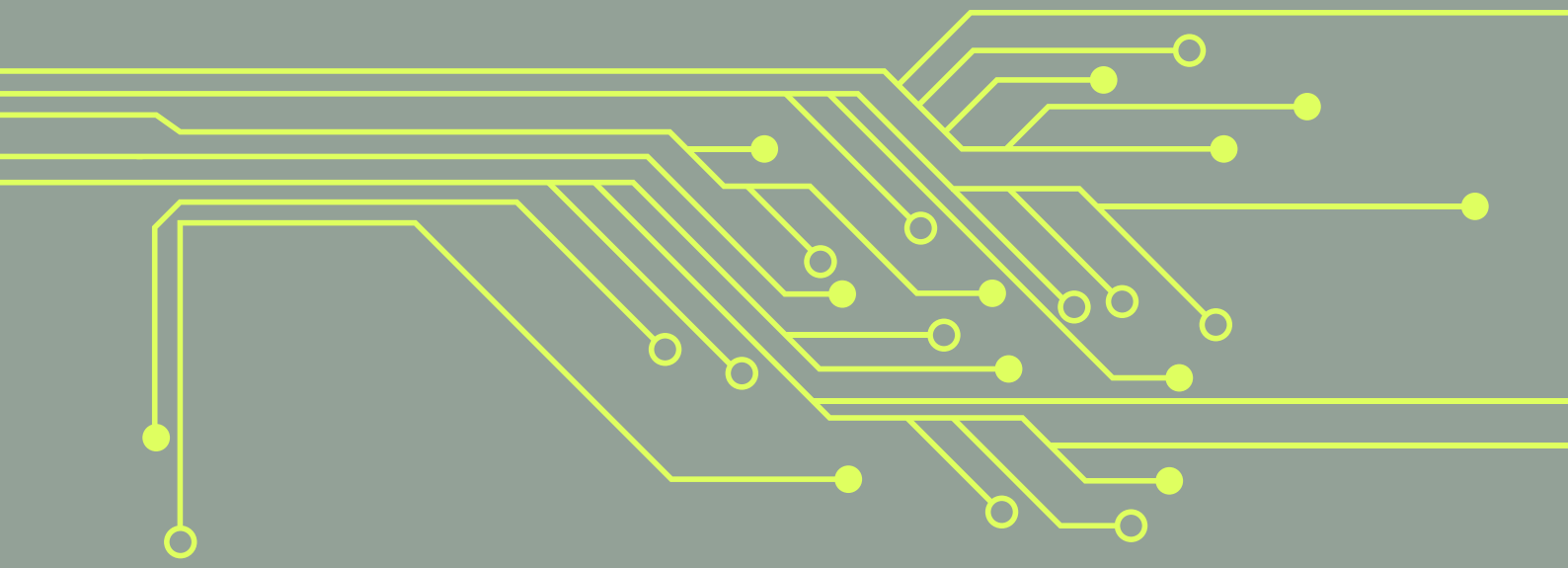
The Forum for Ethical AI

# Democratising decisions about technology
## A toolkit



RSA
21st century enlightenment

The Forum for Ethical AI

# Democratising decisions about technology: A toolkit

# About

## About the RSA

The RSA (Royal Society for the encouragement of Arts, Manufactures and Commerce) believes in a world where everyone is able to participate in creating a better future. Through our ideas, research and a 30,000 strong Fellowship we are a global community of proactive problem solvers, sharing powerful ideas, carrying out cutting-edge research and building networks and opportunities for people to collaborate, influence and demonstrate practical solutions to realise change.

The RSA has been at the forefront of social change for over 260 years. Today our work focuses on supporting innovation in three major areas; creative learning and development, public services and communities, and economy, enterprise and manufacturing.

## About the RSA Tech and Society programme

Tech and Society is a programme of work from the RSA that explores how to increase the agency that people have over the way that organisations design and employ technology. Through deliberative methodologies and innovative conversations, we seek to bring together programmers and citizens, technologists and regulators to place cutting-edge developments in service of the greater good.

## About the Forum for Ethical AI

As decisions are increasingly automated or made with the help of artificial intelligence, machines are becoming more influential in our lives. These machines are generating a range of predictions, such as the likelihood of a defendant reoffending or the job performance of a candidate based on video interview. In some cases, these predictions could lead to positive outcomes, such as less biased decisions or greater political engagement, but there are also risks that come with ceding power or outsourcing human judgment to a machine.

The RSA's Forum for Ethical AI ran a 'citizens' jury' to explore the use of AI in decision-making. Drawing on the model of the RSA's Citizens' Economic Council, we convened participants to grapple with the ethical issues raised by this application of AI under different circumstances and enter into a deliberative dialogue about how companies, organisations, and public institutions should respond.

This report tells the story of that project.

## Acknowledgments

The RSA would like to thank everyone who participated in our citizens' jury, which was invaluable in developing this research, as well the advisory board (see Annex A) who we worked with to ensure robust methodology and ongoing independence in our findings and output. We thank project partners DeepMind for their support in funding the citizens' jury process. Particular thanks should go to Diane Beddoes who was the lead facilitator for the citizens' jury itself and who aided Brhmie Balaram and Kayshani Gibbon in designing the citizen jury process.

Thanks to our RSA colleagues for their ongoing input and help, including Asheem Singh [who also co-authored the report], Josie Warden, Zayn Meghji, Kenny McCarthy, Riley Thorold, Charlotte Holloway, Ash Singleton, Will Grimond, Toby Murray, Amanda Kanojia, Sarah Darrall and Jake Jooshandeh. We are grateful to Studio LP for their design work and to Jocie Juritz for her accompanying animations, available to view on the RSA website.

# Executive summary

## Democratising decisions about technology: Notes from the field

When do citizens think it is appropriate to use automated decision systems? What is the most effective way of engaging citizens with questions on new technologies? How can we ensure information around technology is clear, accessible and actionable?

These were the questions the RSA sought to answer by convening the Forum for Ethical AI. The forum comprised a citizens' jury, that met to deliberate the spread of automated decision systems (ADS) in our private and public lives and in our institutions.

### Key definitions

**AI (artificial intelligence):** The field of computer science dedicated to solving cognitive problems commonly associated with human intelligence.

**ADS (automated decision systems):** Computer systems that either inform or make a decision on a course of action to pursue, about an individual or business that may or may not involve AI.

ADS are common in the private sector. They are also the subject of experiments by public bodies. In the latter, ADS' use strays further into being a *public interest* issue. We sought to better understand what would happen if we therefore treated it as a matter of *public deliberation*.

Some of the key lines of enquiry for our investigation were:

- How might, and how do, citizens discern when it is appropriate to use ADS?
- What sort of reassurance and/or engagement do the public need from those researching, designing and deploying AI?
- What are citizens' red lines with respect to these technologies - and how are these drawn?

## The RSA Forum for Ethical AI: A citizens' jury

The citizens' jury took place over four days, recruiting between 25-29 citizens from a diverse, representative group.

The purpose was to develop understanding of the topic, facilitate an informed dialogue between diverse and sometimes divergent views, and to enrich discussions about the issue with conclusions from citizens.

The RSA also partnered with YouGov to carry out a survey on public attitudes to AI and ADS and so the citizens' jury was an opportunity to further explore the issues surfaced by the survey.

### The YouGov survey: The People vs ADS

The RSA's YouGov survey that accompanied the citizens' jury provided food for thought for ADS enthusiasts and detractors alike – and further insight into the potential benefits of public engagement.

We found that there was public scepticism and lack of awareness concerning the use of ADS; 32 percent of people are aware that AI is being used for decision-making in general, and this drops to 14 percent and 9 percent respectively when it comes to awareness of the use of ADS in the workplace and in the criminal justice system.

On being made aware, the people we surveyed were overwhelmingly opposed to most uses of AI, particularly its use in recruitment and criminal justice; 60 percent oppose or strongly oppose the use of automated decision systems in these domains.[1]

## Citizens' insights

The citizens' jury examined the use of ADS whilst considering behavioural insights, cultural norms, institutional structures and governance, economic incentives and other relevant factors that make up the context into which technology arrives. Scenarios and personae were adopted to help immerse participants in the issues.

1 Figures are from YouGov Plc. Total sample size was 2,074 adults. Fieldwork was undertaken between 16 and 17 April 2018. The survey was carried out online. The figures have been weighted and are representative of all UK adults (aged 18+)

## Technical and ethical questions

With the relative advantages and disadvantages of value-based judgements in mind, and key terms defined, the jury began to ask technical questions about AI and ADS, some of which cut immediately to the ethical heart of the matter:

"How do you retrain an algorithm?"

Juror comment from RSA Forum for Ethical AII

"How can algorithms make accurate predictions about a community if you don't have an inclusive and representative data-set?"

Juror comment from RSA Forum for Ethical AI

"I want to know how far are they going to take this machine learning… Is it going to be 'Terminator-style'?"

Juror comment from RSA Forum for Ethical AI

"What kind of jobs are likely to be replaced by automated decision-making?"

Juror comment from RSA Forum for Ethical AI

Framing the issues around transparency, explainability and clarity particularly resonated with the jury, provoking some insightful questions:

- If ADS is developing and learning on its own, how does it impact on explainability and accountability?
- Is there always going to be a human who is responsible for what decision has been taken by a computer?

### Transparency, explainability, clarity

1. **Transparency**, in the context of ADS, involves being told that there is an ADS being used and how.
2. **Explainability** is about sharing meaningful information about the logic of the ADS in a way most people understand.
3. **Clarity** is about consequences and involves communicating the 'so what' to participants in order to ensure shared understanding of lines of accountability.

The jury agreed that the explainability of ADS should be appropriate to the audience in question, with more onus on the explanation if outcomes are potentially negative. Conversely, gaps in knowledge and explanation may be acceptable if the output is accepted as sufficiently advantageous.

## ADS in action: Recruitment

Case studies were a significant part of the engagement. The Forum for Ethical AI invited experts to present use-cases of ADS in the sectors of recruitment, healthcare and criminal justice.

In the use-case of ADS in recruitment, we asked how the jury would feel about being subjected to ADS with regard to a key career decision, and how these technologies might affect those in the workforce already. Again, the issue of explainability surfaced:

"It's tricky [to evaluate this application] if you don't know how the algorithm has been trained."

Juror comment from RSA Forum for Ethical AI

Citizens were open to the idea of using ADS to determine whether they received a pay rise or promotion. They argued that they would welcome an unbiased assessment of their performance. However, the importance of transparency was underlined, both in knowing that ADS was being used, and in the transparency of the criteria being used by the ADS where possible.

"If a company is predominantly white and male, will it favour these characteristics?"

Juror comment from RSA Forum for Ethical AI

## ADS in action: The NHS

There was broad support for ADS and AI in the NHS,[2] which appeared to be grounded in the high level of public trust in the institution, as well as an awareness of resource constraints and the ability of ADS to alleviate some pressure.

"It's not the system that's biased, it's the people operating the system."

Juror comment from RSA Forum for Ethical AI

The emphasis in the NHS case was not so much on transparency, but on the possibility that data collected for training purposes may find its way into other uses, such as deciding on insurance claims.

## ADS in action: Criminal justice and facial recognition

The importance of trust was raised again in a discussion about the use of facial recognition technologies by the police. The extent of concern over the issue seemed to reflect a generally low level of trust in an institution that some jurors felt strongly about.

"It's not the system that's biased, it's the people operating the system."

Juror comment from RSA Forum for Ethical AI

2 See our upcoming work in partnership with NHS England, which focuses on how the citizens conditions can be met when integrating AI in healthcare.

In all cases where it was raised, citizens felt it important for there to be *proportional* human oversight and involvement where significant decisions were being made. It was thought that *empathy* was also an essential part of delivering such decisions, not to mention the *opportunity to engage in dialogue* over that decision.

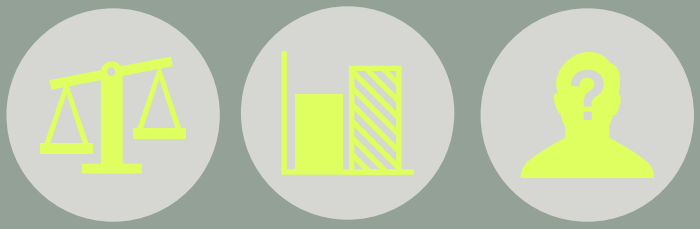
## Recommendations

As ADS proliferates, deliberation is increasingly important. It is both an ethical and a pedagogical tool in a fast-changing landscape. The citizens' jury convened by the Forum for Ethical AI was one example of deliberation in action. That is why, alongside this summary, we present a toolkit for organisations seeking to deploy their own ethical processes around the proliferation of AI.

Consider the following key conditions that jurors would like to see built-in to the processes surrounding ADS in an institutional context.[3]

3 For more on engaging public voice in the design, creation and use of AI, see Brhmie, B., Greenham, T. and Leonard, J. (2018) Artificial Intelligence: Real Public Engagement, London: The RSA.

### Conditions and considerations at the design stage

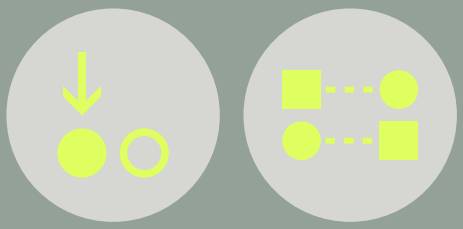


1. To ensure *equality and diversity within ADS*, the data set used for training must be unbiassed.
2. The data gathered for training should not exceed requirements for use.
3. Data should be anonymised to defend against misuse.

### Conditions and considerations at the creation stage

1. There should be *robust policy and legislation* in place to ensure organisational and technical responsibilities with respect to testing, monitoring and audit.
2. In order to be credible, audits should be carried out independently and externally.

### Conditions and considerations at the application stage



1. In the use of ADS, steps should be taken to *ensure accountability*. For example, by establishing a legal requirement for explanation and a right to appeal.
2. Policy and assurance should be empathetic and proportionate with the severity of potential negative impacts, for example where there is a potential for algorithmic prejudice.[4]

4 For more on the intersections of AI and disempowerment, see Singh, A. and Darrall S. (2019) From Precarity to Empowerment: Women and the Future of Work. London: RSA.

The following, moreover, were adjudged to be key questions for deliberation around institutional implementation of ADS going forward.

**Technical Questions**

Is it safe in the way my details are being shared?

Has enough data been used to train the system so that it's accurate?

How do you train algorithms?

How can we protect systems against hackers?

**Accountability-related Questions**

Will I know that an ADS is being used?

How (or is) ADS regulated?

How can I challenge ADS?

What's the government's role in all this?

Who is profiting?

Are there any legal requirements around accuracy?

Who is determining the ethical standards?

**Societal Questions**

What impact will ADS have on broader social structures and interactions?

How does ADS benefit me?

Is this system fair?

How do people without access to tech get represented in the data?

## Lessons in deliberation

Finally, there were learnings for those interested in using wider citizen engagement to surface these complex issues.

There are a number of benefits to using citizens' juries to deliberate on issues related to ADS. The process we undertook seemed to assuage some of the more alarmist concerns around ADS, while surfacing nuanced concerns around transparency, explainability and clarity.

Experts we engaged with felt they had gained a clearer understanding of participants' reactions to ADS, and their reasons for these reactions.

The Forum for Ethical AI's method tallied with most standard models for citizens' juries. There may be aspects of our approach that could be improved upon and so we present its findings in full in the following report as a contribution to the growing body of knowledge in this crucial area.

We recommend further experimentation in this space. In a future increasingly shaped by radical technologies, we can either work towards a scenario in which the public feels without agency and left behind, or a scenario in which informed public opinion is trusted and acted upon by leaders and innovators. The latter would open up a future in which AI and other radical technologies build a new ground for human flourishing beyond private profit. Deliberation has a key role to play in that second future.

The RSA will therefore continue to convene the Forum for Ethical AI and build upon this tranche of knowledge gained at the intersection of technology, public interest and deliberation. We will trial alternative methods of citizen engagement in the context of the ethics of ADS and AI. We will also dive deeper into the institutional case studies discussed in this iteration. A rich seam of people-powered knowledge and human potential lies therein.

# Foreword by Matthew Taylor

There are few more important issues facing us than how to ensure technological change benefits not just those with the resources to exploit it but humanity at large. This paper, combining two topics central to the RSA's work – democratic engagement and technological change - has important implications.

At the micro level of the regulation of specific practices, the report identifies several issues which informed members of the public want to see addressed. The citizens' panels find that for the potential benefits of automated decision systems to be realised, citizens and service users want reassurance about transparency, accountability and human engagement.

At the intermediate level of institutions, the juries revealed not only the desire and capability of ordinary citizens to engage with sometimes complex technology-related questions but also the value that experts gained from citizens' participation. This underlines the need to embed forms of deliberation in the development of technology policy and governance including in major tech companies themselves.

At the macro level our work provides a way of addressing a poignant contrast in societies like ours. On the one hand, scientific discovery and technological change is widespread and rapid, offering the potential for major advances in areas ranging from tackling climate change to improving public services. On the other hand, pollsters find unprecedented levels of pessimism among citizens, while the RSA's own work on economic insecurity reveals that most employees feel technological change will lead to a deterioration in their working lives.

Our citizens' juries on AI may be experimental and relatively small scale, yet they show the need for an approach to technological possibility that is participatory in its method and progressive in its goal. The danger of pessimism is that it become a self-fulfilling prophesy. To ensure that technology not only benefits humanity but is also seen to be doing so we need to inform, engage and empower citizens as guardians of the future. The fascinating outcomes of the RSA's citizens juries shows such an aspiration is both necessary and achievable

**Matthew Taylor**
*Chief Executive, RSA*
*October 2019*

# 1. Designing a deliberative dialogue

Over the course of 2018, the RSA held a series of deliberative dialogues as a means of exploring the ethics of artificial intelligence (AI) with participants representative of society. Our dialogues convened citizens and expert stakeholders (those with knowledge on AI applications within a certain field, or legal frameworks governing its use) to deliberate, reflect and come to conclusions on public policy issues relating to AI.[5]

The deliberative dialogues took the form of a citizens' jury, where participants were asked give their verdict in response to a question posed by the research team, much like in a court of law. The question citizens were asked was: "Under what conditions, if any, is it appropriate to use automated decision systems?"

In this introductory section, we explained why we engaged the public on this topic, what the process entailed, and why we believed organisations should respond to the jury's conclusions. Through sharing our approach, we hoped that others would consider how they could design their own dialogues to engage the public on emerging technology.

## Why we engaged the public on the ethics of AI

Advances in artificial intelligence have spurred vibrant debate about the ethical implications of how emerging technology is developed and deployed. This debate is wide-ranging, encompassing concerns about safety, malicious use, data rights and protection, algorithmic accountability and AI's socio-economic impact. However, this debate tends to happen between expert stakeholders, largely carried out in academic journals and via industry events. Citizen voices have been notably absent from discussions and there appears to be a lack of urgency among businesses and other institutions to seek public opinion. As shown in polling below, there is a lack of awareness among the public of the current use of ADS as well as a scepticism about its use, particularly in criminal justice and recruitment. If citizens lack knowledge and buy-in to technology, this could further enhance an already seen 'tech-lash' against its use, even forgoing any benefits. The aim of the RSA is to break this barrier.

Many of these organisations operate under tremendous pressure; businesses are often racing against competitors to make breakthroughs and launch new products, while public sector bodies have long been constrained by resource and demand for efficiency above all else. Neither is in a position where they feel able to slow down and build in the time it takes to meaningfully engage the public in their processes.

With this in mind, we set out to make a practical intervention. We undertook a citizens' jury to establish the importance of engaging more deeply with the public's views.

## Scoping the dialogue

In consultation with our advisory group, we agreed that the focus of this jury would be on automated decision systems (ADS), which are computer systems

5 We refer to the following publication for our definition of a public dialogue: Bussu, S., Davis, H., and Pollard, A. (2014) The best of Sciencewise reflections on public dialogue. London: Sciencewise.

that either inform or make a decision on a course of action to pursue about an individual or business.[6] Although ADS does not always use AI, these systems increasingly draw on the technology as machine learning algorithms can significantly enhance the accuracy of predictions.[7]

There were a number of reasons why we chose to deliberate on ADS, chief among them their increasingly pervasive use across both private and public domains.

ADS has long been used in the private sector to determine credit ratings, for example, but as machine learning improves predictive power ADS is being experimented with by public bodies for a range of decisions..

In the UK, public bodies are exploring the use of AI to help make decisions regarding planning and managing new infrastructure, reducing tax fraud, rating the performance of schools and hospitals, deploying policing resources, and minimising the risk of reoffending.[8] These systems have been aptly described as 'low-hanging fruit' for government and we anticipate more efforts to embed them in future, albeit with controversy.[9]

Even as machine learning offers promise, the current reality leaves more to be desired in some instances.

For example, facial recognition technology, which is being trialled by police forces in London and Cardiff to help make decisions about whom to stop on the street or in a crowd, has had a remarkably low accuracy rate so far. While systems vary between forces, accuracy rates have yet to surpass single digits, hovering just below 10 percent in the best of cases. The system used by South Wales Police is particularly notable for the number of false positives it has returned – more than 2,400 in 15 deployments since June 2017. In other words, it mistakenly identified innocent people as suspects 91 percent of the time.[10]

Similarly, the system used by the Metropolitan Police during Notting Hill Carnival in 2017 was condemned for its poor performance; out of the 102 suspects it identified, 99 matches – or 98 percent of what the system flagged – proved to be wrong and even the correct matches did not warrant arrest.[11]

While there are questions about whether the police should continue deploying this technology given its lack of precision, the concerns stretch beyond rates of accuracy. These tools may become more sophisticated over time but can still raise alarm because of the contexts in which they're used and the groups that may be disproportionately affected.

6 These computer systems include algorithms, statistical models, and utility functions. This definition draws from the following sources: Karlin, M. (2018) Towards Rules for Automation in Government. Supergovernance, [blog] 2 February 2018; Rahwan, I. (2017) Society-in-the-Loop: Programming the Algorithmic Social Contract. Ethics and Information Technology, 20 (1), pp.5-14.

7 Craig, C. et al. (2017) Machine learning: the power and promise of computers that learn by example. London: Royal Society.

8 The National Infrastructure Commission is considering how the UK maintains its infrastructure, by using data and AI to predict when repairs will be required. HMRC to use AI to enhance decision-making in casework. The Cabinet Office's Behavioural Insights Team is testing machine learning to predict, and rate, the performance of schools and GPs to make decisions about inspections. Kent Police Department is using PredPol, a crime forecasting tool premised on machine learning, to predict hotspots for criminal activity and make decisions about where to patrol accordingly. Durham Constabulary is deploying a Harm Assessment Risk Tool (HART) to help make decisions about whether to refer arrestees to their Checkpoint programme, which aims to reduce reoffending.

9 Martinho-Truswell, E. (2018) 'How AI Could Help the Public Sector.' Harvard Business Review, 29 January.

10 Carlo, S., Krueckeberg, J. and Ferris, G. (2018) Face-off: The lawless growth of facial recognition in UK policing. London: Big Brother Watch.

11 Ibid.

For example, some civil liberties groups have called into question the decision made by the Metropolitan Police to repeatedly test dubious technology at a popular event in the black community. These groups are prompting us to consider whether this technology is ultimately just, or if it is further entrenching existing biases and exacerbating inequalities.

Other uses of ADS have been criticised for being dehumanising. The academic Virginia Eubanks reveals instances where the data collected by these systems has been very intimate and served to intensify state surveillance of the poor in particular. She specifically reviewed the use of ADS as part of the welfare state in the US and concluded that these systems were problematic because they enable the ethical distance needed "to make inhuman choices about who gets food and who starves, who has housing and remains homeless, whose family stays together and whose is broken up by the state."[12]

While it is important to acknowledge this darker side to ADS, there may be legitimate reasons for businesses and public bodies to press ahead with its use. Obvious benefits could include:

- Time and cost savings that reward customers and citizens
- Fairer outcomes for citizens. Data-driven technology may actually be more consistent and objective than humans when making decisions.

Reflecting on this, we wanted to better understand what, if anything, would make the public feel more comfortable with the rising use of ADS.

How do citizens discern when it is appropriate to use ADS? Does it depend on motivations, the circumstances it is used in, or who (or which sector) it is deployed by? What sort of reassurance are they seeking from the chain of people researching, designing and deploying ADS (and AI more broadly) if they're going to trust these systems to inform significant decisions in their lives? What are their red lines when it comes to the use of this technology, and how do they determine these?

## What the process entailed

The jury took place over four days in 2018 (12 and 13 May, 2 June and 13 October 2018). We followed a well-established process (see Figure 1) as well as general principles for recruiting a diverse range of participants and ensuring that there was a variety of views in the room. We had between 25-29 citizens involved throughout the deliberation, drawn from across England and Wales and representative of a range of ages, abilities, ethnicities and socio-economic backgrounds. We also aimed for a more or less equal split between participants in terms of their attitudes towards technology, accounting for positive, negative and neutral perspectives before they began the process.

Some may question how impactful a public dialogue can be given the size of the groups assembled, particularly for citizens' juries like this one. However, dialogue serves a different purpose than a survey, which draws on a statistically significant sample of the population to gauge opinions. As explained by Involve, a leading organisation for deliberative democracy in the UK, dialogue enables a diverse mix of participants with an array of views and values to discuss the issues. The process means participants learn about the issue (eg from written

12 Eubanks, V. (2017) Automating Inequality: How high-tech tools profile, police, and punish the poor. New York: St. Martin's Press.

information and experts), listen to and share with one another as they further develop their views, draw carefully considered conclusions, and communicate those conclusions to inform the decision-making of policymakers and other relevant stakeholders.[13] Deliberative processes are particularly useful in facilitating a nuanced discussion over contentious issues. They ask participants to create a shared space of respect and trust where they can share personal stories, uncover ideological assumptions and empathise with one another, deliberating as a citizen on behalf of society rather than as an individual.

At the outset of the project, we partnered with YouGov to run an online survey on public attitudes towards AI and ADS,[14] and this jury provided an opportunity to delve deeper into the issues that were surfaced by the survey. Our jury also directly connected people with decision-makers with the aim of influencing their approach to how we manage the expansion of ADS.

Using citizens own words, the jury involved:

"Deeper understanding of what's involved – what AI is and ADS. In a way, I have more questions and uncertainty because we learnt more, I have new questions because I have deeper insight into what is going on."

"We're a representation of society, we've got a diverse group and we're giving you public honesty."

"There's been a variety of inputs from the group and I hope everyone has had a chance to speak. Hearing different things from different people means I've factored in different ways of thinking about these things."

"We have developed our thoughts and our opinions have converged."

## The citizens' jury approach

The citizens' jury method has been part of the deliberative democracy toolkit for some time. The form is relatively well-established. A small group of randomly selected citizens, representative of the demographics in the area, come together to reach a collective decision or recommendation on a specified policy issue through informed deliberation. While there is not a definitive definition or checklist of criteria that determine whether a process can be considered or called a citizens' jury, it is possible, based on the work of Involve and others, to identify conditions of best practice within this field. Table 1 outlines these conditions and compares them with the activities undertaken by the RSA.

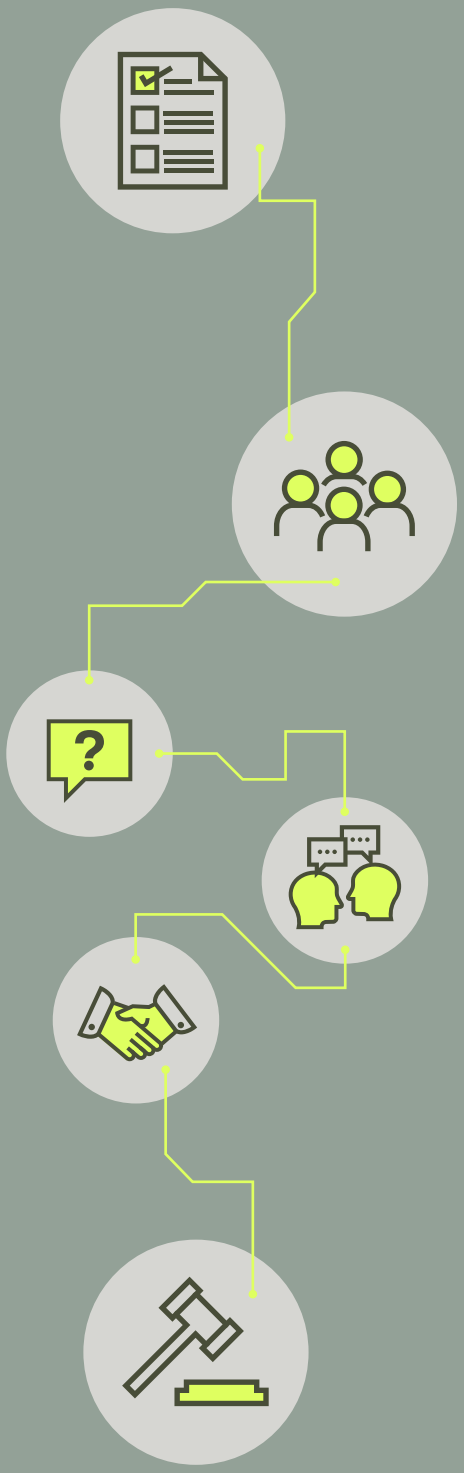
13 Bussu et al. (2014) Op cit.

14 For the results of the survey, please see the Forum for Ethical AI's position paper: Artificial Intelligence: Real Engagement

## Table 1: Conditions of a citizens' jury and comparison with process undertaken by RSA

| Condition | Description of best practice | Description of RSA activity |
|---|---|---|
| Small group, representative of the population | Jurys usually comprise of between 12–24 people. The group size should be small enough that genuine discussion may take place amongst the participants and that all voices are heard, but large enough that the group can suitably reflect the diversity of the community it is representing.<br><br>Jurors are usually selected at random and should broadly resemble the demographic of the community they are representing, eg a national population, using characteristics such as gender, age, ethnicity etc.<br><br>Jurys usually last between 2–7 days. These days may be consecutive or at intervals, depending upon the topic being discussed and jurors involved. For more complex issues a longer jury may be required so that jurors have time to learn about the subject being discussed. | The Jury comprised of 29 jurors.<br><br>The jurors were selected to broadly reflect the makeup of the population of England and Wales on the criteria of gender, age, socio-economic backgrounds. The jury panel was not strictly representative of this population.<br><br>The RSA engaged a market research company to recruit the jurors.<br><br>The jury took place over 4 days: One long weekend (Friday evening until Sunday), and two Saturdays. The first two sessions were held one-month apart. The final session was held four months later.<br><br>In addition to the jury, two workshops were held with a specific demographic considered to be more exposed to the issues discussed, young BAME males. The citizens' jury format is not the best mechanism by which to hear from minority groups as the roughly representative sample of the wider population means their unique experiences are likely to be drowned out within group discussions. |
| Clear question with defined scope | The issue that the jury will deliberate should have a defined scope and is usually set as a clear question.<br><br>Jurors usually provide a report at the end of the deliberation outlining their recommendations. | The jury was set a clear question to frame the deliberation.<br><br>Jurors' deliberations were summarised by the RSA after each session and fed back to them for agreement at the start of the next session.<br><br>The final report has been written by the RSA and is based upon the summaries agreed with the jurors. |
| Independent facilitation | The commissioning body should engage an independent, professional facilitator(s) to deliver the jury process.<br><br>The commissioning body will not be directly involved in the delivery of the jury but will be involved in setting the question to be answered. | The RSA independently designed and managed the research project and jury process, funded by DeepMind.<br><br>The RSA delivered the jury process and engaged an independent and experienced facilitator to support with the design of the process and to host the jury days. |

| Condition | Description of best practice | Description of RSA activity |
|---|---|---|
| High quality and unbiased information | Jurors should receive information and/or opinions on the issues being discussed from expert witnesses.<br><br>This information and the expert witnesses should represent the spectrum of key perspectives on the issues being discussed.<br><br>The identification and selection of the information or opinions delivered, and the expert witnesses delivering them, should ideally be overseen by an independent advisory group whose role is to protect against bias.<br><br>Time should be allocated for jurors to ask questions to the expert witnesses.<br><br>Jurors should be provided with enough time during the process to study the information and issues being discussed. Depending on the complexity of the issues, this may require more or less time. | Jurors heard from 24 expert witnesses across the course of the sessions. These expert witnesses explained key information to the jurors, who were then able to question the witnesses.<br><br>The expert witnesses and the subjects they covered were selected by the RSA with support from an independent advisory group. This group was selected by the RSA and included experts in the deliberative process and in the content of the issues discussed. |
| Democratic discussion and decision-making | Skilled, professional facilitation and moderation to ensure that all participants are included in discussions and that final decisions are arrived at democratically.<br><br>Facilitators and moderators should remain neutral on the issues being discussed.<br><br>During the process jurors should be educated about uncovering their own biases and about the importance of critical thinking in the jury setting. | Facilitation was led by a professional facilitator experienced in running deliberative discussion.<br><br>Support facilitation was provided by RSA staff. These staff had undergone training from the lead facilitator on remaining neutral in discussions whilst also guiding discussion towards decisions.<br><br>Time was allocated within the process for jurors to reflect on and uncover their own biases and to consider how to ensure these did not impact their discussions. |
| Remuneration for taking part | It is best practice to pay jurors a stipend or per diem (eg between £80-£120) plus travel and/or accommodation expenses if required. This increases the likelihood of individuals being able to take part or engage in the full process. | Jurors were compensated and had their travel, accommodation, breakfast and lunch paid for. For taking part in the first session (Friday to Sunday) the jurors each received £175, for the second one-day session they received £80 and for the final one-day session they received £100. |



## Figure 1: The RSA Forum for Ethical AI's citizens' jury – our journey

**1. Defining the problem**

Jurors within a citizens' jury are asked to give their verdict, or answer, in response to a question, much like in a court of law. In this case, the jurors will be answering a specific question that poses a problem, in order to inform government and corporate policies. The question they will be asked is, 'Under what conditions, if any, is it appropriate to use an automated decision system?'

A citizens' jury is best used to resolve contentious issues (with many trade-offs and more than one probable or realistic response). The answer is not pre-determined by those convening the jury.

**2. Selecting the jury**

A small group of citizens are randomly selected from a 'community'; in this case, 25-29 citizens from across England and Wales. This group is not intended to be representative of these national communities, but is recruited to be as diverse as possible to capture a wide range of views.

**3. Deliberating as a jury**

- Citizens spend a period of time learning about and discussing the problem from many different angles. Similar to a traditional jury, expert witnesses are summoned to enhance citizens' understanding of the different elements to the problem.
- Citizens are then asked to enter into an open dialogue, commit to listening to others, and provide responses with consideration for the wider community (in contrast to focus groups and most consultations where individuals are asked for their own opinion). This is to encourage citizens to strive towards a consensus and/or a compromise in the best interests of society, rather than for themselves as individuals.
- Finally, the jury draws its conclusions, providing an answer to the question set and a clear steer or recommendation(s) for government, businesses, and civil society organisations to take forward. This answer will take the form of a statement.

**4. Acting on the answer**

Institutions and organisations, including companies, with influence and authority typically respond directly and publicly to the citizens' conclusions. In this instance, the RSA will be holding an event in autumn 2018, reconvening the citizens, so that they can have the opportunity to hear, and discuss, reflections from key stakeholders on their conclusions.

Source: The RSA set out this process with reference to Participedia.net and: Chwalisz, C. (2017) 'The people's verdict: Adding informed citizen voices to public decision-making'. London: Policy Network

## Why organisations and institutions should respond to the jury's conclusions

From our accompanying survey, we know that only 32 percent of people are aware that AI is being used for decision-making in general, and this drops to 14 percent and 9 percent respectively when it comes to awareness of the use of ADS in the workplace and in the criminal justice system.

On the whole, people are not supportive of the idea of using AI for decision-making, and they feel especially strongly about the use of ADS in the workplace and in the criminal justice system (60 percent of people oppose or strongly oppose its use in these domains).

The inherent scepticism of the public should not be taken lightly. Regardless of the benefits of AI, if people feel like they are more likely to be victimised by technology rather than empowered by it, they may resist innovation – even if this means that they lose out on those benefits.

The NHS England Care.data data-sharing project is a good illustration of a scheme that had majority expert support and huge potential public benefit that had to be shelved after widescale condemnation of the lack of public awareness and consent.[15] More recent plans for Local Health and Care Records (LHCRs) led by five exemplar sites, is placing public trust at the heart of their approach, referencing the valuable insight of citizen juries and the need for "a deep and genuine conversation" with members of the public.[16]

We are arguably already in the midst of a 'techlash', but the public's unease with big business also extends to government.[17] According to the 2019 Edelman Trust Barometer, nearly half of the general public distrust government.[18] As we have argued previously, there appears to be a need for a radical overhaul of the way in which organisations and institutions include and devolve power to citizens over crucial decisions, such as how to (if at all) roll out new technology that has the potential to be life-changing for many.

We advocate for a meaningful realisation of what it means for society to be 'in-the-loop' or, in other words, for the public to be more involved in the development, deployment and oversight of these systems. We produced this toolkit because we want to encourage organisations to take into account what the public expects at this early stage in their familiarity with AI and ADS. We hope that in doing so, organisations will recognise the need for their users to feel more in control of the technology to which they are subjected and be able to demonstrate that they are genuinely deserving of the public's trust.

15 Triggle, N. Care. Data: How did it go so wrong? BBC News. Available at: www.bbc.co.uk/news/health-26259101

16 Singh, I. (2019). So, what is a Local Health and Care Record anyway? NHS Digital. Available at: digital.nhs.uk/blog/transformation-blog/2019/so-what-is-a-local-health-and-care-record-anyway and Fulton, J. (2018). One London local health and care record exemplar – creating the data sharing ecosystem. Available at: digitalhealth.london/one-london-local-health-and-care-record-exemplar-creating-the-data-sharing-ecosystem

17 "Techlash": Financial Times word of the Year. Available at: www.ft.com/content/76578fba-fca1-11e8-ac00-57a2a826423e

18 For more information see: www.edelman.com/trust-barometer

### Box 1: Takeaways on designing a deliberative dialogue

There are many ways to engage the public in a dialogue about AI and ADS, each with their own benefits and trade-offs. Following on from the RSA's Citizens' Economic Council, the Bank of England has decided to set up citizens' panels; each of the Bank's 12 regional agents will hold two meetings a year with 24 citizens, so that they can listen to peoples' experience of the economy across the country.[19] However, we chose to use a jury format, setting a clear question for discussion so that we could sharpen the focus of conversation, moving away from speaking about the ethics of AI in general terms to gather some practical conditions for the use of ADS. Dependant on the objectives of the dialogue - whether to gain understanding of the publics' position on an issue, generate new ideas, break deadlock or increase participation – there are various factors to consider within its design.[20]

**Who needs to be involved?**
A good point to begin thinking about this is to ask who the technology will affect and whether the objective can be resolved in one conversation or will require continuous feedback. Can a group of 12 from one geographical region resolve the issue in a week, or does it require 100 participants from across the UK meeting several times a year?

**What evidence needs to be provided?**
Making sure citizens feel comfortable and capable of engaging in dialogue is key, so if the group needs to reach a technical level of understanding before making recommendations, they will need to go through a learning phase. Things to consider are: what types of information expert witnesses should provide; will citizens need to do 'homework'; what kind of activities will best help people learn and keep them engaged? Organisations should also be aware of potential bias in the information they present, and we would recommend working with an independent facilitator and advisory group to help design and run a public dialogue.

**How do the results get used?**
This is potentially the most crucial point and links back to the objective. When would deliberative dialogue be most useful, is it during the design, creation and/or application stage of the technology? What is the tricky question that needs answering and how will the results be incorporated into improving the system? The danger here is if the results of the dialogue are not acknowledged or acted upon then it could harm rather than engender public trust. There is more on this in the latter stages of this report.

19 Bank of England. 2019. Available at: www.bankofengland.co.uk/outreach/citizens-panels

20 See Nesta's report Evidence vs Democracy for more detail different types of public dialogue and design considerations. Available at: www.alliance4usefulevidence.org/assets/2019/01/Evidence-vs-Democracy-publication.pdf

# 2. Citizen insight: Discussing radical technologies

Our starting point for designing this jury was a belief in the importance of examining the use of automated decision systems within their broader social and economic context. In practice, this means that the jury took into consideration behavioural insights, cultural norms, institutional structures and governance, economic incentives and other contextual factors that have a bearing on how ADS might be used.

In this chapter, we provide an overview of how we communicated with citizens about AI and ADS, and shared insights into how they learned about and engaged with relevant concepts throughout the process. As this process was experimental in nature, this chapter necessarily leans into certain details of the experimental parts of the process.

## How citizens negotiate the difference between how people and machines make decisions

### Kicking off: Making decisions 'as people' vs decisions made by machines

There are many ways to kick-off a citizens' jury process. We began by reflecting on how we make decisions as people, prior to considering how machines might play a role in decision-making. In small groups, citizens deliberated on the differences between making a decision based on fact and making a decision based on emotions and/or values, followed by a discussion of both the positive and negative implications of emotional or value-driven decision-making.

**Box 2: Prompts for discussion on how we make decisions as people**

- What factors are important to you when you make a decision?
- What information do you draw on to help you (and from where does this come)?
- How do you use this information?
- How do you take into account the possible consequences of your decision?
- Does the way you make a decision (eg the information you look for, the weight you give to the possible consequences) change for different types of decision? How? Why?
  - PROBE: Explore the differences between a decision about which bank you would use to open a savings account and whether or not an elderly parent should go into residential care.
  - PROBE: Explore the differences between a decision about which hospital to have an elective operation in and what to serve to guests coming over for supper.
- If someone said: Explain your decision to me, what would you say?
- Is making a decision the same as acting on that decision? Why/why not?
- What role do our values and emotions play in decision-making? Why might it be useful (or not useful) to make decisions informed by our values and/or emotions? Why/why not?

Citizens considered how they had arrived at recent decisions, recognising that there were a variety of approaches to decision-making. For example, some citizens admitted to being more impulsive, while others are more methodical. Even among those who made decisions based on 'gut', there were differences; some instinctively made a choice and then mulled over the rationale, whereas others brought their intuition into play if logic wasn't enough to convince them of a particular course of action. Those who take a more methodical approach noted drawing on insights similar to the information used or produced by an ADS, including facts or data, risk factors and/or previous history.

When discussing making decisions about other people, some felt strongly about trusting one's instincts while others argued that turning to data could be advantageous (eg matchmaking websites for finding a compatible partner). There was also acknowledgment that decisions could be made out of habit and be influenced by both individual and societal values or contexts (ie decisions about how to treat women). Evidently, how a person may initially react to the idea of a machine supporting the decision-making process depends on how they tend to make decisions as individuals.

However, in spite of these differences, it was widely accepted that there wasn't a single approach to decision-making that is fail-proof. It was readily agreed that mistakes are made, although a few citizens pointed out that 'wrong' decisions can also be an important part of an individual's learning process or have an unintended, but positive impact. When asked about whether there was a certain threshold of confidence in a decision that must be achieved to enact it, citizens found it challenging to specify a level – it is not easy or typical to measure one's confidence or calculate the probability of a decision being right before it's agreed. In some cases, citizens try to ensure that they're making the best decision by consulting others, particularly if the decision is significant (for example a life-changing decision about a career transition).

#### Introducing the role of automated decision systems

From that human standpoint, we moved into the more complex definitional questions.

### Box 3: Defining key terms with our jury

With the help of Jess Montgomery, Senior Policy Adviser at The Royal Society key definitions were explained to the jury:

**Data**

Each of us creates different types of data every day:

- The words or phrases we search for in Google
- The things we buy, and the points we collect using loyalty cards
- The services we use from corporations or government, among others.

When collected for use by computers, these data sources might be in the form of text, numbers, audio files, or images (and more).

The last decade – even the last five years – has seen huge growth in the amount and type of data that we create and capture.

**Algorithms**

At their core, algorithms are sets of instructions, which tell a computer how to carry out a particular task.

We can think of algorithms as being like a recipe – a series of instructions that result in an output.

In recent years, computer scientists have made a number of technical advances, which means that the algorithms we have today are more powerful than they were previously.

These advances have progressed a field called machine learning.

**Machine leaning and AI**

Machine learning is the technology that allows computers to learn from data.

Previously, algorithms had to be programmed with each step in order to carry out a function. If we stick with the recipe analogy, computers had to be programmed with each ingredient, how to put them together, and exact cooking instructions.

Recent technical advances mean that today's machine learning algorithms can learn how to solve a problem, if given enough data.

We already interact with many machine learning systems each day, without necessarily realising it. For example, if we think about Siri, Cortana, or any of the other virtual personal assistants you have on your phone, they are based on a technology called voice recognition, which has utilised machine learning to improve hugely in recent years.

When most people think about AI, they usually think of systems that mimic human intelligence. This isn't what machine learning does.

- Machine learning can be thought of as a form of 'narrow AI' – it creates functions that have a specific type of intelligence, or can carry out a specific task intelligently.
- This does not, however, match the suite of capabilities that can be carried out by people. Experts think such systems are at least decades away from being created, and it is not clear that it will ever be possible to create them.

Automated decision systems are defined as computer systems that either make or help humans make a decision by generating predictions (such as different probabilities of risk). These predictions or other outputs are based on data analysed by an algorithm. It was clarified that not all systems use AI or machine learning algorithms, but systems increasingly may do so in order to make increasingly more accurate predictions. We also made the point that it is rare for decisions to be fully automated, humans ultimately still have oversight in many cases.

These systems are sometimes referred to as 'decision-support systems' or data-driven systems, but we are using Automated Decision Systems, the term favoured in GDPR and in ICO guidance.

## Clarifying AI and ADS in practice

Our citizen engagement produced a rich seam of insight for those interested in convening deliberative processes around AI and ADS. Based on the questions citizens generated for our experts about what AI and ADS are and do, we have a better understanding of what businesses and other organisations should keep in mind when they're clarifying or explaining the use of these systems to their users.

**Box 4: Citizens' technical questions about AI and ADS**

- Is there someone who writes a programme at the beginning (or, in other words, how do we maintain control over these systems)?
- Is machine learning another word for AI?
- How do you retrain an algorithm?
- How do you decide what data goes in in the first place, so that you get the right outcomes?
- How can algorithms make accurate predictions about a community if you don't have an inclusive and representative data-set?
- Are datasets refreshed?
- How do we best protect ADS from hacking? The more [technical systems] we introduce, the more vulnerable we are (eg to data hacking in hospitals).
- Jessica (from Royal Society) said that AI is expert at one thing (narrow intelligence) – is that likely to change in the future, so that intelligence becomes wider (general)?
- I want to know how far are they going to take this machine learning, how far can they go? Is it going to be Terminator-style?
- With regards to machine learning: What if it learns in a way it wasn't intended to?
- Can you stop AI from learning at a certain point so we can maintain control over it?
- What can be done to stop AI from becoming malicious (or being used maliciously)?
- What kind of jobs are likely to be replaced by automated decision-making?

Some of these questions could also be interpreted as legal and regulatory questions.

We also have insights concerning the challenges posed by citizens. It was difficult for some citizens, particularly of older ages, to move past the idea that with the emergence of AI, machines no longer need to be programmed. Some struggled with this concept because they found it difficult to imagine how we might retain control over these systems if they are self-learning. The following were highlighted as key areas of concern and control:

- Clarity about what citizens have control over, such as the data the system trains itself on.
- Accountability for these systems, noting that some organisations are making efforts to explain why certain data (ie of a sensitive nature) has been used in an ADS and demonstrating that they have considered what the implications might be.
- High quality testing before use, ie at that point the system will stop being trained on new data.
- With regard to hacking, organisations can confirm that they are adhering to data protection laws to keep users' information safe.
- In some cases, it may be possible for organisations to go the extra mile with transparency and disclose what kind of data they're encrypting, although this is something that regulators can also scrutinise on behalf of users.

**Box 5: Citizens' legal and regulatory questions about AI and ADS**

- Who's in control of deciding which data is used to train the machine?
  - How is it determined (by developers and/or users) that the data is genuine, unbiased, and legally obtained?
- If the machine is learning from the data, what if it comes up with outcomes we don't like or find morally impermissible (eg if left-handed people were found to default on their loans more)? We might be morally uncomfortable about this, so how do we plan for that?
- How do we best protect these automated-decision systems from being misused or abused?
- How is AI regulated?
  - What existing regulation applies to machine learning?
- Has the law kept pace with advances in technology?
  - Does there need to be more investment in changing our laws to adapt to this new technology?
  - How can you make new laws when you aren't sure what will happen?
- How can the law be used to facilitate positive uses of AI?
- What is an example of regulation from another industry that is comparable?
- What, if any, are the risks to over-regulating AI?
- Are there any laws which govern ADS specifically? What safeguards are in place?
- Are there any legal requirements about the thresholds of accuracy for acceptability, and who decides that?
- Can we opt out of being subjected to ADS?
  - What happens if you do opt out?
- Who is regulating ADS?
- How can we be sure that regulators are working in our best interests (as opposed to the interests of business)?
- Who is deciding that we need this, who is driving it, and who is deciding who is profiting?

A number of citizens expressed that they wanted more agency over ADS and whether they were subjected to it, prompting several questions about the possibility of opting out. It would be helpful if organisations could clearly communicate whether it is possible to opt out of being subjected to ADS, what the implications of opting out would be, and how to opt out if you can and do choose to. For example, one of the experts pointed out that private sector companies are not obliged to provide a service to all and may not serve an individual who has opted out of the use of ADS.

Another noted that machines make many decisions that affect us (some of which may not be considered significant), so it may not be possible to 'meaningfully' opt out or to opt out entirely.

Citizens also expressed concern at the presence of systems of redress: Organisations could clearly outline their processes for redress and accountability if a user wants to challenge or appeal a decision. They indicated that they'd like more clarity about how to make a complaint, and how to escalate their concerns if they are still unhappy with the resolution offered by the organisation (or even the regulator).

### Distinguishing the decision-making capabilities of machines

We captured some of the initial reactions that citizens had to the use of ADS before collectively drawing up a list of pros and cons.

There was general agreement that ADS is faster, more efficient and potentially more consistent than humans, there was some scepticism about whether ADS is less biased than human decision-makers.

> "If it's only learning from what humans have previously done, wouldn't the biases be inherent in the system?"

> "Human error does exist in machines as well."

There was also concern about whether making decisions primarily based on statistics (probabilities of risk) was either accurate or morally acceptable. Due to doubts about how accurate an ADS could be, citizens also questioned whether it was appropriate for any decision to be fully automated.

> "So, everything's based on statistics? Statistically someone is more likely to do that, statistically you're likely to do that… that's bullshit!"

> "It makes predictions which don't take the cultural or social context into the picture."

> "If they [ADS] were 100 percent accurate, sure, let them decide. But if they're not, I wouldn't be happy for them to make a final decision."

> "Can a machine ever be 100 percent accurate if a human designs it?"

Some questioned whether, rather than leading to fairer outcomes, ADS may inadvertently result in the opposite by treating everyone homogenously.

> "It puts everybody on a level playing field. But that's not fair. To be fair we need to treat people differently. Human decision making takes into account lots of different factors like gender, jobs and all that. Humans take into account many factors when trying to create equality."

> "It's narrow-minded – it can only consider the data it has [whereas humans can use wider judgment]."

Citizens wondered if ADS is simply informing or assisting decision-makers or if it is really making a decision because the human is too reliant on its output. Relatedly, there was some apprehension about whether ADS was de-skilling humans and allowing people to evade responsibility.

> "It's not informing; it's making decisions for you. People are lazy."

Citizens expressed that trust in the system would depend on how it was designed, whether there was transparency about the system, and if was the outcomes were monitored and assessed in order to ensure that these were the 'right' outcomes.

> "When calculators came out, we would add it up again to check that the calculator was right."

## Table 2: Citizens' views on the pros and cons of ADS

| Advantages of ADS | Disadvantages of ADS |
|---|---|
| **Function** | |
| • Potential to be more accurate<br>• Objective with data available<br>• Could be unbiased<br>• Not corrupt (can't be bribed or persuaded)<br>• Consistent (because it is unemotional, decisions are based on the facts)<br>• Cheaper | • Could be hacked into/open to abuse<br>• Could be biased based on what data was used<br>• Could result in critical system failure, malfunction at the wrong moment |
| **Deployment** | |
| • Works 24/7<br>• Labour-saving<br>• Doesn't need personal care<br>• Faster to learn and more efficient than humans | • Deskilling of decision-makers<br>People assuming that the computer is right [when it isn't]<br>• Loss of human contact, for example not seeing GPs [for a diagnosis] or other decision-makers<br>• Wouldn't take into account mistakes or understand the human condition; may not take into consideration full context<br>• There is no reasoning with a computer<br>• Not empathetic |
| **More broadly** | |
| • Potentially available to rich and poor | • Could become exclusive to those who own the technology<br>• Corrupt governments could control/ manipulate populations with ADS |

## Initial considerations about the transparency, explainability and clarity of ADS

### Box 6: Citizens' questions about the transparency of ADS

- How do we know whether an ADS has been used?
  - Should there be a label or symbol of some sort to indicate the use of ADS?
- How do we know what sort of information has been taken into account by the ADS?
  - Should it be explicitly stated at the end of [providing] a decision, in the terms and conditions, etc??

Finally, let us consider three concepts. When convening deliberative processes around ADS, transparency, explainability and clarity are central to the discussion.[21] Some definitions:

1. Transparency (to be told that there is an ADS being used).
2. Explainability (to be told meaningful information about the logic in a way most people understand).
3. Clarity about the consequences (the 'so what' in order to ensure accountability by complaining or lodging an appeal as appropriate).

### Box 7: Citizens' questions about the explainability of ADS

- Why can't we explain automated decision systems?
  - Is every computer system explainable regardless of whether it uses AI or not?
- What does GDPR mean for explainability?
- Who decides what information is provided in an explanation?
- How accessible does an explanation need to be?
- Should all automated decisions be explainable (ie considering associated costs and burdens)?
- I don't get how explainable things have to be at the moment. What is the status quo?

It is eminently possible to explain some automated decisions, but others might be more difficult to clarify (particularly in detail) depending on the predictive technology being used (for example, ADS using deep learning) and how complex the system is.

There could also be legal reasons why an explanation is withheld (ie IP issues). GDPR, which sets a higher bar for organisations using our data (ie to make automated decisions), requires that individuals are provided with 'meaningful logic' about the decision.

There is debate about what meaningful logic entails, but according to one expert, it is usually possible to work out a simple explanation.

21 The method we used to explore these concepts was a series of expert talks followed by a 'World Café' session. See World Café method, available at: www.theworldcafe.com/key-concepts-resources/world-cafe-method/

In our session, it was also acknowledged that GDPR does not require explanations for many of the systems discussed in the jury, such as a healthcare professional using a machine for advice about a diagnosis. This is because the right to explanation only applies to solely automated decisions, and at the moment these systems are semi-automated or simply providing decision support. Citizens would only have the right to object and challenge a system (based on the explanation) if it were made by solely automated means, such as a decision about care insurance premiums.

In terms of the explanation itself, relaying the technical 'ins and outs' of an ADS to the general public isn't very practical; technical details tend to better serve as a description. Organisations are obliged to provide an explanation in accordance with the audience, and thus ought to take into account the technical knowledge of the person to whom they are giving an explanation.

In cases where an explanation isn't possible, we may still be able to accept that the output is correct despite the gaps in knowledge. In these cases, the more important question to consider is if the output serves the common good. These are concentric ethical questions and they require deliberation and care in their exposition.

### Box 8: Citizens' questions on the accountability of ADS

- Who is accountable (according to the law)?
  - Is there always going to be a human who is responsible for what decision has been taken by (or with the aid of) a computer?
- Is it possible for companies to deflect accountability and what is this level of risk (ie under current legal or regulatory system)?
- If an ADS is developing and learning on its own, how does it impact on explainability and accountability?
- Is there always a human-in-the-loop, or at the very end of the decision-making process?
- At what point does ethics need to be considered, and by whom?
  - Is this a broader problem about business ethics, rather than specific to AI/ADS?
- What sort of complaints or appeals process is in place to hold an organisation to account for an automated decision?
  - How quick is this process?
  - Does access to justice differ between people (ie if you're poor, can you still hold these organisations to account)?

### Box 9: Takeaways on communicating about radical technology

When citizens responded to our questions about AI and ADS they immediately reached for stories from their own lives to help understand and contextualise their opinions. Talking about data or ADS in the abstract didn't carry much meaning, but once the jury could see how ADS have tangible effects on the real world, they became much more emotionally involved and concerned with ensuring they were used for public benefit.

**Accessibility**
Experts who described practical applications and used analogies to explain concepts, were intuitively understood much faster. One example already given above, is that of an algorithm as a cake recipe, another was to compare the need for ADS regulation to the development of building regulation over time to keep society safe. In one of the sessions we asked citizens to experiment with different uses of ADS including: AutoDraw, a predictive drawing tool; the virtual assistant Siri; and a chapter of Harry Potter written by a predictive text keyboard.[22] This was intended to demonstrate the use cases of AI, but interestingly also did something to assuage some citizens' fear of general AI as they saw how difficult it was to perfect just one application of narrow AI.

**Time**
Clarifying the difference between AI and ADS in practice took some time to sink in, and certain concepts were challenging to move beyond, therefore requiring more time for consideration.

- Anxiety over machines displacing humans in their jobs was pervasive throughout. In order not deviate from the jury question, some citizens required constant reassurance about the role of ADS as a support tool, intended to enhance the performance of people at work. There are strategies for ensuring that associated concerns with the technology don't dominate or distract from key questions; this may include trying to contain discussion through an early and dedicated session about the impact of the specific technology on jobs. We found that it was important to acknowledge juror concerns on the matter, but facilitators were able to successfully bring the conversation back on topic by reminding them of their responsibility to answer the jury question, which was displayed on the wall for all the sessions.
- We noticed that the jury were much more confident in their understanding by the second full day, asking the experts insightful and difficult questions after they had had a night to digest everything from the day before. The timing of intense learning sessions, with plenty of opportunities for reflection is another important point when considering design.

Citizens share similar concerns to experts. Several experts commented that they were pleased to be questioned on concepts such as accountability, transparency and explainability as they felt their work and their own concerns over these issues were being validated by the wider public. Organisations should expect interrogation of the technology they are developing, not only by the active policymaker audience, but the public also expect a response to these questions.

22 For more information see: www.autodraw.com and botnik.org/content/harry-potter.html

# 3. Case studies: Recruitment, healthcare, criminal justice

While citizens did recognise the potential of emerging technology while learning about AI and ADS, they also expressed some concern. This is evident from the questions they asked about malicious use, bias and accountability, for example, as well as the list of cons they came up with. However, their concerns became clearer and more practical to engage with when we moved from discussing this technology in a broad sense to reviewing particular applications.

As part of our dialogue, we invited experts to present case studies on the use of ADS in three sectors; recruitment, healthcare and criminal justice. As citizens interrogated these applications, experts could then respond to specific issues that citizens raised. In this chapter, we provide an overview of how citizens processed various uses of ADS, what challenges they identified and how dialogue enabled experts to address citizens' concerns in a meaningful way, or indeed highlight challenges that required experts to go back to the drawing board. We do this by referring specifically to the process we deployed in our citizens' jury by way of example – though other ways in other contexts may draw these subjects out.

## Recruitment

### Exploring the use of ADS for hiring workers

The expert presenting this case study was an employment lawyer whose firm works with organisations that develop AI tools as well as organisations that use them. He specialises in issues of discrimination, including those that pertain to the use of ADS in the workplace.

The lawyer highlighted that AI was being used predominantly in two ways:

1. To more efficiently replicate decisions that humans would make. For example, the legal sector is making use of algorithms which can review contracts at a quicker pace and make recommendations about what lawyers should then focus on, such as particular clauses that indicate risk.
2. To more objectively make, or help make, decisions that would otherwise be prone to human fallibility and bias. For example, across industries, HR practitioners are turning to algorithms that can help them make decisions about who to shortlist for an interview or to hire as an employee.

Focusing on the second use, he explained that reviewing job applications is particularly labour intensive and that it may be unrealistic to expect that an individual, let alone a team of recruiters, would approach each application in the exact same way. Even factors such as the time of day an application is reviewed can affect the recruiter's judgment. AI may therefore be appealing because "it doesn't get tired, sick or distracted." Organisations that often attract a large volume of applicants may wish to use AI to rapidly process applications, trusting that it will be consistent in its approach to selection. However, they may run into problems if the algorithms learned to reinforce unconscious biases and/or

it isn't clear how they reached their conclusions, particularly if there appears to be discrimination (eg on the basis of gender).

While the majority of AI-based recruitment tools scan applications, he also took note of newer products on the market, including HireVue, which provides a video interviewing platform and uses voice and facial recognition technology to take audio and visual cues, such as tonality and expressiveness, into consideration as part of candidates' assessments. These visual cues are interpreted as predictive of performance based on a comparative assessment of the company's existing employees.

Following these examples, the employment lawyer presenting this case study encouraged our citizens to examine ADS from two perspectives, asking:

1. How would you feel if you knew a machine was making, or helping to make, a key decision on your career?
2. What about people who are in the workforce already, whose jobs may be affected by machines taking on these sorts of tasks? Is that something as a society we want, and what kinds of limits might be needed?

Some citizens were positive about the use of ADS in recruitment, referring to the expert's examples about how assessments could change based on timing or mood, while others expressed more uncertainty without greater transparency about how these systems work.

> "Some of the things [AI is used for] that are bigger – jobs, its role in interviews, etc – are more contentious. We're a little less comfortable, maybe because we don't know what they're basing decisions on?"

> "It's tricky [to evaluate this application] if you don't know how the algorithm has been trained."

Several citizens emphasised the importance of employers trying to find the 'right fit' rather than just matching the right qualifications. Some were particularly keen on personality being considered and suggested that only other humans could assess this trait. Companies like HireVue try to address this by screening the first round of applicants using video technology, which provides candidates with an opportunity to demonstrate their disposition at an early stage before then being brought in for an in-person interview. However, this sort of technology was met with suspicion; like other experts, citizens doubted the veracity of the behavioural science supposedly underpinning the technology.[23]

There was also some trepidation about the potential for ADS to further embed biases. There was concern that the criteria being taken into consideration by ADS might be too narrow and therefore exclusionary.

> "If a company is predominantly white and male, will it favour these characteristics?"

> "Who is determining what a 'top performer' is? This is hard to define for many jobs other than sales, for example."

23 AI Now 2018 report notes that: "These claims are not backed by robust scientific evidence, and are being applied in unethical and irresponsible ways that often recall the pseudosciences of phrenology and physiognomy. Linking affect recognition to hiring, access to insurance, education, and policing creates deeply concerning risks, at both an individual and societal level." Whittaker, M. et al. (2018) AI Now Institute, New York University. Available at: ainowinstitute.org/AI_Now_2018_Report.pdf

To mitigate against bias, citizens wanted to know the criteria that was being used to calculate candidates' scores; they also thought this would be helpful for providing feedback to unsuccessful candidates. However, with support from the expert, they also acknowledged that there may not be much transparency in recruitment processes at present and that too much transparency can enable gaming of the system.

Citizens also had questions about accountability in terms of policing and overseeing how these algorithms function as well as in relation to appealing decisions. With regard to the latter, it was recognised that some degree of explainability is necessary in order to mount a challenge.

Notably, citizens appeared to be more open to the use of ADS to make decisions related to existing employees rather than for scouting new hires. Citizens were warm to the idea of using ADS to determine whether they received a pay rise or a promotion, as opposed to either decisions being solely in the hands of line managers. One person voiced her preference for ADS because she believed her employer was discriminating against her on the basis of her age. Our expert pointed out that if these systems were being used to help make a decision about promotion, they could ultimately indicate whom to let go of. He noted that some employers, such as Amazon, are using tracking and monitoring technology to inform these sorts of decisions.[24]

Reflecting on the citizens' deliberation, he noted that he found the dialogue about ADS in the workplace far more emotive than he had anticipated. He concluded that people should still drive recruitment while receiving some limited support from AI.

From the citizens' responses, it's apparent that transparency, explainability and accountability all play a part in how trusting people are of these systems. Explanations in particular seemed important to citizens as a means of reassuring them that these systems were being deployed responsibly and served their interests.

#### Explainability of ADS for hiring workers

In a subsequent session in partnership with the Alan Turing Institute and the Information Commissioner's Office (ICO), we focused on what sort of explanation citizens might want if they were subjected to a hiring decision made with the support of ADS.

At the beginning of the session, our expert from the ICO responded to citizen's concern that ADS may struggle to analyse or appreciate human qualities such as creativity. He assured them that the aim, and the benefit, of using an AI system would be to recognise those qualities by analysing phases, structures and patterns that would correlate with the best employees. He expressed that AI should go beyond tick-box exercises, helping humans grapple with complexity rather than simply checking off how many A-levels a candidate has, for example.

With regard to the explanation, citizens were keen on receiving specific feedback about why they didn't qualify for the next round, so that they could reflect on any constructive criticism and improve their future prospects. Considering that feedback currently tends to be limited at the initial stages of applying, citizens were optimistic that ADS could improve an organisation's

24 For more information see: www.theverge.com/2019/4/25/18516004/amazon-warehouse-fulfillment-centers-productivity-firing-terminations

capacity for responding to unsuccessful candidates earlier, and with timelier and higher quality feedback. In general, if ADS is being used citizens would expect a more efficient turnaround in response time than they would otherwise expect of an individual. Although citizens emphasised the need for a simple, accessible explanation about the rationale for a decision, some also said they'd like the option of more technical information (ie through a link) and that they would potentially have more confidence in the system if the organisation could disclose how long it had been in place.

#### Governing the use of ADS for hiring workers

When discussing accountability for ADS, the idea of externally auditing this technology was raised. Citizens noted that if the technology was developed in-house, they would be more trusting if a regulatory body could be relied on for oversight and took responsibility for setting guidelines that organisations would be expected to follow. These guidelines were open to interpretation but could stipulate requirements (ie for testing and evaluating systems, training staff to use ADS, etc) and clarify procedures (ie the complaints or appeals process). If the technology was procured from another company, it was assumed that the organisation would be even less likely to have knowledge of how to manage an ADS responsibly, further underlining the need for an external body and statutory guidelines.

## Healthcare

#### Exploring the use of ADS for assessing patients

Johan Ordish, a Senior Policy Analyst from the PHG Foundation at the University of Cambridge, joined us to provide an overview of the use of AI in healthcare, ranging from research to care.

He explained that there are three broad applications of AI in healthcare:

1. **Automating tasks.** For example, the NHS 111 system is experimenting with AI for patient triaging by using natural language processing to recognise words that indicate urgency and redirecting callers accordingly.
2. **Analysing large datasets.** For example, AI is being used to review the 2.5m scientific articles that are published each year in order to make recommendations that are specific to an individual's healthcare profession.[25]
3. **Predicting conditions through complex pattern recognition.** Emerging apps like SkinVision are identifying users' likely conditions and making related suggestions for redress.

Johan also shared examples of AI being used to assist medical professionals by identifying existing drugs that could treat rare diseases (Helix), distinguishing cancerous tissue for surgeons to cut through (iKnife) and illuminating where tumours may be (InnerEye).

The focus of his case study was InnerEye, an algorithm which has been developed to automatically highlight anatomy in a computerised tomography

25 See similar example at: www.researchgate.net/publication/334209824_Unsupervised_word_embeddings_capture_latent_knowledge_from_materials_science_literature

(CT) scan with the aim of supporting pathologists and radiologists to recognise and treat tumours. The aim of training algorithms on anatomy would be to quicken the analysis of scans, thereby enabling more regular scanning of patients, accurate tracking of disease progression and the detection of any swelling (an indicator of a serious condition even in the absence of a tumour). It is a form of ADS in the sense that it is predicting where a tumour most likely is so that pathologists and radiologists can then act, making decisions about whether the patient requires treatment and how to proceed.

The use of AI and ADS in healthcare as described by Johan was positively received. Citizens were impressed by the majority of examples, particularly iKnife and InnerEye, although they were more sceptical of automated 111 triaging. It seems that citizens were more enthusiastic because they already had a high level of trust in the NHS (which is generally true of the public, as supported by findings from the Edelman Trust Barometer) and, as part of this, faith in the existing regulatory system. There was also general awareness of the intense financial pressure and resourcing constraints within the NHS, which seemed to fuel greater sympathy for the use of ADS.

> "If the NHS decides it is cheaper and more effective and efficient, it is better. It frees up doctors for other services."

However, some citizens did follow up with important questions about the possible consequences of relying on ADS. For example, citizens asked whether ADS would lead to over-diagnosis since it is risk-adverse and prone to false positives. Our expert noted that this was a possibility, but that it highlights the need for healthcare professionals to work in tandem with ADS and agree the actual, or final, diagnosis. Some citizens wondered whether ADS might be used to decide which patients to invest in (ie in terms of providing expensive therapy or treatment), which could disadvantage older people, for example. In this case, our expert countered that healthcare professionals in the NHS already make similar decisions and questioned why citizens found it less acceptable for a machine to help make that decision as opposed to a doctor deciding on their own. A debate also ensued about how this technology might change doctor/patient interactions; one citizen, who is a psychiatric nurse, argued that people value being able to speak directly with doctors, whereas another citizen mused that this is a culturally normative expectation we've been raised with and that future generations may feel differently.

In contrast with the use of ADS for recruitment, citizens seemed less concerned with transparency of ADS in healthcare, although this may depend on the decision and what kind of impact it would have on the patient. Citizens clearly felt that the use of ADS within healthcare is likely to have significant impact and the potential benefits should not be missed out on. However, they were cautious when it came to discussing data-sharing agreements, understanding the need for high quality data to train systems but wanting to ensure anonymity and that the data they had chosen to share with the NHS would not later be used against them by a third party (in insurance claims for example).

Citizens expressed uncertainty about who should be held responsible if a mistake was made by a system, debating whether it came down to the regulator or if different parties should share responsibility. However, they ultimately expressed more confidence in the use of ADS in healthcare than other sectors because of their trust in the NHS and its existing regulatory practices.

### Explainability of ADS for assessing patients

Reflecting on how positive the reception had been to ADS in healthcare thus far, we decided to test a potential use of ADS that could prove more controversial during our session on explainability with the Turing Institute and ICO. We asked citizens to consider the hypothetical use of AI to predict whether a patient would benefit from a particular form of treatment and, more specifically, what sort of explanation they would need about a resulting decision. The use of AI for this purpose was more contentious than the use of AI in pathology, for example, because there was a perception that some patients would lose out [on treatment] if ADS was deployed. When the stakes were raised, citizens were clear about the need for using sensitive language and conveying empathy when communicating about a decision of this significance. Citizens thus preferred to be informed of decisions of this nature face-to-face rather than through formal correspondence.

In terms of the explanation itself, citizens noted that they wanted to know more about the policies or guidelines the decision followed (for example, 'if the probability of a patient benefiting from this treatment is below X threshold, the practitioner should not proceed with treatment'; 'The NHS has set this policy because of X'). They wanted information about what, if any, alternative options they had, and to be informed of how they could challenge or complain about the decision. Some citizens expressed that a counterfactual explanation would not be appropriate in this case, as the patient's ability to modify the factors resulting in that decision would be limited.

> "In this case [recruitment], it's about changing things. But in a medical case its acceptance of the decision which is important."

### Governing the use of ADS for assessing patients

Given the potential life or death scenarios that ADS may be used for within healthcare, generally citizens valued accuracy over explainability although that was heavily dependent on the use of robust and nonbiased datasets. They called into need auditing procedures for the datasets used to train MedTech as well as regular auditing of the outcomes, and transparency over whether predictions were more likely to be inaccurate for certain groups due to smaller sample sizes.

More so than in any other sector, the jury expressed a strong desire for human oversight over significant decisions and were adamant that the outcome was delivered by someone capable of empathy. Citizens wanted the NHS to use ADS to support, and not replace staff, highlighting the importance of training so that staff feel comfortable using these systems and that they are able to question the outcome of a decision should it go against their professional judgement. Indeed, engaging healthcare professionals and building workforce capacity and capability have been cited amongst the top three factors needed to enable AI within health and care by experts working across the AI ecosystem in England.[26]

26 The ASHN Network (2019). Accelerating Artificial Intelligence in health and care: results from a state of the nation survey. Available at: www.kssahsn.net/what-we-do/our-news/news/Documents/AI-Strategy.pdf

"It's both. Definitely in the delivery but even with the diagnosis, there are lots of factors and people will want that explained to them. I suppose I'm OK with AI being part of the process, but a human has to be part of the process too."

"If you're just doing it on ADS, 'we've decided your genetics mean this won't work for you' – then no, I wouldn't accept it. What if the ADS is malfunctioning?"

## Policing and criminal justice

**Exploring the use of ADS in policing and the criminal justice system**

Citizens deliberated various uses of ADS in the policing and criminal justice system, although the primary focus was on Durham Constabulary's Harm Assessment Risk Tool (HART). To present on HART, we invited Sheena Urwin, Head of Criminal Justice at Durham Constabulary, and Marion Oswald, an academic at the University of Winchester (at the time of delivering case studies), to join us. Sheena oversees the implementation of HART, while Marion has played a role in creating a decision-making framework for the deployment of HART and other algorithmic assessment tools in policing.

### Case study

HART was developed to aid decision-making by custody officers when assessing the risk of arrestees reoffending.[27] It uses a random forest model to divide arrestees into three groups based on their risk of reoffending. Arrestees who are forecast as moderate risk of reoffending are eligible for the Constabulary's Checkpoint programme. The programme is an 'out of court disposal', which refers to a way of dealing with arrestees who commit low-level offences that don't require prosecution. HART is not intended to be used in a determinative way; rather, it supports custody officers in "ensuring the most expensive and punitive options are targeted on the 'right' offenders."[28]

Sheena explained that the Checkpoint programme prefaced HART, and was designed to address the 'revolving door' in the criminal justice system. The revolving door describes low-level offenders cycling in and out of the system. These offenders commit an offence, go through the court system, come out with a conditional discharge for a length of time, and then restart this cycle because they have reoffended before that time is up. Once an offender has been through the revolving door a few times, they may be sentenced to prison. Checkpoint enables the Constabulary to enter into a formal, four-month contract with offenders as opposed to prosecuting them; if the offenders work with 'Navigators' to comply with the contract, they can avoid a criminal conviction.

HART helps better target people who are caught up in the revolving door. It assists custody officers in selecting those most likely to benefit from the programme – arrestees who are at moderate risk of reoffending. In partnership with Cambridge University, a random forest model was built to assess risk using 34 predictive variables. It is largely based on prior offending data, but also

27 For more information see: www.tandfonline.com/doi/pdf/10.1080/13600834.2018.1458455
28 Ibid.

includes gender, post codes and intelligence count. A custody officer takes the risk score into consideration when making the ultimate decision about whether someone is referred to Checkpoint.

HART is being subjected to independent validation, as well as considered with reference to ethical and legal frameworks. Marion expanded on some of the questions they had taken into account when reviewing HART, posing them to the jury for reflection:

- What is the data that goes into the algorithmic model, and is it relevant to the decisions being made?
  - For example, gender – how relevant is that to the decision that's being made?
- What about the outputs?
  - For example, what effects do the outputs have on particular groups; is there any sort of discriminatory effect based on the way the data is manipulated?
  - What do you then do with that risk assessment? From a statistical point of view, these outputs are saying that this particular individual is part of a group, not necessarily saying anything about this individual's future. How is that [the individual's future] assessed using this model?
- How do you challenge the outputs?
  - If you want to argue back, how do you do that if you can't understand what is really going on, especially in a criminal justice environment?
- What kind of effects do these tools have on the human decision-making process?
  - Traditionally, it's the custody sergeant who makes these decisions. Now they're being told to use this tool as well, but what effect does that have on the way the decisions are making made? Are they still fair?

During deliberation, citizens probed the experts on the information used by the ADS to calculate risk. There was some concern about whether the use of postcode data may result in biased outcomes, and citizens questioned what safeguards were in place to prevent bias. Some citizens expressed scepticism about HART's rate of accuracy (62 percent) and debated what the acceptable threshold for use should be. Citizens wanted to know whether custody sergeants could overrule decisions, and what the use of this tool would mean for jobs (ie as in whether it would replace custody officers). The experts repeatedly assured the jurors that HART was a decision-support tool, meaning that the outcome is not acted on autonomously but is considered by a custody sergeant alongside other relevant information. This reassurance also addressed concerns voiced by some citizens that the ADS could not be reasoned with through dialogue as they imagined might be possible with a custody sergeant, judge or other relevant decision-maker.

In subsequent sessions, we explored the use of facial recognition technology in policing. The main concern jurors had was that the police force was institutionally racist and that would therefore affect the way ADS is used. While there was some recognition of the potential to minimise biased stop and search practices and the number of people being inaccurately targeted, they thought it was more likely to exacerbate rather than solve racial prejudice against minority

groups. Furthermore, they questioned why innocent members of the public should be subject to surveillance, considering this to be a limitation on social freedoms:

> "It's not the system that's biased, it's the people operating the system. The police chose to use the system at the Notting Hill Festival – that was a choice."

> "Everyone would get screened as part of the facial recognition – so what if you haven't done anything wrong, why should you be screened?""

#### Additional insight from our workshops on facial recognition tech

In addition to the citizens' jury, the RSA ran two workshops with young BAME people, mainly men, who are disproportionately likely to be impacted by the use of facial recognition technology within policing.[29] The citizens' jury format is not the best mechanism by which to hear from minority groups as the roughly representative sample of the wider population means their unique experiences are likely to be drowned out within group discussions. Instead we collaborated with Dr Adam Elliott-Cooper, an experienced facilitator who has a background working with community organisations that monitor and scrutinise the misuse of police power, to design 2 to 2.5-hour workshops so that participants would feel comfortable sharing amongst their peers. The workshops incorporated some basic information on ADS and case studies of their use in policing, as well as a session about citizens' rights in relation to stop and search, with aims to empower participants to be better prepared to respond to uses of this technology. The first of these workshops was held in November 2018 with students at Goldsmiths University, and the second in February 2019 at Hackney Community College in partnership with Voyage Youth who recruited young people from their network. We spoke to over 20 participants across both workshops about their experience of policing through stop and search and how they thought the application of AI technology would affect their experience of the criminal justice system.

#### Experiences and attitudes towards policing and stop and search

Both groups acknowledged police presence is intended to keep society safe but reflecting on their own personal experiences they considered the majority of interactions they had had with the police to have been negative. They felt that they were racially stereotyped and stop and search disproportionately targeted them: "We're even stopped more in places we're not the majority. Somewhere there is no knife or gang crime." Adam explained that in 80 percent of stop and search instances the police don't find anything, the most common reason for a "positive outcome", in other words arrest, is marijuana possession and the second most common reason is that "you fit a description".

Several participants cited the lack of diversity within the police force as a barrier to intelligence-led rather than assumption-led policing. One person mentioned that even the few black police officers they had interacted with had been aggressive or used unnecessary force, questioning the legitimacy of the force's institutional culture. Given the humiliating and often frightening

29 For more information see: commonslibrary.parliament.uk/home-affairs/security/police-use-of-live-facial-recognition-technology-challenges-and-concerns/

situation of being stopped and searched, participants said good policing would embody empathy and respect acknowledging the individual beyond stereotypes.

When participants reflected on whether facial recognition technology would improve bias within the police force, a key concern was the level of accuracy of the technology. One participant wondered whether a legal case could be brought against them using the image of somebody else. Many facial recognition products are known to be less effective at accurately evaluating people with darker skin tones, and females. A study by researchers at MIT and Stanford University reviewing products from Microsoft, IBM and Face++, found all three to be 99-100 percent accurate at evaluating light male faces and the worst performing when it came to identifying darker females, with the largest gap between the two being 34.4 percent.[30] Similarly, as to the need for a diversified police force, people commented that there needed to be greater diversity within the sector to prevent unconscious bias amongst those designing and developing these types of technology from putting minority groups at risk at being misidentified. Considering the significant impact stop and search can have on those being targeted, several individuals suggested the technology shouldn't be used unless perfect.

> "Is it acceptable to use if it has an accuracy of over 70 percent? It's easy to say 70 percent if you're never at risk of being stopped and searched."

If accuracy could be completely assured, a couple of participants suggested that using this type of technology may result in fairer policing as everyone is subject to the same screening process rather than certain groups being targeted. They suggested it could be more objective than a police officer as: "AI doesn't have feelings, and discrimination is led by feelings."

Most importantly however, people drew attention to the structures within which facial recognition technology is used. They suggested that regardless of the accuracy of facial recognition technology, if the organisation using it or the society it is being used in is biased then it can be used in a discriminatory way: "It matters who is on the database to begin with."

> "Whatever system is used, if the institution using it is biased it will be too. I'm looking more at the people behind it than the system itself."

### Explainability of ADS in policing and the criminal justice system: keeping a human in the loop

The jury were slightly more open to the use of Durham Constabulary's HART as compared to facial recognition technology, perhaps because it was seen to be preventative rather than punitive use of ADS, but it was still very controversial particularly in its use of postcode data. Within the explanation, jurors thought there needed to be a good description of what the Checkpoint programme itself was, the eligibility criteria of the programme and what data points were used in HART. They also wanted to know what the significant data points were in determining someone's eligibility, and the main reasons for placing them into a certain risk category. This was another situation in which many jurors said they

30 Buolamwini, J. (2018). Gender Shades. Available at: gendershades.org. See also: medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced

would be unhappy with a counterfactual explanation as historical criminal records or postcode data are factors that individuals have little or no ability to change about themselves.

> "He needs to be told it's an automated system, he needs to know what criteria are being used, and he needs to know what data specific to him is being input."

In order to trust the decision, jurors wanted some assurance of accountability, through audits to check decisions for accuracy and a structured appeals process that had a **human in the loop**. Depending on whether the decision was positive or negative and the severity of the crimes the person in question had committed, jurors suggested this would be another circumstance in which a decision should be delivered by a human and discussed in person rather than via written correspondence.

> "I think this is the right form in the right cases; this could be a drink driving offence, no one is going to skip town because of a driving offence; but if you're dealing with someone who is more of a risk, then a decision should be made that an in-person visit is made to that person."

> "I think we should be told that the accuracy is checked. If the desk sergeant was to do these checks himself, would he come to the same decision as the machine?"

#### Governing specific tech (such as recidivism risk assessment tools and facial recognition technology)

There were mixed views on the use of ADS in the criminal justice system, particularly with the use of certain data points, but there was some support for ADS being used to inform a decision rather than make the final call. Jury members and the participants in the workshops on facial recognition emphasised that reviewing datasets to make sure they are unbiased is only one step towards ensuring these tools are fair. The impacts on different groups need to be evaluated and the technology regulated to prevent abusive or lethal discrimination.

> "I'd be comfortable in the criminal system to have AI as a backup, and an aid, for the judge. I can get extra information from the machine to look at what I already know."

> "But you have to first understand why there's crime in that area."

# 4. A toolkit for institutions and citizens

Throughout this report we have seen what is possible when citizens are actively engaged in a deep and considered fashion in the complex business of ADS in a variety of contexts. The suggested questions and prompts, in sum, we hope provide a working model that can be adapted to a greater or lesser extent by organisations looking to engage in conversations around the ethics of ADS in their operational systems.

The citizens' jury process is one way of getting to this point. In the next chapter we will consider the pros and cons of citizens' juries as a whole in the spectrum of engagement methods that bring deliberation and democracy into the arena of radical technologies.

For now, this chapter rounds off our account of the citizens' jury process by providing a summary of the conditions that jurors would like to see built-in throughout the process of ADS.

## Conditions and considerations at the design stage

An important condition in this stage, as highlighted by the jurors, is the need to ensure that equality and diversity is built into the use of ADS, by ensuring against bias in the training of the data. Jury members were concerned that ADS would be less accurate for minority groups if the datasets used to train the algorithms were not sufficiently diverse or contain enough examples from those minority groups.

This is demonstrated by the differing levels of accuracy at evaluating lighter and darker skinned people, and people of different gender by facial recognition technology. Jurors also placed importance on protection against hacking and ensuring data security.

Legal requirements for appropriate data collection and the right to privacy should be enacted through the following mechanisms:

- Data collection not to exceed requirements for use (ie no 'general interest' data)
- Only fully anonymised data to be shared, nothing that identifies individuals. This condition was particularly relevant to the healthcare case studies where jurors saw huge potential public benefit from data-sharing but wanted to ensure individual privacy.

Jurors felt it was crucial to ensure 'human-in-the-loop': ADS prediction only, not decision making. In many of the cases, jurors were adamant that ADS be used in conjunction with a trained professional. They wanted doctors or custody officers to benefit from the support of predictions from ADS, but for those predictions to come with a specified degree of accuracy and for the human-in-the-loop to have agency to consider other factors in making a more holistic decision. Some citizens mentioned wanting a human touch in terms of empathy, but also in ensuring oversight over decisions.

## Conditions and considerations at the creation stage

A key condition for the creation stage is to ensure that organisational and technical responsibilities are set down in policy and legislation, for example,

expectations for testing, monitoring and auditing. Further, there is a need for robust monitoring to prevent malfunction.

Establishing standards for auditing use of ADS is critical. Jurors were concerned that without regular audits of the outcomes of ADS, they could end up being used even if the predictions they were making were inaccurate or discriminatory. To prevent this from occurring, citizens wanted organisations using ADS to commit to regular auditing to make sure they were still providing the intended outcomes. The guidance on standards for those audits should be provided by sector-specific bodies, and who is to be responsible for conducting audits should be specified prior to ADS being rolled out widely

There is a need for independent external audits of ADS use. Jurors suggested that a third party independent could be responsible for conducting audits to ensure accountability for organisations that were worried about sharing intellectual property during an auditing process. They thought this may add an extra layer of credibility and perhaps boost public confidence in ADS that had been independently audited.

Additional conditions include ensuring that there is always a robust back-up system in place to address malfunction or poor use, and ensuring the protection of individuals who may be negatively affected by ADS due to membership of a specific group (eg race, religion, postcode). At this stage, the condition refers to the responsibility of organisations developing ADS.

### Conditions and considerations at the use (application) stage

At the use stage, jurors were clear that there are a number of conditions to be considered:

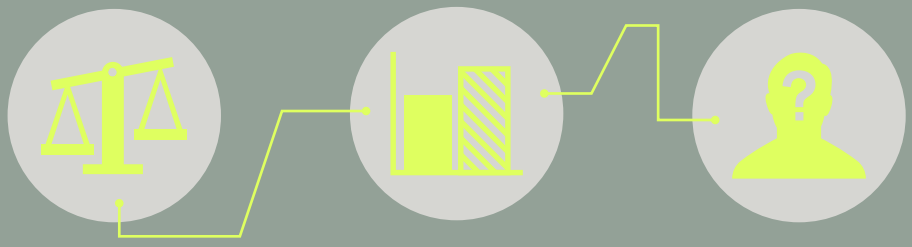
- Policy and assurance (eg privacy, security) must be proportionate to the severity of potential negative impacts.
- Staff training and monitoring on the use of ADS, including in ethical issues.
- Clear and accessible information to customers, clients and the public about when and how ADS is being used.
- Legal requirements for equality and diversity in relation to the use of ADS (ie to protect against discrimination).
- Legal requirement for the explanation of specific ADS decisions.
- Right of appeal against particular decisions made using ADS.
- Establish policies to protect against unemployment resulting from the use of ADS.

In addition, there is a need to establish standards for training in the use of ADS. Jurors were concerned that society may fail to enjoy the benefits of ADS if they are misused or neglect to be used in practice. To ensure those using ADS are comfortable and feel able to challenge a decision if it goes against their professional opinion, jurors wanted them to have adequate training and support during implementation. What these standards look like would need to differ by sector and application.

Lastly, there is the condition for organisations to provide a clear explanation of why a particular decision has been reached. Jurors had different expectations for an explanation dependant on the sector and application of the ADS. In some, for example, they thought counterfactual explanations would be useful, but in others be potentially offensive.

## Figure 2: A toolkit for institutions and citizens

**Conditions and considerations at the design stage**

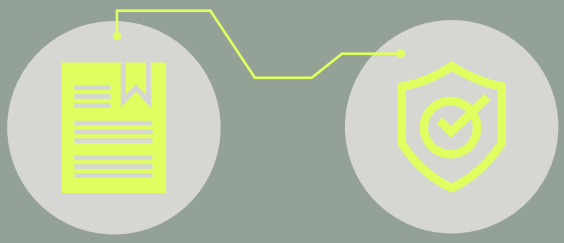


Ensure equality and diversity within ADS, the data set used for training must be unbiassed.

Ensure the data gathered for training should not exceed requirements for use.

Data should be anonymised to defend against misuse.

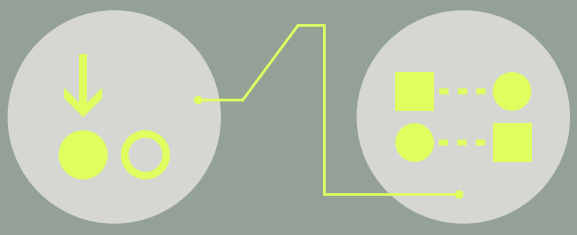
**Conditions and considerations at the creation stage**



There should be robust policy and legislation in place to ensure organisational and technical responsibilities with respect to testing, monitoring and audit.

In order to be credible, audits should be carried out independently and externally.

**Conditions and considerations at the application stage**



In the use of ADS, steps should be taken to ensure accountability. For example, by establishing a legal requirement for explanation and a right to appeal.

Policy and assurance should be empathetic and proportionate with the severity of potential negative impacts, for example where there is a potential for algorithmic prejudice.

**Questions our citizens recommend their peers ask about ADS**

| |
|---|
| What impact will ADS have on broader social structures and interactions? |
| It is safe in the way my details are being shared? |
| Will I know that ADS is being used? |
| How or is it regulated? |
| How can I challenge it? |
| How does it benefit me? |
| Is this system fair? |
| Has sufficient data been used to train the system so that it's accurate? |
| What's the government's role in all this? Is there any drive to promote such systems? Is there any investment in place? Is there something happening in the background that we don't know about? |
| How can we protect systems against hackers? |
| How do people without access to tech get represented in the data? |
| Will expertise of AI become more varied in the future? |
| Who is determining the ethical standards? |
| Who's profiting? |
| How do you train algorithms? |
| Are there any legal requirements around accuracy? |

What is notable about these recommendations on ADS, as prescribed by the jurors, is that they all have elements of requirements for transparency, accountability and explainability woven through every stage of the process, from design through to application.

The citizens in our jury emphasised the importance of ADS being fair on diverse groups at all stages and placed a high value on ADS aiding human decisions, not making them autonomously. The theme of scrutiny and evaluation is pervasive throughout the feedback and conclusions of jurors, from the establishment of audit standards to the continued training and monitoring around the use of ADS. The jurors recognised the need for flexibility and agility in these conditions as the capability of ADS proliferates.

# 5. Conclusion: Deliberating the future

It is important to note that no two processes for engagement and deliberation are exactly alike. We round off this report therefore with some broader reflections on the role of citizens' juries and other deliberative methods in the development of ADS and radical technologies.

### Lessons in deliberation: On citizens' juries

In the first chapter of this report we outlined a standardised approach to citizens' juries - and small differences between this and the actual approach undertaken by the RSA. Building on that exercise we note that the following in the RSA's approach may have influenced deliberations in a way that future deployment of such methods may seek to rectify.

### Influence of jury timeline on deliberations

The timeline and structure of the jury may have influenced the deliberations. The intensive first session (Friday to Sunday) was designed to give jurors time to feel comfortable with the process and to build trust in one another and with the delivery team. It was also designed to provide enough concentrated time for jurors to hear from experts and get to grips with the complexity of the technology being discussed. However, subsequent sessions took place weeks and months later and future commissioners may wish to consider whether our timetable was optimal in light of the complex subject matter under consideration.

### Selection of the jury

To select the jury the RSA worked with a market research organisation. This enabled the curation of a group of jurors who broadly reflected the population of England and Wales in the demographic make-up. Different market research companies deploy different recruitment methodologies; this should be a consideration for future deliberative democrats.

### Summary of juror deliberations

Between sessions the RSA summarised the deliberations of the jurors, in order to capture key insights and themes. At the start of the following session these summaries were fed back to the jurors to ensure that they agreed with the summaries, and to provide an opportunity for alterations to be made. These summaries form the basis of this report.

Whilst these summaries should form an accurate reflection of the deliberations and recommendations and have been agreed by the jurors as being such, it is important to note the role of the facilitation team in synthesising the discussions of the jurors. Alternatives to the approach we took – such as empowering citizens to keep their own records - could be part of an alternative jury set-up.

### Lessons for deliberative democrats: Beyond citizens' juries

It should also be noted that there are many deliberative processes beyond citizens' juries. While a complete analysis of each kind of method is beyond the scope of this report, recognising the breadth of possible approaches and their appropriateness for different sorts of inquiry around ADS is an important part of our reflections on this report and our consideration of future work.

The starting point for choosing between methods is 'purpose'. Is the entry point of a deliberative inquiry to, say, improve governance, or build trustworthiness, or gain insight? What is the commissioning organisation trying to achieve?

Different purposes yield different methodological contentions. Inquiries into the ethics of ADS and AI from a governance perspective may focus on broader or narrower questions than an inquiry that seeks to garner more generalised insight.

The following table captures a number of these additional methods and contains some thoughts on their future applicability in the context of ethical AI.

## Table 3: Some methods and future applications of deliberative methods in the context of ethical AI

| Purpose | Method | Description | Advantages | Disadvantages | Examples of future lines of inquiry |
|---|---|---|---|---|---|
| Insights > trust | Future search | Brings together a large group of stakeholders who all have power or knowledge on the topic at hand. It involves a series of meetings among participants where the focus of the meetings shifts from their past experiences with the topic, to the present, on to the ideal future. Participants then agree on common ground for the future. | High degree of buy-in among participants<br>Rich discussion due to experiences of participants<br>If successful common ground is reached this can be motivating for participants. | Requires follow-up to ensure it is carried forward<br>Needs relatively a lot of time and resource<br>Those outside the participants could find it hard to match the energy. | 'How can we create a better social care system using AI'?<br>'Where might AI and the criminal justice system intersect in future?' |
| | Appreciative enquiry | Uses a core group of participants to create a shared vision of the future, based in what has previously worked and what could work in the future. The aim is to create design interventions which will work towards the vision agreed. These design ideas are then shared with wider stakeholders to ensure their feasibility. | Helps to create community involvement and engagement<br>Creates strong vision for the future<br>Encourages participants to work in partnership with each other and wider stakeholders. | It is a fairly loose methodology<br>There is less focus on the 'problem' as some may expect<br>There are relatively less people involved than other methods. | 'Where could we introduce AI in the classroom to most improve the education of young people?' |
| Trust > governance | Citizens' jury | Citizens' juries use a series of workshops spread over 3-5 days, whereby the jurors are posed a specific question and consulted by those with a high degree of knowledge on the subject at hand. They then use the information gained to reach a conclusion on the question(s) posed. | -Able to engage citizens in a relatively technical or complex problem<br>Extended deliberation on the subject<br>Representative of the population. | Relatively small number of participants<br>The issue discussed is often framed in a top-down manner and may not be the most pertinent questions to the participants. | 'Should the use of ADS be permitted in the criminal justice system?'<br>'Should individuals be given the right to withdraw themselves from the use of AI or ADS in public services?'<br>'To what extent can and should the use and application of AI and ADS be explained to citizens when used in an institutional context?' |
| | Citizens' assembly | Similar to citizens' jury, the assembly is used as a way of allowing citizens to learn about and deliberate on policy issues. However, citizens' assembly are often larger and looser in their prescribed methodology. The assembly follows three stages: learning, deliberation and decision making. | Often high profile<br>It can draw out diverse opinions on a subject<br>Decision makers are brought face-to-face with citizens<br>Participants can give a more knowledgeable opinion on a subject following the learning phase. | Can be a highly complex exercise, requiring the management of many participants<br>Can be a costly and long process to organise and run<br>Danger that it may be seen as a token exercise without real outcomes. | 'Should we introduce regulatory legislation on the use of AI or ADS in public services? If so, what should we introduce?' |

## Deliberation as a principle

The RSA's deliberative dialogue proved hugely illuminating in drawing out new insights and recommendations and developing an appreciation that the jurors collectively came up with nuanced and practicable questions and approached. We contend that deliberation in whatever form, is crucial to unpicking these key questions in business, technology and society.

As such, it is critical that the voices of citizens are part of an ongoing iterative dialogue[31]. As contexts change and technologies evolve, we need to ensure that citizen voices are embedded in the decisions on what technologies are used on us as citizens. When systemic bias and organisational groupthink can take root, it may be that it requires the voice of citizens to hold up the mirror and show an organisation what they previously were unable to identify or recognise.

This report is one part of a nascent movement for change. We want it to be the makeweight for further experimentation by ourselves and others. As we've seen in the launch of the Information Commissioner's Office and Alan Turing Institute's interim report[32], or in the ongoing work of the Ada Lovelace Institute[33], there are lots of opportunities where citizens juries and other deliberative approaches can effectively generate rich, qualitative data about the effect AI and ADS have on society and new ideas for what we might do together to ensure the benefits of technological advances are shared and risks properly mitigated. We are, when it comes to deliberating on AI and ADS, near the beginning of an important journey.

Through this process we've also sought to clarify where and how to use deliberative dialogue to maximum benefit. For sure, these methods cannot be afterthoughts; they require in-depth, long-term planning, integration and facilitation if they are to achieve their considerable potential. It is with this in mind that we reflect on the words of The RSA CEO, Matthew Taylor, from when The RSA first began applying the citizen-led approach in this arena 18 months ago:

> "On the one hand, unless the public feels informed and respected in shaping our technological future, the sense will grow that ordinary people have no agency – a sense that is a major driver in the appeal of populism. At worst it could lead to a concerted backlash against those perceived to be exploiting technological change for their own narrow benefit. On the other hand, if those who will shape our technological future – from politicians and officials to corporate leaders and technologists themselves – trust, understand and act on informed public opinion, AI could prove to be a powerful tool to open up new opportunities for human fulfilment."

31 For instance, see our previous report on public engagement with AI: Balaram, B, Greenham, T. and Leonard, J. (2018) Artificial intelligence: real public engagement. [online] London, UK: RSA. Available at: www.thersa.org/discover/publications-and-articles/reports/artificial-intelligence-real-public-engagement. Also see work by Simon Burall: Burall, S. (2018) Rethink public engagement for gene editing. nature, [online] 21 March. Available at: www.nature.com/articles/d41586-018-03269-3. See also: Jasanoff, S. and Hurlbut, J. B. (2018) A global observatory for gene editing. nature, [online] 21 March. Available at: www.nature.com/articles/d41586-018-03270-w. See also: Chilvers, J., Pallett, H. and Hargreaves, T. (2017) Public Engagement with Energy: broadening evidence, policy and practice. [online] London, UK: UK Energy Research Centre. Available at: 3sresearch.org/2017/10/31/broadening-public-engagement-with-energy/

32 Information Commissioner's Office (2019) Project ExplAIn Interim Report. [pdf] London: Information Commissioner's Office. Available at: ico.org.uk/about-the-ico/research-and-reports/project-explain-interim-report/

33 Patel, R. (2019) Public deliberation could help address AI's legitimacy problem in 2019. Ada Lovelace Institute, [online] 8 February. Available at: www.adalovelaceinstitute.org/public-deliberation-could-help-address-ais-legitimacy-problem-in-2019/

# Annex A – Advisory board

**Simon Burall**
Senior Associate of Involve

**Rumman Chowdhury**
Responsible AI Lead at Accenture

**David Edmonds**
Senior Research Associate at the Oxford Uehiro Centre for Practical Ethics

**Paul Mason**
Director of Responsive Programmes for Innovate UK

**Catherine Miller**
Director of Policy at Doteveryone

**Maja Pantic**
Research Director at Samsung Artificial Intelligence Centre

**Beth Singler**
Junior Research Fellow in Artificial Intelligence at Homerton College, University of Cambridge

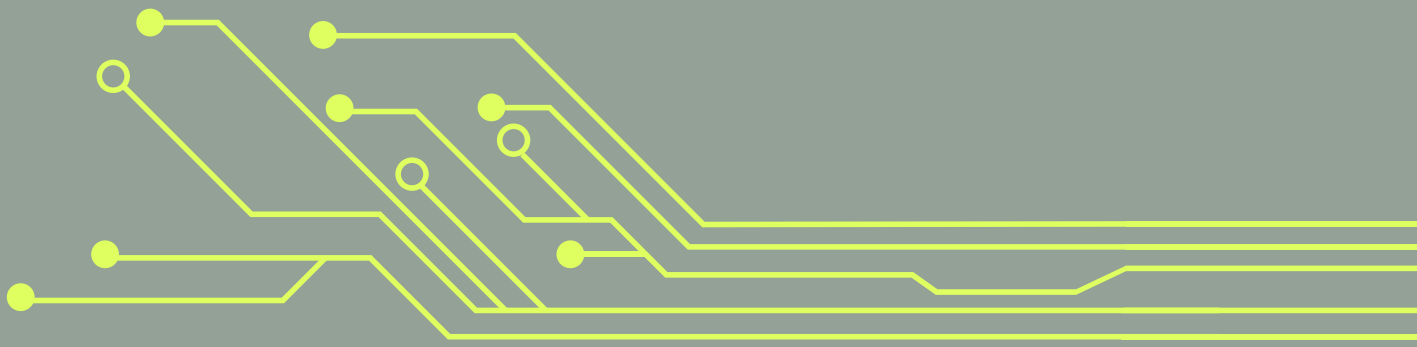
**Wendy Tan-White**
Adviser to BGF Ventures

**Ian Walden**
Professor of Information and Communications Law and Director of the Centre for Commercial Law Studies, Queen Mary, University of London

# Annex B – Partnership with DeepMind

The RSA worked with DeepMind throughout this project, who played a role as commissioner of the jury though played no role in the process of devising or delivering the jury.

DeepMind have published a blog on their involvement with the project at: deepmind.com/blog



RSA
21st century enlightenment

8 John Adam Street
London WC2N 6EZ
+44 (0)20 7930 5115



**www.thersa.org**

ISBN 978-1-911532-35-4