens' jury can tell policymakers what members of the public think once they become more informed about a policy problem. In a citizens' jury, a broadly representative sample of citizens are selected to come together for a period of days, hear expert evidence, deliberate together, and reach conclusions about questions they have been set. The method was devised by Dr Ned Crosby in the 1970s. He went on to set up the Jefferson Center, which produced the Citizens' Juries Handbook[3], the method followed by Kyle Bozentko and Sarah Atwood of the Jefferson Center when designing and running the juries in Coventry and Manchester.

Citizens' Juries are a form of "deliberative democracy", based on the idea that individuals from different backgrounds and with no special prior knowledge or expertise can come together and tackle a public policy question. A citizens' jury is a particularly relevant method for informing public bodies making value judgements. Some organisations have used citizens' juries to *make* policy decisions, even though members of juries are not elected and cannot be made accountable for decisions. For example, Melbourne City Council appointed a citizens' jury to determine how to allocate its A$5 billion budget, and the council is implementing virtually all of the jury's recommendations. A Citizens' Council was commissioned by the Irish government on whether to change the Irish Constitution on abortion recommended change, leading directly to the national referendum on the subject.

Expert witnesses

Expert witnesses were chosen to provide relevant information to the members of the jury to enable them to answer the jury questions. Each witness gave a presentation and then answered questions posed by the jurors.

The expert witnesses were issued with a brief prior to preparing their presentations. It is published at: www.bit.ly/GMPSTRCCitizensJuries. Their slides were reviewed in advance by the oversight panel who recommended changes to the slides which were made prior to the start of the juries.

The following is the information provided (in ring binders) to jurors about each witness.

Day	Expert Witness		Topic
Day 1 PM	Prof. Sofia Olhede	A professor in the Department of Statistical Science at University College London with a special interest in how data and algorithms are impacting our daily lives.	Balancing artificial intelligence and explainability
Day 2 AM	Rhiannon Webster	A specialist information lawyer and a partner at DAC Beachcroft in London.	Law concerning data protection and artificial intelligence

Day 2 PM	Dr. Andre Freitas	A lecturer and researcher in the School of Computer Science at the University of Manchester and specialist in artificial intelligence.	Making the case for artificial intelligence performance
Day 2 PM	Prof. Alan Winfield	Professor of Robot Ethics in the Department of Engineering, Design and Mathematics at the University of the West of England who researches, and develops standards in robot ethics.	Making the case for transparent and explainable artificial intelligence
Days 3 & 4	Dr Allan Tucker	A lecturer and researcher at in the Department of Computer Science at Brunel University London with a special interest in the application of artificial intelligence in medicine.	Interpreting and helping to explain how artificial intelligence is being applied in the four jury scenarios.

Scenario witnesses

The four scenarios presented members of the jury with situations with which were unfamiliar to most jurors (e.g. few jurors knew people who had had kidney transplants or who were awaiting criminal trial). For this reason, four "scenario witnesses" were recruited and interviewed. A recording of the interview (three videos, one audio only) were played to each jury as they considered each scenario. The scenario witnesses were asked questions to prompt them to provide relevant information about the importance of decision accuracy (e.g. in stroke diagnosis) and of explanations for each scenario. The four scenario witnesses were:

- Scenario 1: a Stroke UK employee who supports patients with strokes
- Scenario 2: a human resources professional who recruits staff for a major insurance company
- Scenario 3: a patient awaiting a new kidney who also supports others on kidney dialysis awaiting a kidney transplant
- Scenario 4: a professor of criminology.

Some scenario witnesses asked to remain anonymous, and so no names are given above.

The oversight panel

The oversight panel was appointed to help monitor and minimise bias. The panel reviewed the citizens' jury design, and much of the detailed jury documentation, including the jury questionnaires, the videos viewed on day one of the jury, and the slides from the presentations by the expert witnesses, resulting in some changes to these materials. The oversight panel members, chosen for their knowledge of the topic and lack of conflict of interest in any particular jury outcome, were:

- Reema Patel, Programme Manager – Data Ethics & AI, Ada Lovelace Institute
- David Leslie, Ethics Fellow, The Alan Turing Institute;
- Soren Holm, Professor of Bioethics, University of Manchester.

The brief for the oversight panel is available at: www.bit.ly/GMPSTRCCitizensJuries. Each member of the panel completed a questionnaire about bias, published at the same webpage. All three panel members were either "mostly satisfied" or "fully satisfied" that the two juries were designed with the aim of minimising bias, and that this aim was achieved. One panel member expressed concerns over whether the scenarios used were sufficiently nuanced, and of the use of anthropomorphic language (e.g. "AI learns"), and recognised that attempts were made to address these issues.

Citizens' jury project team and funders

The project manager was Dr. Malcolm Oswald, Director of Citizens Juries c.i.c. and an Honorary Research Fellow in Law at The University of Manchester. He worked closely with the jury funders, the jury facilitators, oversight panel, and expert witnesses. Kyle Bozentko, Executive Director of the Jefferson Center facilitated both juries with his colleague Sarah Atwood. They also led the design of the jury process. Chris Barnes and Amanda Stevens recruited and supported the jurors, and jury process.

The juries were commissioned and paid for by the NIHR Greater Manchester Patient Safety Translational Research Centre (a partnership between The University of Manchester and Salford Royal NHS Foundation Trust: Director – Prof Stephen Campbell) and the Information Commissioner's Office. Prof. Niels Peek from the University of Manchester and Alex Hubbard of the ICO constructed the four scenarios and all of the jury questions (with support from Malcolm Oswald). Carl Wiper from the ICO was also involved in commissioning the juries.

Appendix 2: The Jury Questions

The two juries were tasked with responding to four scenarios (on days three and four), each of which had three associated questions. On day five of the jury proceedings, they tackled three general questions. All of these scenarios and questions are set out below. The juries were designed to prepare and enable the jurors to answers these questions.

Scenario 1 - stroke

Scenario

There are more than 100,000 strokes in the UK each year – that is around one stroke every five minutes. About 11% of patients die immediately or within a few weeks as a result of the stroke, making stroke the fourth biggest killer in the UK. Almost two thirds of stroke survivors leave hospital with a disability.

Rapid and accurate diagnosis of stroke greatly increases chances of survival and recovery of the patient. This is highly specialised work which ideally should be done by neuroradiologists with many years of training and experience. However, these experts are not available in each hospital, 24 hours a day, 7 days a week, and in practice diagnosis is often done by non-specialist emergency medicine doctors.

As diagnostic data are accumulated from previous stroke patients, automated decisions systems could provide stroke diagnosis that is fast, and always available in each hospital.

There are three automated decision systems for the NHS to choose from – system A, system B, and system C. Each system uses information about a patient's acute symptoms (for instance paralysis and loss of speech), their medical history, and neuroradiological images (such as CT-scans of the brain) to identify patterns that indicate whether he or she has had a stroke; the type of stroke; its location; and its severity.

- System A – Expert System

  This system uses an algorithm that was developed with help from experienced neurologists and neuroradiologists, and aims to follow the same reasoning as they would do. In practice it does not reach the same level of accuracy as they would, but the algorithm is completely transparent in the way it reaches its conclusions: for each individual case it can provide specific rules that were applied to reach a conclusion. It has an overall accuracy rate of 75%, which is comparable to what most emergency medicine doctors would achieve.

  This means that in 25% of cases, someone might be classified as having a stroke while they were not or vice versa, or the type, location, and severity of the stroke might be misjudged.

- System B – Conventional machine learning

This system uses an algorithm that was established through machine learning from a large set of patient data, collected at English hospitals. This algorithm reaches (human) expert level performance, but it is not very transparent in the way it reaches its conclusions: it can only tell us which features, in general, are important and which are not.

It has an overall accuracy rate of 85%. This means that in 15% of cases, someone might be classified as having a stroke while they were not or vice versa, or the type, location, and severity of the stroke might be misjudged.

- System C – Deep Learning

This system uses advanced AI derived from the same set of patient data as System B. However it has "taught itself" from the data which features were best able to distinguish strokes from non-strokes, and best able to distinguish different types of stroke, their location, and their severity. This algorithm is not transparent in the way it reaches conclusions: it is unable to provide any explanation that is understandable by humans.

However it has an overall accuracy rate of 95%, which is better than human experts perform. This means that in 5% of cases, someone might be classified as having a stroke while they were not or vice versa, or the type, location, and severity of the stroke might be misjudged.

	System A	**System B**	**System C**
Accuracy	75% (A&E doctor's level)	85% (human expert level)	95% (beyond human level)
Transparency	Full explanation	Partial explanation	No explanation

Scenario 1: Questions

1. How important is it for a patient to receive an explanation of an automated decision about stroke diagnosis?

□ Very important

□ Fairly important

□ Not very important

□ Not at all important

□ Don’t know

2. If system C was chosen by the NHS, almost no explanation would be provided. How much does this matter?

   □ Very much

   □ Quite a lot

   □ Not very much

   □ Not at all

   □ Don’t know

   Why (up to three reasons)?

3. Which automated decision system do you think the NHS should choose? Explain the factors affecting your choice.

   System A – Expert System

   System B – Conventional machine learning

   System C – Deep Learning

Scenario 2: Recruitment

Scenario

When running recruitment campaigns, a private sector organisation currently tasks its staff with manually screening all job applications received, and making decisions about which applications to shortlist for interview. This often involves several members of staff reading through thousands of job applications. Sometimes, the organisation receives so many applications that its staff cannot (or do not) properly review every application.

So that the organisation is able to screen every application and free up staff to focus on other work, for future recruitment campaigns, it plans to use an automated decision system to screen job applications and make shortlisting decisions.

The automated decision system will use data about existing employees to be programmed, or to “learn”, which qualities contribute towards being a high performing employee. This may include traditional qualities such as relevant experience, skills and qualifications. But there may also be other qualities that the system is programmed, or “learns”, to treat as important such as particular writing styles, personality traits and personal interests.

The system will then be used to screen the CVs, covering letters and application forms of individuals applying for jobs. It will predict the likelihood of applicants becoming high performing employees. Based on these predictions, the system will decide whether to place them in 1 of 2 classifications:

- *Application accepted*
  Applicants predicted as likely to become high performing employees are placed in this classification and are accepted for interview.
- *Application rejected*
  Applicants predicted as unlikely to become high performing employees are placed in this classification and are rejected for interview.

There are 3 automated decision systems for the organisation to choose from:

- *System A – Expert System*

This system uses an algorithm that was developed with help from experienced recruitment officers, and aims to follow the same reasoning as they would do. In practice it does not reach the same level of accuracy as they would, but the algorithm is completely transparent in the way it reaches its conclusions: for each individual case it can provide specific rules that were applied to reach a conclusion.

When tested on existing data about recruitment, this system was shown to have an overall accuracy rate of 75%. This means that 25% of the time its predictions were incorrect (e.g. predicting that an applicant would be unlikely to become a high-performing employee when in reality they did, or vice versa).

The accuracy of this system is comparable to that of a typical recruitment officer.

- *System B – Conventional machine learning*

This system uses an algorithm that was established through machine learning from a large set of recruitment data, collected by the organisation. This algorithm achieves (human) expert level performance, but it is not very transparent in the way it reaches its conclusions: it can only tell us which features, in general, are important and which are not.

When tested on existing data about recruitment, this system was shown to have an overall accuracy rate of 85%. This means that 15% of the time its predictions were incorrect (e.g. predicting that an applicant would be unlikely to become a high-performing employee when in reality they did, or vice versa).

The accuracy of this system is comparable to that of a very experienced recruitment officer.

- *System C – Deep Learning*

This system uses advanced AI, derived from the same set of data as System B. However it has “taught itself” from the data. This algorithm is not transparent in the way it reaches conclusions: it is unable to provide any explanation that is understandable by humans.

However it has an overall accuracy rate of 95%, which is better than human experts perform. This means that 5% of the time its predictions were incorrect (e.g. predicting that an applicant would be unlikely to become a high performing employee when in reality they did, or vice versa).

	System A	**System B**	**System C**
Accuracy	75% (recruitment officer level)	85% (human expert level)	95% (beyond human level)
Transparency	Full explanation	Partial explanation	No explanation

Scenario 2 Questions

1. How important is it for an applicant to receive an explanation of an automated decision about accepting / rejecting a job application?

   □ Very important

   □ Fairly important

   □ Not very important

   □ Not at all important

   □ Don’t know

2. If system C was chosen by the organisation, almost no explanation would be provided. How much does this matter?

   □ Very much

   □ Quite a lot

   □ Not very much

   □ Not at all

   □ Don’t know

Why (up to three reasons)?

3. Which automated decision system do you think the company should

choose? Explain the factors affecting your choice.

System A – Expert System

System B – Conventional machine learning

System C – Deep Learning

Scenario 3 - kidney transplantation

Scenario

Chronic kidney disease (CKD) is a condition characterized by a gradual loss of kidney function over time. It may be caused by diabetes, high blood pressure, and other disorders. If kidney function gets worse, wastes can build to high levels in your blood, leading to complications like anaemia (low blood count), weak bones, and nerve damage. About 2.6 million people (6.1%) in the UK live with CKD. There is no current cure for CKD.

Eventually, a person with CKD will develop permanent kidney failure, and they will need dialysis or a kidney transplant in order to survive. Dialysis is the removal of waste products and excessive fluids from blood using a machine, and typically needs to happen three times per week for at least 3 hours, placing an immense burden on the patient. A kidney transplant is a better option than dialysis, as patients will have a normally functioning kidney after the transplantation, enabling a relatively normal life.

Because there is a mismatch between demand for and supply of kidney transplants, patients often have to wait for many months (or even years) on dialysis before receiving a kidney transplant. There are currently about 8000 patients on this waiting list in the UK. There has to be a good 'match' between kidney donor and recipient in terms of blood type, immune system, and many other factors in order to maximise the chances of survival of the transplanted kidney. Determining whether donor and recipient match is usually done by experienced renal consultants. However in about 15% of cases the match is not ideal and the transplanted kidney does not survive more than 5 years – and some of them stop functioning within days or weeks. The transplanted kidney is removed and discarded and these patients will have to go back to dialysis.

The NHS wants to deploy an automated decision system for finding matches between kidney donors and recipients so as to make good use of the kidneys and avoid mismatches between donated kidneys and recipients. Each time a new donor becomes available, the system will use data about the kidney and about the potential recipient to determine, for each patient on the waiting list, whether the risk of transplanting the donor's kidney to that patient is 'high', 'intermediate', or 'low'. Only matches that are classified as low risk are eligible for actual transplantation. If the system indicates that there are multiple low risk matches for the same donor, younger patients will be prioritised.

It is hoped that with this system, a larger number of transplanted kidneys will survive longer. There are three automated decision systems to choose from – system A, system B, and system C.

- System A – Expert System

This system uses an algorithm that was developed with help from experienced kidney doctors, and aims to follow the same reasoning as they would do. In practice it does not reach the same level of accuracy as they would, but the algorithm is completely transparent in the way it reaches its conclusions: for each individual case it can provide specific rules that were applied to reach a conclusion.

It has an overall accuracy rate of 75%, which is a little lower than what is currently achieved in practice across the NHS (and lower than that achieved by the top specialists). This means that 25% of the time its predictions were incorrect (e.g. predicting that the kidney would last at least 5 years for the selected patient when in reality it didn't).

- System B – Conventional machine learning

This system uses an algorithm that was established through machine learning from a large set of patient data, collected at English hospitals.. This algorithm achieves (human) expert level performance, but it is not very transparent in the way it reaches its conclusions: it can only tell us which features, in general, are important and which are not.

It has an overall accuracy rate of 85%. This means that 15% of the time its predictions were incorrect (e.g. predicting that the kidney would last at least 5 years for the selected patient when in reality it didn't).

- System C – Deep Learning

This system uses advanced AI, derived from the same set of patient data as System B. However it has "taught itself" from the data which features were best able to distinguish successful matches from non-successful matches. This algorithm is not transparent in the way it reaches conclusions: it is unable to provide any explanation that is understandable by humans.

However it has an overall accuracy rate of 95%, which is better than human experts perform. This means that 5% of the time its predictions were incorrect (e.g. predicting that the kidney would last at least 5 years for the selected patient when in reality it didn't).

	System A	**System B**	**System C**
Accuracy	75% (below human expert)	85% (human expert level)	95% (beyond human level)
Transparency	Full explanation	Partial explanation	No explanation

Questions

1. How important is it for a kidney patient and their family to receive an explanation of an automated decision about why the patient could or could not be matched?

   □ Very important

   □ Fairly important

   □ Not very important

   □ Not at all important

   □ Don’t know

2. If system C was chosen by the NHS, almost no explanation would be provided. How much does this matter?

   □ Very much

   □ Quite a lot

   □ Not very much

   □ Not at all

   □ Don’t know

   Why (up to three reasons)?

3. Which automated decision system do you think the NHS should choose? Explain the factors affecting your choice.

   System A – Expert System

   System B – Conventional machine learning

   System C – Deep Learning

Scenario 4: Criminal justice

Scenario

In an effort to reduce recurring low-level crime, a UK Police Force is setting up a rehabilitation programme to address the underlying issues that lead people to commit multiple minor offences (non-violent / non-sexual) and to prevent them from reoffending.

Artificial intelligence software will be used to identify the individuals to be offered a place on a rehabilitation programme rather than face trial. The individuals selected by the software will be those charged with a minor offence who are considered unlikely to go on to commit a serious offence in the next six months. Individuals that successfully complete the programme will not be prosecuted and will not receive a criminal conviction for the offence they were charged with, and no further action will be taken against them. However, if they fail to complete the programme, they are liable to face prosecution. Individuals offered a place on a rehabilitation programme may refuse it and choose to face prosecution.

The Police Force plans to use an automated decision system to make decisions about who to refer to the rehabilitation programme. It wants to include individuals likely to go on to commit further minor offences but exclude individuals likely to go on to commit a serious offence (violent / sexual).

The automated decision system will use information about an individual's offending history, age, gender and geographical area, as well as any available information about them from local agencies (such as social services) and national databases (such as the Police National Computer). It will analyse this information to predict the likelihood of the individual committing a serious offence over the next 6 months. Individuals will be placed in 1 of 2 classifications based on this prediction:

- *Eligible for rehabilitation programme*

Individuals predicted as unlikely to commit a serious offence in the next 6 months are placed in this classification and referred to the rehabilitation programme.

- *Ineligible for rehabilitation programme*

Individuals predicted as likely to commit a serious offence in the next 6 months are placed in this classification and referred for prosecution for the offence charged.

There are 3 automated decision systems for the Police to choose from:

- *System A – Expert System*

This system uses an algorithm that was developed with help from very experienced Police Custody Officers, and aims to follow the same reasoning as they would do. In practice it does not reach the same level of accuracy as they would, but the algorithm is completely transparent in the way it reaches its conclusions: for each individual case it can provide specific rules that were applied to reach a conclusion.

When tested on existing data about reoffending, this system was shown to have an overall accuracy rate of 75%. This means that 25% of the time its predictions were incorrect (e.g. predicting that an individual would commit a serious offence when in reality they didn't, or vice versa).

The accuracy of this system is comparable to that of an average Police Custody Officer.

- *System B – Conventional machine learning*

This system uses an algorithm that was established through machine learning from a large set of criminal offence data, collected by the police and local agencies. This algorithm achieves (human) expert level performance, but it is not very transparent in the way it reaches its conclusions: it can only tell us which features, in general, are important and which are not.

When tested on existing data about reoffending, this system was shown to have an overall accuracy rate of 85%. This means that 15% of the time its predictions were incorrect (e.g. predicting that an individual would commit a serious offence when in reality they didn't, or vice versa.)

The accuracy of this system is comparable to that of a very experienced Police Custody Officer.

- *System C – Deep Learning*

This system uses advanced AI, derived from the same set of data as System B. However it has "taught itself" from the data. This algorithm is not transparent in the way it reaches conclusions: it is unable to provide any explanation that is understandable by humans.

However it has an overall accuracy rate of 95%, which is better than human experts perform. This means that 5% of the time its predictions were incorrect (e.g. predicting that an individual would commit a serious offence when in reality they didn't, or vice versa.)

	System A	**System B**	**System C**
Accuracy	75 (Custody officer level)	85% (human expert level)	95% (beyond human level)
Transparency	Full explanation	Partial explanation	No explanation

Questions

1. How important is it for an individual to receive an explanation of an automated decision about referral to a rehabilitation programme?

   □ Very important

   □ Fairly important

   □ Not very important

   □ Not at all important

   □ Don't know

2. If system C was chosen by the police force, almost no explanation would be provided. How much does this matter?

   □ Very much

   □ Quite a lot

   □ Not very much

   □ Not at all

   □ Don't know

   Why (up to three reasons)?

3. Which automated decision system do you think the police force should choose? Explain the factors affecting your choice.

System A – Expert System

System B – Conventional machine learning

System C – Deep Learning

General Questions

1. Should automated decisions be explained to individuals? Choose one of the following options.

   □ Some form of explanation should be offered to individuals in all contexts, even if there is a reduction in accuracy.

   □ Some form of explanation should be offered to individuals in certain contexts, even if there is a reduction in accuracy in those contexts.

   □ Automated decisions should be explained to individuals with as much detail as practically possible without reducing accuracy.

   Identify the three most important reasons for accuracy and the three most important reasons for explainability.

2. Should automated decisions be explained to individuals when used in contexts where a human decision-maker would not usually be expected to provide an explanation? Choose one of the following options.

   □ Automated decisions should be explained to individuals in all contexts, even if human decisions would not usually be explained.

   □ Automated decisions should be explained to individuals in certain contexts, even if human decisions would not usually be explained in those contexts.

   □ Automated decisions should not be explained to individuals in contexts where human decisions would not usually be explained.

   Should human and automated decisions require similar explanations?

   What are the three most important reasons for them to be the same and the three most important reasons for them to be different?

3. Other than explanations, what could be done to build confidence in AI-generated decisions?

Appendix 3: Bibliography

1. Shortliffe, E.H. and M.J. Sepúlveda, *Clinical decision support in the era of artificial intelligence.* Jama, 2018. **320**(21): p. 2199-2200.
2. Armour, A., *The citizens' jury model of public participation: a critical evaluation*, in *Fairness and competence in citizen participation*. 1995, Springer. p. 175-187.
3. Jefferson Center. *The Citizens' Jury Handbook*. 2004 [cited 09 Feb 2016]; Available from: http://www.epfound.ge/files/citizens_jury_handbook.pdf.