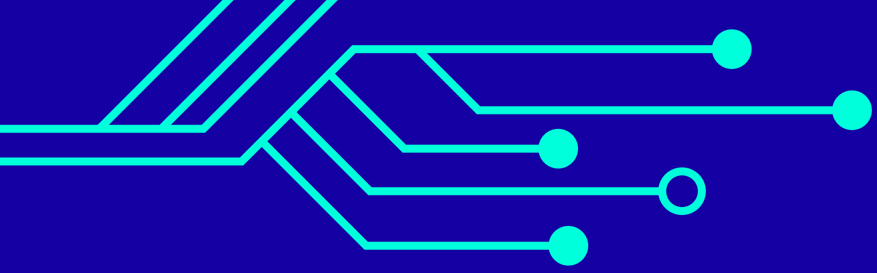
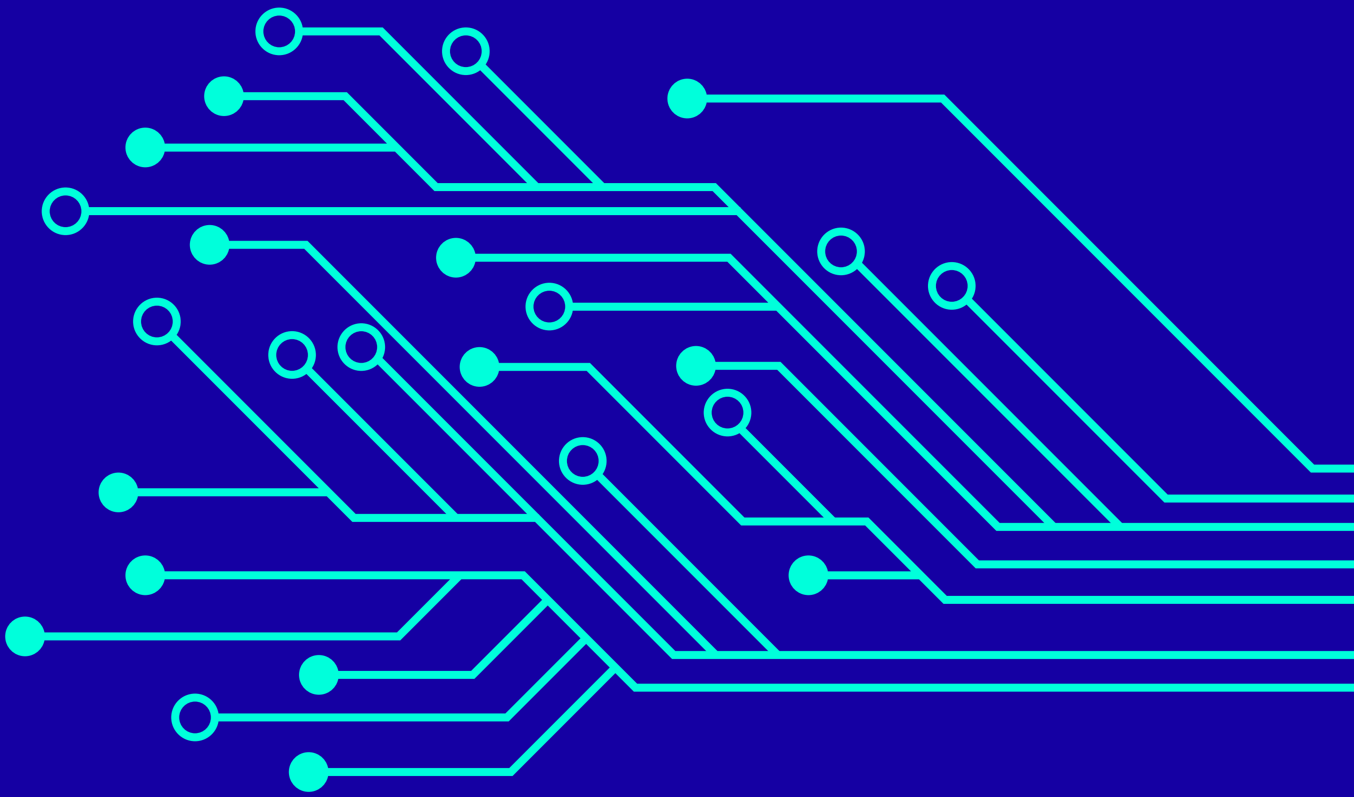


Artificial Intelligence: Real Public Engagement



RSA

21st century enlightenment

Engaging citizens in the ethical use of AI for automated decision-making

About the RSA

The RSA (Royal Society for the encouragement of Arts, Manufactures and Commerce) believes that everyone should have the freedom and power to turn their ideas into reality; we call this the Power to Create. Through our ideas, research and 29,000-strong Fellowship, we seek to realise a society where creative power is distributed, where concentrations of power are confronted, and where creative values are nurtured.

About the Authors

This report was produced by the following members of the programme team for the RSA's Forum for Ethical AI: Brhmie Balaram, Tony Greenham and Jasmine Leonard.

About the Forum for Ethical AI

The RSA and DeepMind are partnering on a new project to encourage and facilitate meaningful public engagement on the real-world impacts of 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 what sort of political messaging is most likely to appeal to a particular group. 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 is designing a citizens' jury to explore the use AI to make, or help make, decisions. Drawing on the model of the RSA's Citizens' Economic Council, we will convene 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.

Acknowledgments

This report would not have been possible without the generous support of DeepMind's Ethics & Society programme, of which the RSA is a partner. In particular, the authors are grateful to Brittany Smith of DeepMind.

We would also like to extend our appreciation to the many RSA colleagues who took time to provide invaluable input and feedback: Kayshani Gibbon, Ben Dellot, Anthony Painter and Matthew Taylor. Thanks goes to Shirin Maani for developing the accompanying interactive tool; Fabian Wallace-Stephens for his support with data analysis, and the team responsible for producing and communicating the report: Thomas Proudfoot, Ash Singleton and Chris Ward.

And finally, thank you to Lauren Chalmers of Studio LP for her role in designing this report.

Foreword

This report, the first emerging from the RSA's work on the ethics of artificial intelligence, speaks to two important themes for us. First, it reflects our approach to technology, which is one that appreciates the great benefits that technological progress can bring, but which resists technological determinism and seeks to shape change to benefit humanity – and the planet – as a whole. Second, the report – indeed the whole project – asks how citizens themselves can be enabled and empowered to influence their shared future.

Public engagement has been a growing focus of the RSA's work. Our Citizens Economic Council, which concluded earlier this year, demonstrated the capacity of a representative group of citizens to engage with economic issues and develop sophisticated ideas. Among its impacts, the project was successful in persuading the Bank of England to set up citizen panels to inform its regional advisors. But the project also made a significant personal impact on the many experts who engaged with it and on the members of the Council. Two comments from council member are typical of the group's experience:

> *"I thought that this is exactly what the country (and the world) needs: hoping that this could be the beginning of people appreciating that they are instrumental in the country's economic decision making processes"*
> - Council member, Satu Jaatinen
>
> *"My participation with the Council is one I could only wish for any other citizens to be part of"* - Council member, Enolia Agbeti

As I will argue in my annual lecture in 2018, it is time to take deliberative forms of engagement, such as the Council and the citizens' jury being organised for this project on AI, from the margins of politics and policy making and embed them in our democratic processes.

This is not only about empowerment, important though that is in these times of mistrust and polarisation. It is also because citizens, given the right support and balanced information, almost invariably demonstrate collective wisdom, getting to the heart of issues and developing well argued conclusions. Given public concern about the impact on people's lives, of cutting edge technologies like AI (concerns highlighted in the survey evidence revealed in this report), it is urgent and vital to hear the voice of informed citizens in shaping norms, practise and policies.

Currently it can feel that the growing ubiquity and sophistication of AI is closely matched by growing public concern about its implications. 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. It is in pursuit of the latter of these outcomes that we present this first report of the RSA project on AI and ethics.

Matthew Taylor,
Chief Executive, RSA

Summary

Technological breakthroughs can be polarising because there are often both benefits and risks. New technology promises us a better way of life, and sometimes it does deliver for the masses. The world has been transformed by medical discoveries like penicillin; revolutionary modes of transport, like trains and planes, and in more recent years, inventions like the internet and the smart phone. But sometimes there are complications or consequences; there have long been concerns about the loss of jobs to automation, but people are increasingly anxious about other risks, such as threats to privacy, security, and psychological well-being, as well as increasing susceptibility to political manipulation and fraud.

Developments in artificial intelligence (AI) are likely to intensify these risks if not handled carefully, and also introduce new ones, such as the possibility of reinforcing systemic biases and exacerbating inequality. Although AI has enormous promise in fields as diverse as education, health and transport, given the risks, a growing chorus of voices is calling for greater consideration of ethics as AI is further developed and adopted more widely.[1] We're at a point now where the backlash to AI may be beginning, and in some cases, rightfully so.

As with any technology, AI's potential to help or harm us depends on how it's applied and overseen. One application that demonstrates this double-edged potential is the use of AI in automated decision systems.

Automated decision systems refer to the computer systems that either inform or make a decision on a course of action to pursue about an individual or business.[2] Automated decision systems do not always use AI, but increasingly draw on the technology as machine learning algorithms can substantially improve the accuracy of predictions. These systems have been used in the private sector for years (for example, to inform decisions about granting loans and managing recruitment and retention of staff), and now many public bodies in the UK are exploring and experimenting with their use to 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.[3]

This technology could have significant social and economic implications, but there has been no 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 decisions about the deployment and regulation of these systems.

From our online survey of the UK population, carried out in partnership with YouGov, we know that most people aren't aware that automated decision systems are being used in these various ways, let alone involved in the process of rolling out or scrutinising these systems. Only 32 percent of people are aware that AI is being used for decision-making in general, and this drops to 14 percent and nine percent respectively when it comes to awareness of the use of automated decision systems in the workplace and in the criminal justice system.[4] On the whole, people aren't supportive of the idea of using AI for decision-making, and they feel especially strongly about the use of automated decision systems in the workplace and in the criminal justice system (60 percent of people oppose or strongly oppose its use in these domains).

The public's doubts about AI have yet to seriously impede the technological progress being made by companies and governments. Nevertheless, perceptions do matter; regardless of the benefits of AI, if people feel victimised by the technology rather than empowered by it, they may resist innovation, even if this means that they lose out on those benefits.

The problem may be, in part, that people feel decisions about how technology is used in relation to them are increasingly beyond their control. Moreover, they may not trust those who are making these decisions, as is clear from the Hansard Society's recent annual audit of political engagement.[5] What this suggests is that there may need to be a radical overhaul of the way in which organisations and institutions include and devolve power to citizens over these decisions. In a liberal democracy, these decisions should be made *with* the public, not just for the public. Such democracies should value innovation, human rights *and* public dialogue and voice. It is possible to imagine a different model where the private sector or the state alone drive innovation in service of the interests each holds dear. However, the RSA, with its ethos of 21st century enlightenment, believes that these alternative models are not the right way to proceed.

The RSA's Forum for Ethical AI is making the case for entering into a public dialogue with citizens about the conditions under which this technology is used. While human rights law serves to protect people from egregious violations, we also need to engage directly with people to address the wider problems of mistrust and disempowerment that can arise when only a few are making critical decisions on behalf of many.

When it comes to automated decision systems, for example, experts have called for the need to go beyond embedding individual or group judgment in these systems and to start encompassing the values of society as a whole.[6] This requires a public dialogue with citizens to resolve trade-offs; for example, trade-offs between privacy and security, or between different notions of fairness.[7] These ethical issues being surfaced by AI may ultimately lead to enacting new laws or policies, but they are also the reason why we should expect organisations and institutions (in both the private and public sectors) to fundamentally change the way they operate, engage with, and are accountable to citizens. In our public dialogue, the citizens might help stimulate thinking about what sort of reform is needed in terms of corporate and governmental structures, products, and services to minimise risks and secure benefits for more people.

The RSA wishes to see profound innovation for the public, by being with the public. If the public are going to unite in favour of innovations like AI, they need to be engaged early and more deeply. They need to feel confident that this technology is being deployed responsibly and will uplift individuals *and* communities at large.

In this paper, we first set out what we mean by ethical AI and why AI needs to reflect the public's values. We propose exploring the use of AI for decision-making with the public, expanding on the proposition for 'society-in-the-loop' systems. We then clarify what the process for public dialogue is, and in particular, long-form deliberation. We follow this by presenting the results of our online survey of the UK population's attitudes towards AI and automated decision systems, which were used to draw out key issues for deliberation with citizens. Finally, we detail what the RSA's public dialogue with citizens will look like in practice and clarify our next steps.

Embedding citizen voice in ethical AI

1

Although established as field of technology since the 1950's, swift progress has been made in recent years to further develop AI. In less than a decade, what were once nascent capabilities of AI, such as computer vision and natural language understanding, have evolved to rival the capabilities of humans.[8]

In some areas, AI has surpassed human capability; for example, when it comes to tasks such as recognising objects, or playing competitive games such as Go and Poker.[9] New approaches to developing AI, most notably deep learning, have accelerated advances in the technology, which in turn have stimulated greater investment and support for the growth of the industry.[10] AI is increasingly being entrusted to do more for us by companies and governments, and in a range of sectors such as healthcare and criminal justice.[11] However, as AI is used in new ways that could have significant consequences for individuals and communities, concerns about the ethics of AI are becoming more urgent.[12]

What do we mean by AI?

AI refers to machines that can perform tasks generally thought to require intelligence.

Most modern AI systems employ a technique known as 'machine learning', in which computers learn how to perform a specific task from examples, data and experience.[13] This is in contrast to traditional computer systems, which are explicitly told how to perform a particular task by human programmers.

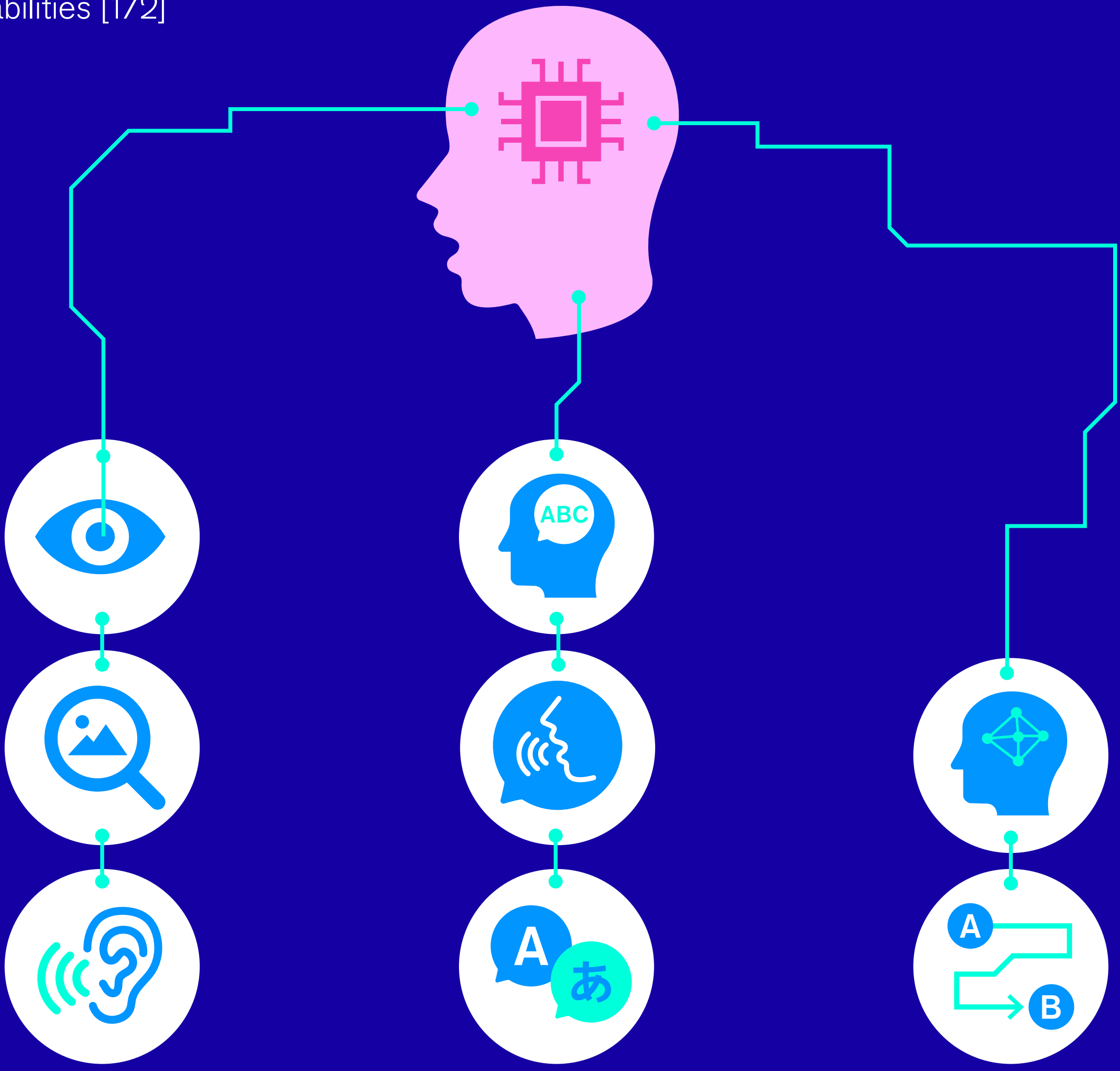
One of the most effective methods of machine learning developed so far is 'deep learning'. Based loosely on the structure of the brain, deep learning algorithms involve many layers of interconnected units which form a 'neural network'. The complexity of deep learning networks makes it impossible to understand exactly how they work, leading them to be described as 'black boxes'.

One task that AI is increasingly being used for is to make predictions about the likelihood of future events occurring. While predictions can be made using a variety of statistical techniques, machine learning is increasingly becoming the preferred tool due to its potential for greater accuracy.

As AI is used in new ways that could have significant consequences for individuals and communities, concerns about the ethics of AI are becoming more urgent.

Figure 1

AI Capabilities [1/2]



1. Perception

- Object recognition: being able to identify objects from visual information (including facial recognition)
- Scene analysis: understanding what’s going on in a visual scene (eg not just being able to identify that there’s a car and a human in a scene, but being able to understand that the car is about to run over the human).
- Speech recognition: being able to pick out speech from a soundscape

2. Natural Language Processing

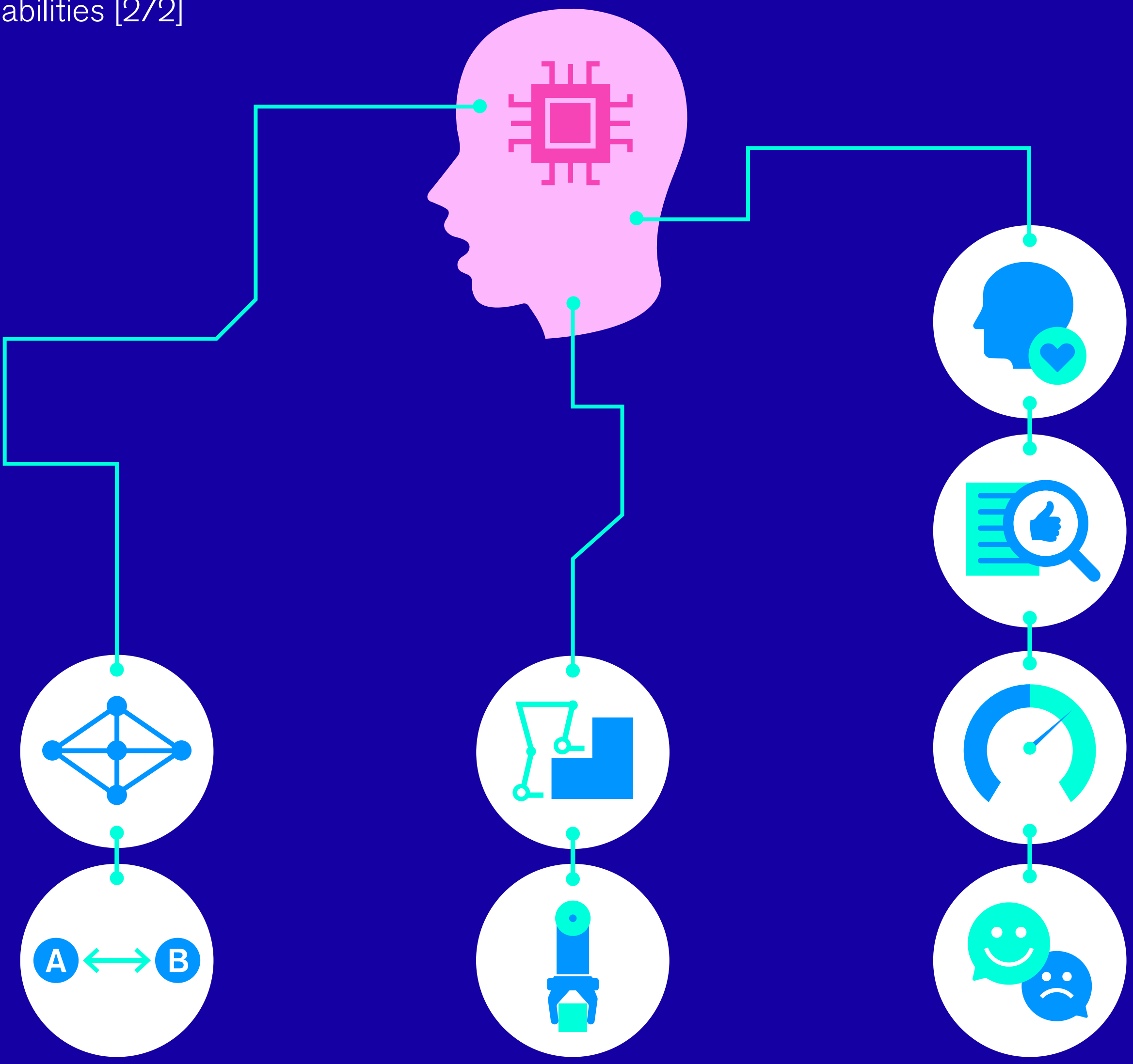
- Understanding language (text and speech)
- Generating language (text and speech)
- Translating from one language to another

3. Reasoning and Planning

- Making logical deductions (eg understanding that if “Socrates is a man” and “All men are mortal”, that therefore “Socrates is mortal”.
- Working out the optimal route to reach a specified goal (eg how to solve a multi-step puzzle in the fewest number of moves; how to drive from London to Manchester in the quickest possible time)

Figure 1

AI Capabilities [2/2]



4. Knowledge Representation

- Understanding the semantic relationships between concepts (eg that a chicken is a type of bird)
- Verbal reasoning (eg understanding that “man” is to “woman” as “boy” is to “girl”)

5. Locomotion and Manipulation

- Being able to move about in a physical environment, and across different physical environments (eg walking on sand and up mountains and up and down stairs)
- Being able to pick up and manipulate physical objects (eg using a pen, tying shoe-laces, shaving)

6. Affective / Emotional Capacities

- Recognising emotions expressed through facial expressions, body language and tone of voice
- Sentiment analysis: understanding the sentiment expressed in speech or text (eg understanding if a tweet contains a pro or anti Brexit message)
- Being able to predict someone's emotional reaction to a given event / action
- Being able to display emotions, (eg generate facial expressions or speech that displays appropriate emotions)

What do we mean by 'ethical AI'?

Ethical AI is garnering much interest, but it's not always clear what this refers to. A broad range of emerging issues have been identified as requiring ethical frameworks or principles in order to steer the development of AI in a socially beneficial manner, including:

- **AI safety:** Ensuring that autonomous systems do not behave in ways that inadvertently harm society.
- **Malicious uses of AI:** Guarding against the misuse of AI by malicious actors.
- **Data ownership and protection:** Overseeing the use of personal data for AI systems.
- **Algorithmic accountability:** Clarifying governance and responsibilities for the use of algorithms, such as in the case of automated decision systems.
- **Socio-economic impact:** Managing social and economic repercussions of AI, such as increased inequality of wealth and power.

Strikingly, many of the authors and organisations that are exploring these issues and developing related frameworks or principles have advocated public dialogue or engagement. For example:

- A report on malicious AI authored by a coalition of organisations, including the Future of Humanity Institute, the Centre for the Study of Existential Risk, and Open AI, advocated for a public dialogue on appropriate uses of AI technology.[14] The authors recommended actively seeking to expand the range of stakeholders and domain experts involved in discussions of the challenges, which they believe should include the general public alongside civil society, businesses, security experts, researchers, and ethicists.[15]
- The international and interdisciplinary research community known as Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) developed principles which suggest facilitating public auditing of algorithms.[16]
- Similarly, the Association for Computing Machinery expressed within their principles for algorithmic transparency and accountability that public scrutiny is ideal, particularly in relation to training data, in order to maximise opportunity for corrections.[17]
- Most recently, AI Now Institute called for public agencies to enable communities to review and comment on their use of automated decision systems, detailing the process as part of 'Algorithmic Impact Assessments'.[18]
- The Partnership on AI, which was founded by leading technology companies and now encompasses a number of academic and non-profit organisations, published tenets committing to educating and listening to the public; an open dialogue on the ethical, social, economic and legal implications of AI, and actively engaging with and being accountable to a broad range of stakeholders.[19]

To build on these ideas, we propose a working definition of what we mean by ethical AI:

> *AI that is designed and implemented based on the public's values, as articulated through a deliberative and inclusive dialogue between experts and citizens.*

We intend this definition to capture a number of elements that we consider to be necessary to achieve deployment of AI technology in a manner that is beneficial to society over the long-term, has moral and political legitimacy, and hence is grounded in widespread popular consent. These are:

1. In both design and implementation, AI is guided by values above short term profit.
2. Values should be based on our best understanding of society's value.
3. The most effective methods for building a shared and considered set of societal values bring together citizens in deliberative and inclusive dialogue with subject experts, such as technologists and philosophers.

Why does AI need to reflect the public's values?

The potential of AI to dramatically transform our lives is enormous, and this shift is already underway. The use of AI in the private sector is comparatively widespread,[20] but now public bodies are increasingly adopting the use of AI systems, expanding their reach and raising the stakes. In large part, AI is likely being used by public bodies to increase efficiency and reduce costs, but in some circumstances it may also be used to improve the fairness of outcomes and minimise biases in systems, particularly those that involve decision-making.

Using computers to make decisions thus far has not always met expectations, and in some instances it has exacerbated inefficiencies and reinforced inequalities. The academic Virginia Eubanks has exposed cases in the US in which the use of automated decision systems has further disadvantaged some of the most vulnerable groups in society. She investigated the use of these systems to determine eligibility for welfare, allocate social housing, and evaluate the risk of child abuse and neglect, finding that the process of how they reached their verdicts was often inexplicable (ie because it was not apparent to what extent algorithmic predictions influence human decision-makers).[21] She revealed that the data collected for these systems could be very intimate and serve to intensify state surveillance of the poor in particular.[22] Eubanks argues that the fundamental problem with these systems is that they enable the ethical distance needed "to make inhuman choices about who gets food and who starves, who has housing and who remains homeless, whose family stays together and whose is broken up by the state."[23]

While many of these systems have used more simple statistical techniques to date rather than AI, public bodies in the UK are exploring, and in some cases, experimenting with 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.[24]

Yet, as Eubanks demonstrates, given that the consequences of some of these systems are far from trivial, it is reasonable to consider whether greater public legitimacy is needed or clearer parameters should be established before they become more widespread as a result of advances in AI. After all, it's not the accuracy of these systems that is the primary concern; it is the ethics of using the systems under particular circumstances or conditions. It's important for the public to have an opportunity to engage with the trade-offs of using AI in these ways and to express their views and values.

When we speak of AI systems based on the public's values, we are referring to exploring how citizens understand the contemporary use of AI and how they apply ethical reasoning to how it should, or should not be, used in the delivery of private or public services.[25] This includes citizens' views on how these institutions should demonstrate transparency and accountability to citizens who will be directly affected by their use of AI.

In public dialogue, it's recognised that underlying values help us to understand why citizens hold particular opinions or perspectives. Values are specifically defined as "(a) concepts or beliefs; (b) about desirable end states or behaviours; (c) that transcend specific situations; (d) guide selection or evaluation of behaviour and events, and (e) are ordered by relative importance."[26] Values underpin people's preferences for one course of action over another, and, in turn preferences are premised on what people believe about how actions will affect the things they value.[27]

While citizens are unlikely to all share the same views, a dialogue can enhance mutual understanding of facts and values, as well as value differences.[28] Although citizens do not have to reach a consensus, there is some evidence that reflection about and articulation of value positions can reduce conflict and enable compromise.[29]

When it comes to controversial uses of AI, the public's views and, crucially, their values can help steer governance in the best interests of society. Citizen voice should be embedded in ethical AI.

What should the public be engaged on?

There are many ethical issues that a public dialogue could address. For example, autonomous vehicles and weapons are currently capturing the public's imagination. Both are being developed and attracting investment in research, although neither is commercially available yet. These systems raise questions about how much power should be ceded to AI over human life.

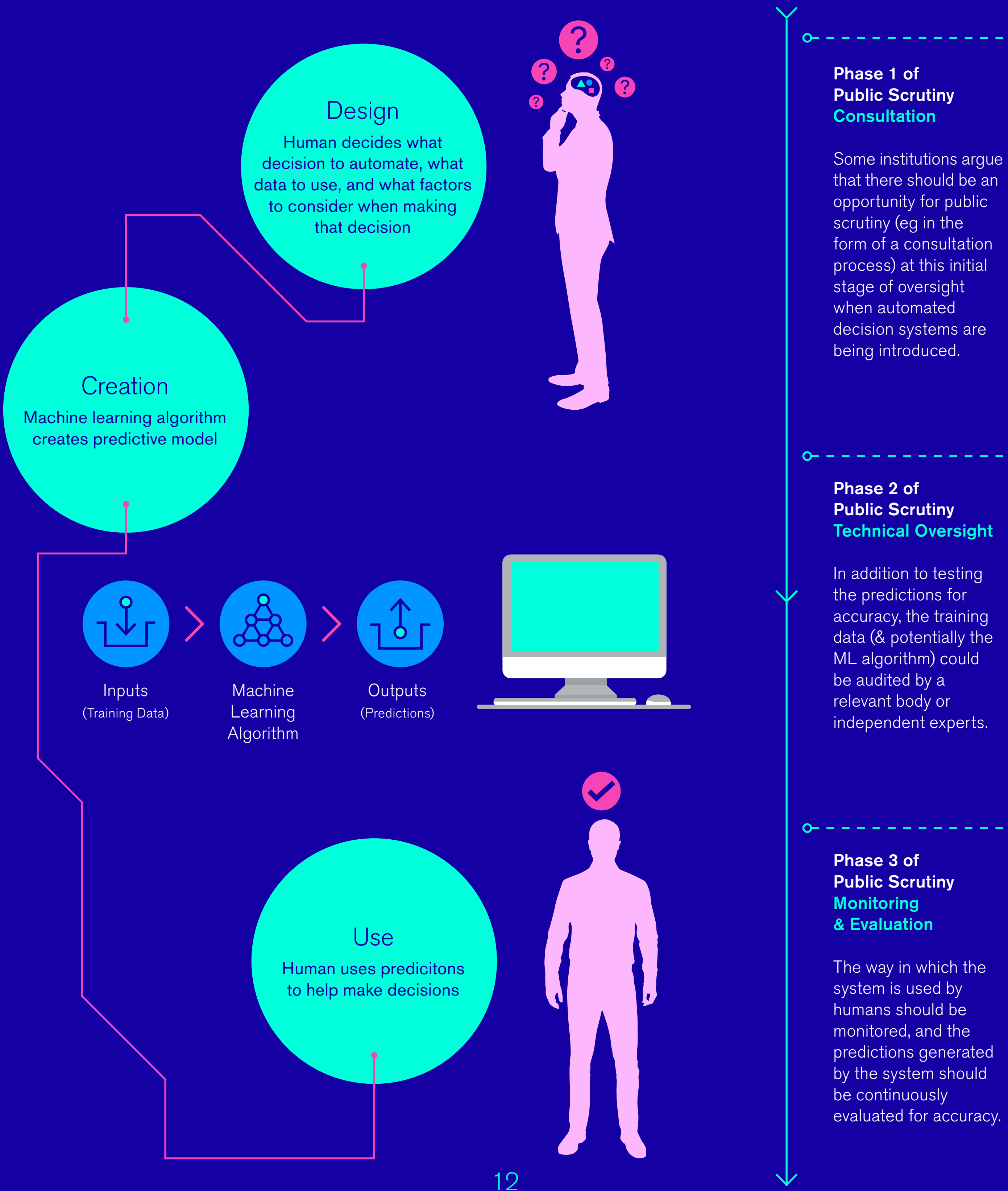
The dual-use nature of AI is also increasingly of concern as it has has become apparent that technology designed with one purpose in mind can be exploited for different and, possibly, more malevolent aims. As researchers have observed, "Surveillance tools can be used to catch terrorists or oppress ordinary citizens. Information content filters could be used to bury fake news or manipulate public opinion. Governments and powerful private actors will have access to many of these AI tools and could use them for public good or harm."[30] An example of this might be the experimentation with 'social credit scores' in China, which are ratings assigned to every citizen based on government data regarding their economic and social status.[31]

All of these issues could inspire meaningful public dialogue. However, the **RSA's Forum for Ethical AI is choosing to apply a process of citizen deliberation to explore the rise of automated decision systems.** These systems have been characterised as 'low-hanging fruit' for government and we anticipate more efforts to embed them in future.[32]

The RSA's Forum for Ethical AI is choosing to apply a process of citizen deliberation to explore the rise of automated decision systems.

Figure 2

Understanding the process of making an automated decision



Design

Human decides what decision to automate, what data to use, and what factors to consider when making that decision

Creation

Machine learning algorithm creates predictive model

Use

Human uses predicitons to help make decisions

Phase 1 of Public Scrutiny
Consultation

Some institutions argue that there should be an opportunity for public scrutiny (eg in the form of a consultation process) at this initial stage of oversight when automated decision systems are being introduced.

Phase 2 of Public Scrutiny
Technical Oversight

In addition to testing the predictions for accuracy, the training data (& potentially the ML algorithm) could be audited by a relevant body or independent experts.

Phase 3 of Public Scrutiny
Monitoring & Evaluation

The way in which the system is used by humans should be monitored, and the predictions generated by the system should be continuously evaluated for accuracy.

A whole systems approach to automated decision-making

2

In the opening chapter, we set out what we mean by ethical AI and why AI needs to reflect the public's values. Now, we explore why public dialogue on the use of automated decision systems specifically would be useful.

Automated decision systems refer to computer systems that either inform or make a decision on a course of action to pursue about an individual or business.[33] To be clear, automated decision systems do not always use AI, but increasingly draw on the technology as machine learning algorithms can substantially improve the accuracy of predictions.[34]

It is important to examine the use of automated decision systems in the broader social and economic context, considering behavioural insights, cultural norms, institutional structures and governance, economic incentives and other contextual factors that have a bearing on how an automated decision system might be used in practice.

A 'whole systems' approach to automated decision-making

At present, it is rare that decisions are fully automated; these systems are typically used as part of a wider process of decision-making that involves human oversight, or a 'human-in-the-loop' (HITL). Iyad Rahwan of the MIT Media Lab describes the use of human operators in HITL systems as potentially powerful in regulating the behaviour of AI. He explains that HITL systems serve two functions: to identify misbehaviour by otherwise autonomous systems and to take corrective action; and/or to be an accountable entity in case the systems misbehave. In the latter scenario, the human operator encourages trust in the system because someone is held responsible and expected to own up to the consequences of any errors (and therefore, is incentivised to minimise mistakes).[35]

Rahwan builds on the concept of HITL, proposing the idea of 'society-in-the-loop' (SITL) systems that go beyond embedding the judgment of individual humans or groups in the optimisation of AI systems to encompass the values of society as a whole. SITL systems do not replace HITL systems but are an extension of them; they incorporate public feedback on regulations and legislations rather than individual feedback on micro-level decisions. They are therefore particularly relevant when the impact of AI has broad social implications; for example, as is the case with algorithms that filter news, wielding the power to politically influence scores of voters.

As part of designing SITL systems, consideration is given to the question of how to balance the competing interests of different stakeholders. Society is expected to resolve the trade-offs between the different values that are embedded within AI systems (for example, as highlighted by Rahwan, trade-offs between security and privacy, or between different notions of fairness) as well as agree on which stakeholders should reap certain benefits and which should pay certain costs.

Our proposition is that public deliberation is an essential component of developing effective SITL systems.

The RSA has previously proposed a similar 'whole systems' approach to resolving social and economic trade-offs in the fields of corporate governance,[36] regulating digital platforms[37] and formulating economic policy,[38] as well as understanding how innovation happens.[39] In relation to the ethical use of AI, we argue that the context in which automated decision systems are used is as important as the design of the systems themselves. Both have a bearing on the outcomes of automated decision systems, and therefore both have a bearing on the social and ethical acceptability of those outcomes.

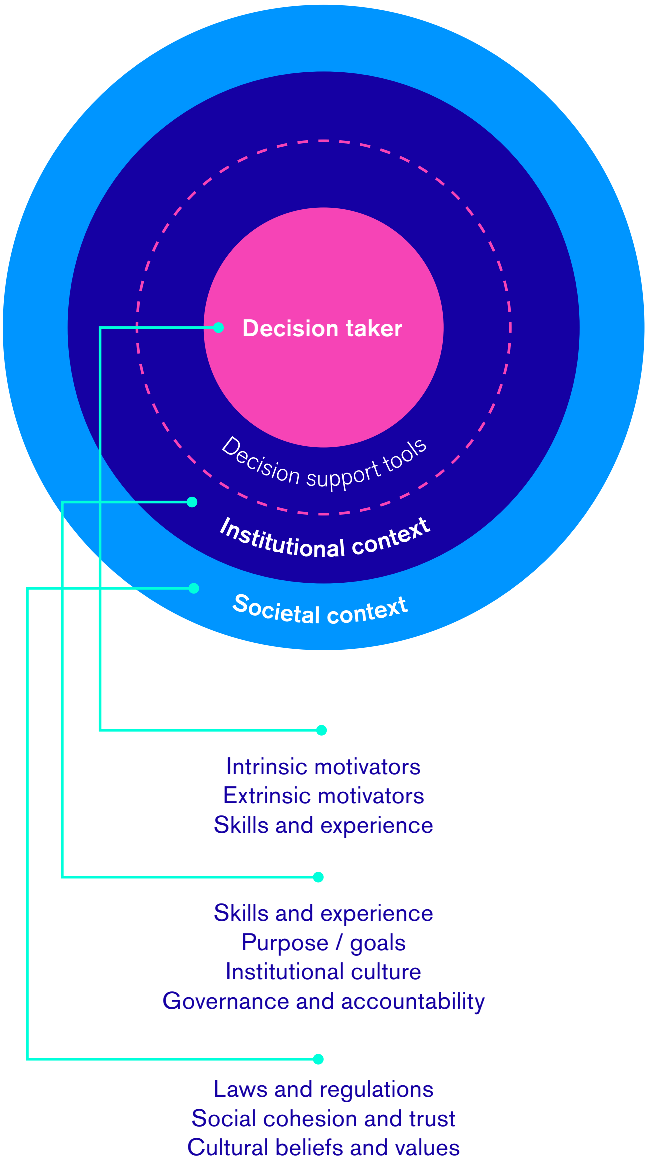
Automated decision systems in context

Drawing on the concepts of HITL and SITL systems, the decision-making context can be conceptualised as three tiers:

1. the decision taker, who may or may not be human;
2. the institution that is ultimately accountable for the decision;
3. and the societal context in which that institution is operating.

The three tiers and key factors influencing each tier are summarised in Figure 3.

Figure 3:
Decision-making in context: a whole systems view



The decision taker

Assuming for now that the decision taker is human, there are many factors that influence a decision other than the raw data on which the decision is based. These include intrinsic factors, such as the individual's own values and beliefs, and extrinsic factors such as the financial and social rewards or penalties faced by the individual as a result of the outcomes of the decision. They also include the individual's skills and experience as applied to managing data and reaching a decision.[40]

The institutional context

The next tier is the institution that is accountable for the decision.[41] The goals of the institution, and the culture and internal incentives that determine how those goals are pursued, have a significant influence on the decision taker. Equally, the governance structure, transparency and accountability of the institution to wider stakeholders and society will in turn influence the institution's internal goals, culture and incentive structures.

Intermediating between the decision taker and the institution may be a suite of decision support tools that are provided by the institution. These may be internal training, manuals or guides, expert systems, or other tools that help the decision taker manage data and follow a rules or principles based process for reaching a decision.

The societal context

Finally, both the individual and the institution will be influenced by societal context in terms of hard factors such as laws and regulations, and softer ones such as cultural norms, moral and religious belief systems, and sense of social cohesion and solidarity. Identifying the societal context may not be easy, especially for global organisations; although it will often approximate to a nation state, it may also be sub-national (eg London, California) or supra-national (eg European, Roman Catholic). For simplicity we have not sub-divided societal context, but the question of how global, national and local cultural norms and laws interact is one to which we expect to return.

A couple of key observations emerge from this conceptualisation of the whole decision-making context.

First, AI can be introduced in two ways. It can be used as part of a decision support tool to help a human make a decision, or it can replace the human decision-taker entirely. Even in this latter case, the AI decision-taker exists within an accountable institution that is ultimately governed by humans.[42] Therefore, there will inevitably be a HITL system bridging the institution and the automated decision system. However, this may not be well defined or governed.

Second, SITL systems bridge the societal context and the accountable institution, reinforcing the idea that they are complementary to HITL systems rather than an alternative to them. Unlike HITL which is baked in to institutional structures, SITL systems are not well developed in existing private, NGO or public institutional structures and so this seems to be the area of most potential and greatest urgency.

In recent months, numerous academics and organisations have suggested detailed processes or made concrete recommendations that reflect the concept of SITL and suggest what it could look like in practice for automated decision systems. In particular, there are several variations of 'impact statements' or 'impact assessments' that explicitly call for public review and engagement in algorithmic governance. For example:

- FAT/ML wrote a 'Social Impact Statement (SIS) for Algorithms' to accompany their principles, advocating that those who create algorithms should also publish a statement about the social impact of the system so that the public can know what to expect.[43] FAT/ML urges creators to draw on their principles and includes a set of questions and steps that should be answered and adhered to when drafting a statement.

- In anticipation of the mayor of New York City, Bill de Blasio, announcing a task force on automated decision systems, AI Now Institute constructed a framework for carrying out 'Algorithmic Impact Assessments (AIAs)'.[44] The researchers note that they directly drew on impact assessment frameworks in environmental protection, data protection, privacy, and human rights policy domains to produce a framework that they hope will similarly help agencies and the public to consider complex social and technical questions as automated decision systems are adopted. AIAs set out a five-stage process of governance with the intention of supporting affected communities and stakeholders to assess the claims made about these systems, and ultimately, to determine where, if at all, their use is acceptable.

- An initial outline of AIAs prompted Michael Karlin of the Treasury Board of Canada Secretariat to contemplate what form a 'Canadian Algorithmic Impact Assessment' would take.[45] Karlin emphasises that the Government of Canada must consider more than protecting the rights of individuals and also balance broader concerns, including how these systems will impact communities, the environment, the ability of individual businesses to succeed, and the health and competitiveness of markets. He drafted a questionnaire as the basis of an AIA that asks two questions of programme officials in government seeking to use automated systems: what impact will the system have on various aspects of society or the planet, and how much judgment will the system will be delegated (ie is there a human-in-the-loop). He remarks that broad expertise is needed to respond to the questions, therefore requiring institutions to collaborate with others. Moreover, he invites comment on his draft AIA, acknowledging that the questionnaire must be scored in a way that is reflective of a diverse set of priorities and worldviews.

These interventions seem promising, and this project adds to the developing field by examining and experimenting with a process of deeper, deliberative engagement that would be appropriate during suggested 'comment' or consultation periods with the public on these systems.

Understanding the role of public dialogue

3

In the second chapter, we described the societal context for decisions made within institutions, introduced the concept of 'society-in-the-loop' for governing the use of AI and raised the potential for requiring impact assessments for the application of AI systems. This is why we argue that there is an urgent need to explore how public dialogue can be applied to the ethics of AI. In this chapter, we expand on what we mean by public dialogue, clarifying the methodology, and what the process can accomplish.

What do we mean by public dialogue?

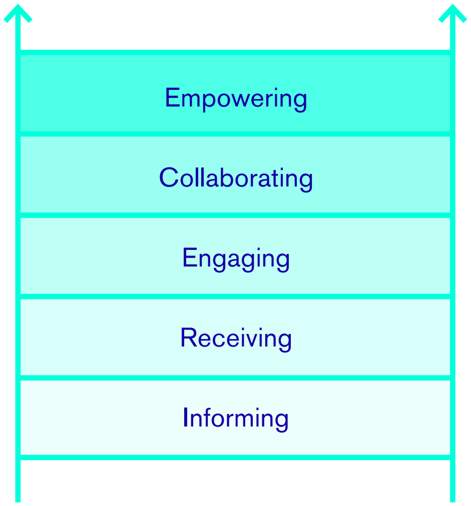
The theory and practice behind such public participation in policy-making is already established, for example, in the field of planning and environmental impact (the Aarhus Convention),[46] science policy (Sciencewise)[47] and health policy.[48] The case for public participation sits within a broader school of post-positivist theory that challenges the notion of neutral and rational technocratic policy making; this theory instead emphasises the normative nature of policymaking and, thus, the need for integrating deliberative dialogue in governance alongside empirical analysis and logical reasoning.[49]

It has been pointed out that in UK policy documents, dialogue is often used as a synonym for conversation, consultation, collaboration, participation, dissemination, and deliberation.[50] However, among practitioners, dialogue refers to a specific form of engagement that typically involves convening citizens and expert stakeholders to deliberate, reflect, and come to conclusions on public policy issues.[51]

Involve, a leading organisation for deliberative democracy in the UK, advises that public dialogue should enable a diverse mix of participants with a range of views and values to learn about the issues (eg from written 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.[52]

There are different degrees of public engagement ranging from the transmission of information to the full concession of decision-making to a public forum or electorate. This was conceptualised by Arstein as a 'ladder of participation' with eight rungs. [53] A more modern interpretation has been developed by the International Association of Public Participation (IAP2) with five different degrees on a spectrum of participation.[54] At the RSA we adopt IAP2's terminology but within Arstein's original visualisation (see Figure 4).

Figure 4:
The Ladder of Participation



Our focus is on long-form deliberative processes. These sit on the top two rungs of the ladder, 'empowering' and 'collaborating', which can include citizens' juries in addition to citizens' assemblies, reference panels, and commissions. In a review of long-form deliberative processes, Claudia Chwalisz distinguishes long-form deliberative processes by the following characteristics:[55]

- Citizens are tasked with helping to resolve a pressing problem that requires navigating multiple trade-offs and considering more than one possible and realistic solution (and this solution is not pre-determined).
- This group of citizens is a small group (in numbers between 24 and 48) who are randomly selected from a local, regional or national community.
- The group spends a generally long period of time (eg a few sessions over the course of two to three months) learning about and deliberating on a policy issue from different angles.
- Citizens are not asked for their individual opinion on an issue, but to deliberate on behalf of their community with the aim of reaching a consensus or compromise.
- The group produces concrete recommendations for decision-makers, who then respond directly and publicly to the proposals.

Crucially, long-form deliberative processes should not be confused with focus groups or consultations. They are not 'one-way' exercises in which citizens are only asked for their own opinions on an issue; rather, they are 'two-way' conversations between experts, decision-makers and the public in which ideas are exchanged (and often respectfully challenged) in order to reach a conclusion in a collaborative manner.[56]

Mass LBP, an organisation that pioneered long-form deliberative processes in Canada, makes the case that such efforts are an innovative approach to public engagement and consultation.[57] The organisation's founder Peter MacLeod argues that policymakers tend to be misguided about how to consult the public, typically convening town hall meetings "when something's gone wrong, or a decision has already been made and an elected official is trying to explain it."[58] These consultations tend to be dry and technocratic, leaving little room to explore people's feelings about an issue.[59]

We can take this analysis of consultations further, observing that they tend to be reactive, reflecting a failure of decision-makers to take the long-term view preferred by citizens.[60] Ideally, public dialogue should be far more proactive, inviting citizens and experts to explore emergent issues of importance.

What does a public dialogue accomplish?

Public dialogue is useful when a topic is controversial or complex (involving difficult choices to make or many trade-offs to consider).[61] It is especially valuable when it raises important ethical and social questions that cannot be resolved with facts alone.[62]

This may seem counterintuitive to some who assume that advanced or specialised knowledge (eg at degree level) must be required to draw meaningful conclusions about these sorts of topics, but there are many examples where people have successfully engaged in very complicated and contentious issues. As highlighted by Involve, these include developing an alternative voting system, redrafting the Icelandic constitution, rebuilding New Orleans, forming domestic violence courts in New York, and managing the Federal Deficit in the US.[63] It is argued that citizen input is needed precisely because these topics are so difficult,[64] and to make progress when some sense of public buy-in or legitimacy is required. If there is a clear question that the public can help answer, citizens' juries are especially ideal.

Table 1:
Advantages vs the limitations and challenges of citizens' juries

Advantages
Enables direct input from citizens on topics of a social and ethical nature
Focused on a single well-defined question
Impartial and objective
Allows citizens to hear from, challenge and question expert witnesses
Provides time for extensive deliberation
Can focus political/organisational attention on public views
Can inform other research into the topic: for example, surveys

Limitations / challenges
Defining the role of the citizen – are they there as an individual or as a representative of society?
Framing the question and evidence neutrally and impartially
Providing an adequate breadth of evidence and opinion
Respecting emotional as well as rational responses to a topic
Limited number of people: potential for selection bias
Ensuring that the jury findings have impact

Source: Compiled by Diane Beddoes, Director of Deliberate Thinking

Some people may question how impactful a public dialogue can be given the scale of the groups assembled for citizens' juries in particular. There may be concerns about whether such small groups are likely to be representative of one's own views and values. It is thus important to clarify that there is a distinction between representation and representativeness. We are asking these citizens to represent their community to encourage them to consider more than their own, individual interests, but we are not claiming that they are statistically representative of that community. Rather, we are suggesting that there are relevant insights to be drawn from a diverse group of citizens who are given the opportunity to enter into an informed and deliberative dialogue. Similar logic underpins the use of juries for criminal trials, in which lay members of the public are chosen to reach a verdict rather than trained legal experts.[65]

Public dialogues which are long-form deliberative processes, such as citizens' juries, will ultimately make recommendation(s) to be enacted. The organisation that convenes the dialogue does not commit to acting on these recommendations; rather, this is expected of the relevant institutions and organisations with influence and authority. However, if these recommendations aren't acted upon, the convenor will explain why this is the case to the citizens, and still publicise the process and findings widely to help broaden and enrich public debate.

We can now see that contemporary proposals for public dialogue on AI stand on an enormous body of academic literature and established practice. However, while the theory and techniques are not new, we sense that there is novelty in the subject matter for dialogue. The pace at which AI is being developed, its potentially pervasive and significant impacts on society, and the sense that the capabilities of AI are outpacing the ability of political and public discourse to keep up with the ethical issues that might arise create a pressing case for prototyping public dialogues on the ethics of AI. This view is reinforced by the recent report of the House of Lords Select Committee on Artificial Intelligence which identified a range of opportunities and risks from the development of AI and concluded that, "[t]he transformative potential for artificial intelligence on society at home, and abroad, requires active engagement by one and all."[66] The RSA's Forum for Ethical AI seeks to make a contribution to bring this active engagement into being.

Public dialogue is useful when a topic is controversial or complex. It is especially valuable when it raises important ethical and social questions that cannot be resolved with facts alone.

Engaging the public on automated decision systems

4

As we set out above, in our dialogue, we are focussing on the application of automated decision systems, and in particular those that make use of AI. As a starting point for further research, we partnered with YouGov to carry out an online survey of 2,000 people, a sample representative of the UK population. In this chapter, we analyse the results and draw out key issues for public dialogue.[67]

We set out to first understand how familiar the general public is with the use of automated decision systems and to what extent they support them (either based on their previous knowledge or on the basic information we provided. For exact wording of questions, please see appendix). Our survey questions gauged the public's familiarity with AI in general and automated decision systems in particular before uncovering their levels of concern and support for these technologies.

See Figure 5: Familiarity with AI

We found that most people are familiar with uses of AI that are widely debated and depicted in the media, or with AI that is designed for consumer or household use. For example, although no one yet owns a self-driving car, the majority of people (84 percent) are aware of them as a use of AI. Similarly, people seemed to be more conscious of AI that they can observe or interact with, such as digital assistants that they can speak to and ask questions (eg Siri and Alexa) and AI that can identify people in photos or videos (eg tag photos of their friends on Facebook). Considering this, we may have expected more familiarity with 'chatbots' (only 46 percent are familiar with chatbots), but there may be a distinction here because it's not always clear when you are exchanging messages with a bot or a human online.[68]

In contrast, people are not very familiar with AI that hums along in the background and may be integrated as part of other systems. Roughly only a third (32 percent) of people are aware that AI is being used as part of automated decision systems.

Whether or not a person is familiar with certain uses of AI appears to depend on its degree of visibility.

See Figure 6: Familiarity with uses of automated decision systems

We also asked people about their familiarity with the use of automated decision systems specifically, listing a range of different examples drawn from real-life case studies. Overall, most people are not very familiar with the use of automated decision systems. People were least familiar with the use of automated decision systems in the criminal justice system – 83 percent were either not very familiar or not at all familiar with its use.

There were also high numbers of people who lacked familiarity with the use of these systems to make decisions about immigration (78 percent were unfamiliar); in the workplace (77 percent); in healthcare (75 percent); and about claims for social support (73 percent).

Of the minority that were familiar with these systems, they tended to be most aware of systems designed for consumer markets (eg in financial services to determine credit ratings) or those with much more coverage in the media (eg for the curation of content and advertisements by social media companies).[69]

See Figure 7: Support for uses of automated decision systems

We wanted greater insight into how supportive people were of the idea of using automated decision systems for each of the specific purposes we outlined.

It appears to us that the less familiar people were with the use of the system, the less likely they were to support it. People were least supportive of the systems in both the criminal justice system (with 60 percent either opposing or strongly opposing its use) and the workplace (60 percent), for example, and either supportive of or indifferent to these systems in finance (27 percent supporting and 28 percent indifferent) or social media (26 percent supporting, 36 percent indifferent).

However, even in cases where people are more familiar with uses of AI, for example – advertising and social media (49 percent familiar) and personal finance (40 percent familiar) – the degree of support does not increase in line with the degree of familiarity (26 percent and 27 percent supporting respectively). Overall, automated decision systems have a low level of public support relative to much higher levels of opposition.

Figure 5

Q: Before taking this survey which, if any, of the following uses of AI were you aware of?

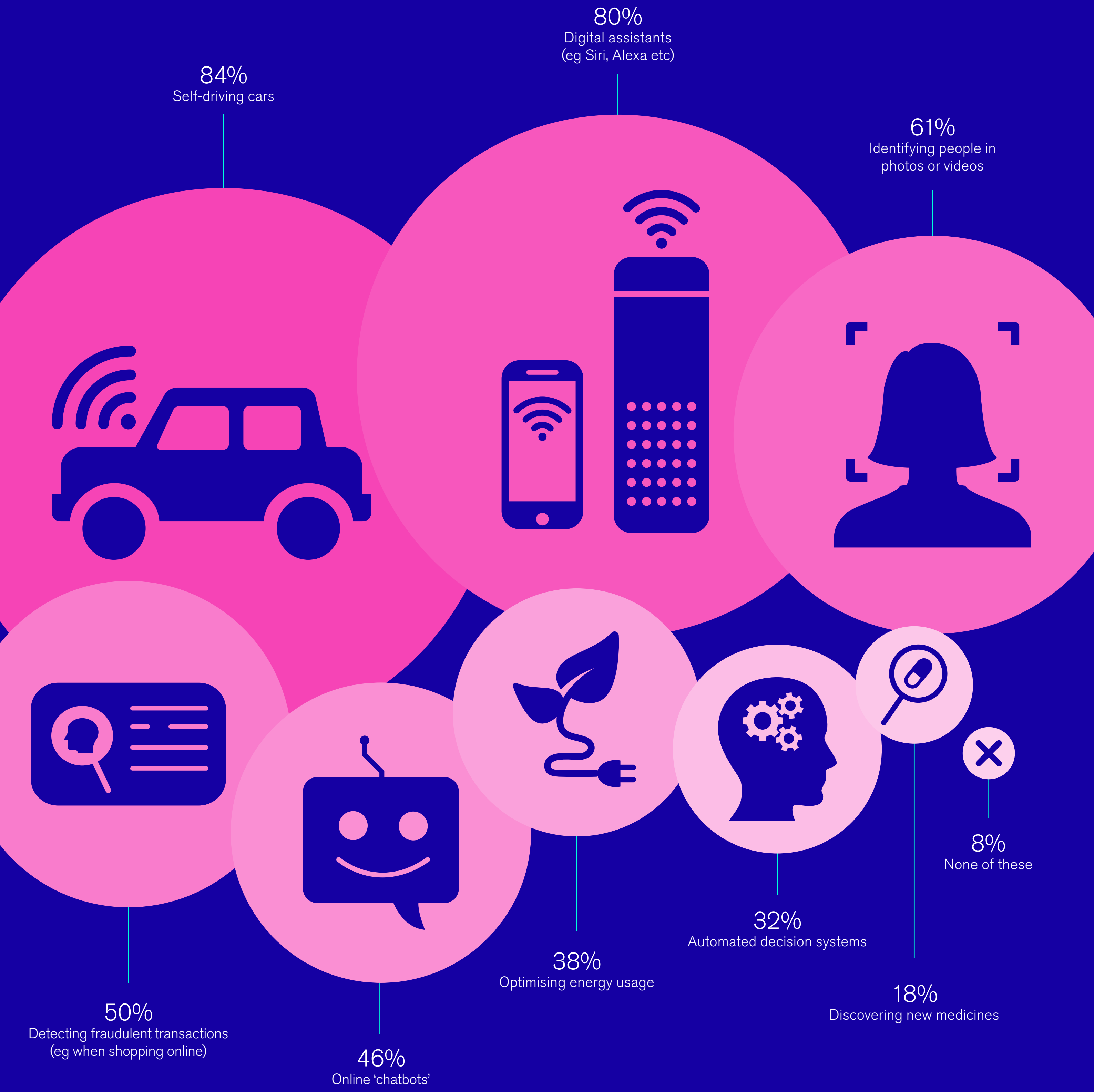


Figure 6

Q: How familiar, if at all, were you with the idea of automated decision systems being used to aid each of the following decisions?

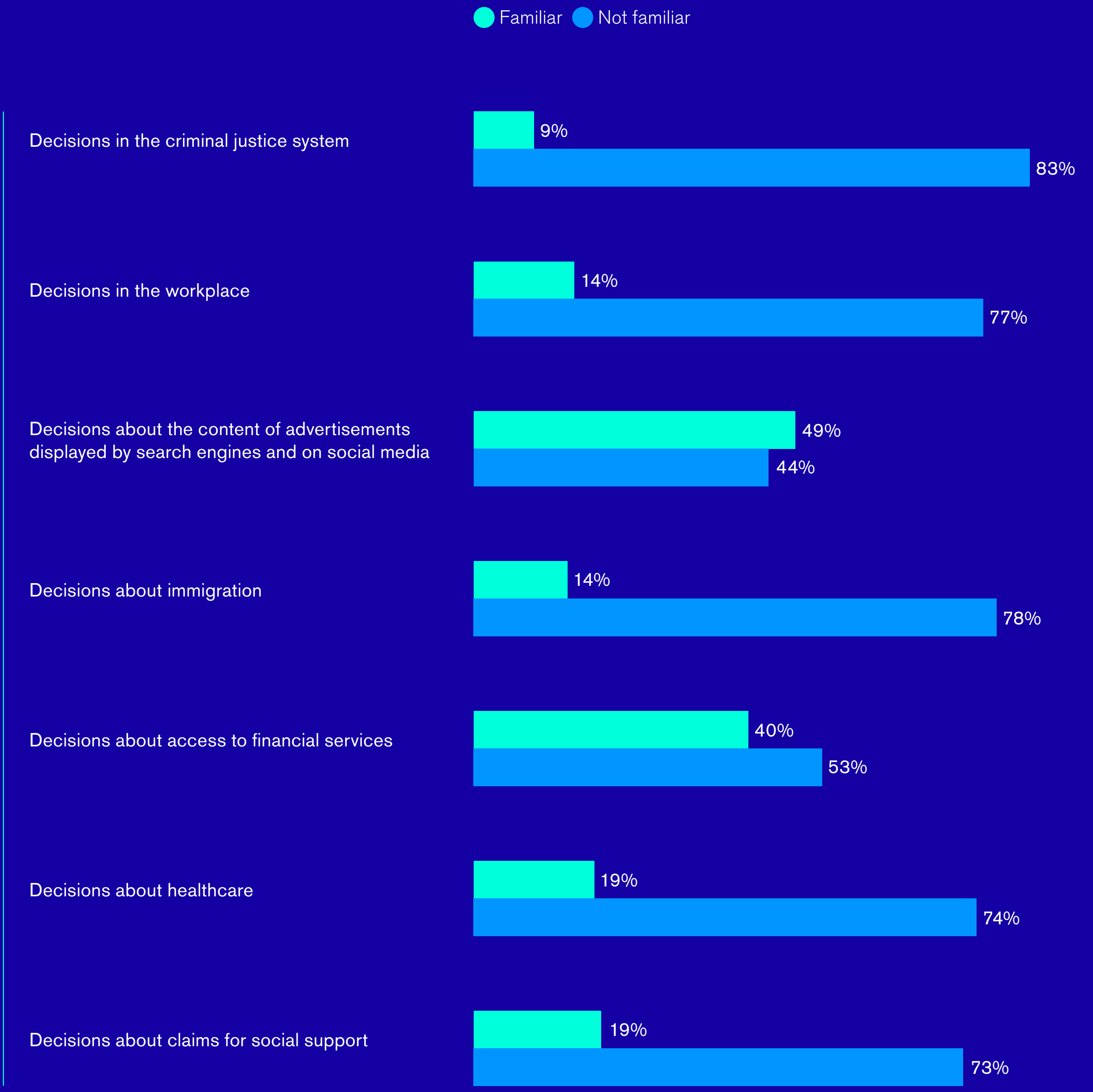
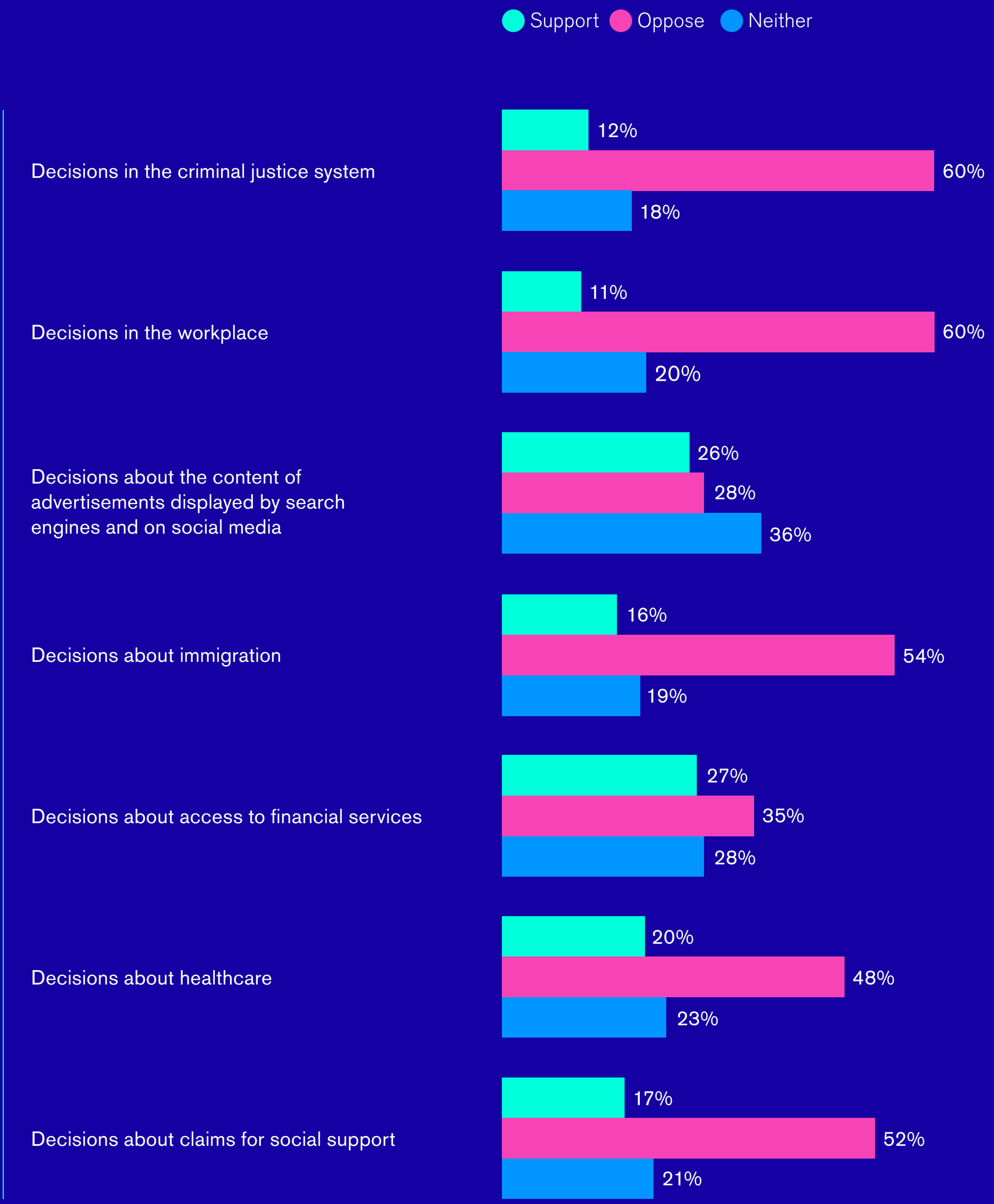


Figure 7

Q: To what extent, if at all, would you support or oppose the use of automated decision systems to aid each of the following decisions?



See Figure 8: Concerns about automated decision systems

To learn more about the reasons for their lack of support, we asked people about what most concerned them about these systems. They were asked to pick their top two concerns from a list of options.

Although we made it clear within the question that automated decision systems are currently only informing human decisions, there was still a high degree of concern about AI's lack of emotional intelligence. Sixty-one percent expressed concern with the use of automated decision systems because they believe that AI does not have the empathy or compassion required to make important decisions that affect individuals or communities. Nearly a third (31 percent) worry that AI reduces the responsibility and accountability of others for the decisions they implement.

These concerns broadly echo that of Eubanks' fears about relying too heavily on these systems when making morally challenging decisions and ceding more power to machines.

There were very few people who are relaxed about the rise of AI in their lives, with only six percent saying they were not particularly concerned about any potential problems with the use of automated decision systems.

See Figure 9: Potential of automated decision systems

We wanted to know which potential benefits, if any, the public is most looking forward to about the use of automated decision systems. They were asked to pick their top two potential benefits from a list of options.

We found that people were most looking forward to improved accuracy and consistency of decisions (31 percent), as well as increased efficiencies and savings made by the use of these systems by governments and companies (23 percent).

Although many researchers have expressed that these systems are promising because they may be able to reduce bias and inequality, only 19 percent of people regarded it as one of the top two benefits. It could be that some people felt that increased accuracy and consistency in decision-making may mean reducing bias by default.

Significantly, about a third (30 percent) of people stated that there was nothing about automated decision systems that they were looking forward to. Older people in particular were much more likely to state this (41 percent of 55+ year olds).

See Figure 10: Increasing support for automated decision systems

To gauge what might possibly increase support for the use of automated decision systems, we asked respondents to consider whether the following actions or policies would make a difference to them. They were asked to select all that apply.

Thirty-six percent noted that their support for these systems would increase if they were granted the right to request an explanation of the organisational steps or processes undertaken to reach a decision with an AI system. Fewer people (20 percent) noted that their support would increase if the technology was only used if it could be explained to the lay person (ie someone with no technical expertise).

A third (33 percent) would feel more supportive if penalties, such as fines, are introduced for organisations who fail to comply with monitoring or auditing these systems appropriately. Notably, 29 percent were not swayed to lend more support by any of these actions or policies. This is very similar to the proportion, 30 percent, that said they were not looking forward to any potential benefits of automated decision systems.

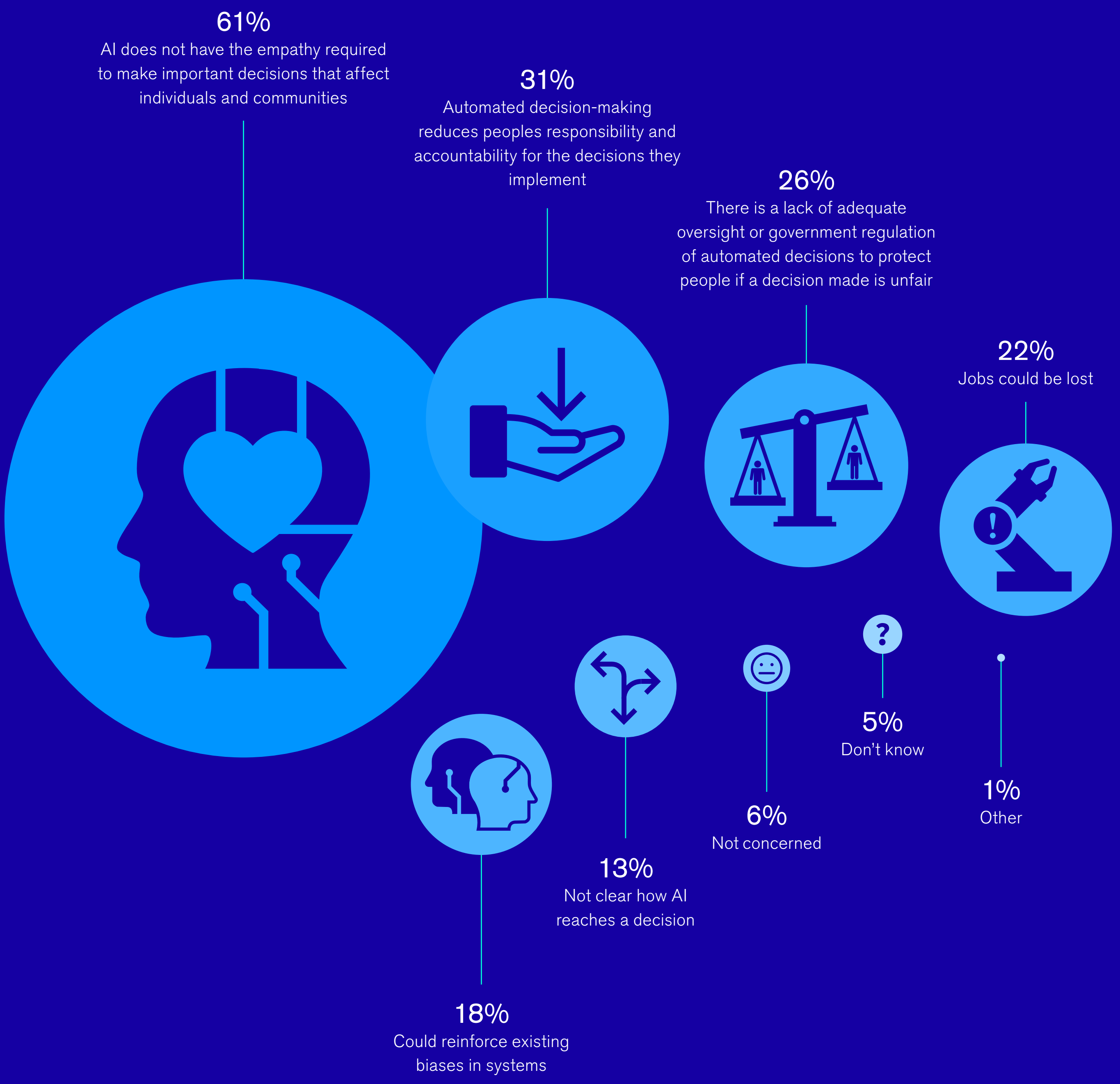
See Figure 11: Comfort levels with fully automated decisions

Finally, as AI becomes more accurate and consistent over time, there is scope for more decisions to be fully automated without the need for human intervention. This means that it would not be necessary for a human to make the final decision, which would be wholly based on the prediction made by the automated decision system. However, humans would still have a role in monitoring and auditing these systems to ensure they are working as intended.

At present, 64 percent of people are uncomfortable with this idea, and of this group, 26 percent are not at all comfortable.

Figure 8

Q: Which TWO, if any, of the following potential problems of using automated decision systems would you be MOST concerned about?



61%
AI does not have the empathy required to make important decisions that affect individuals and communities

31%
Automated decision-making reduces peoples responsibility and accountability for the decisions they implement

26%
There is a lack of adequate oversight or government regulation of automated decisions to protect people if a decision made is unfair

22%
Jobs could be lost

18%
Could reinforce existing biases in systems

13%
Not clear how AI reaches a decision

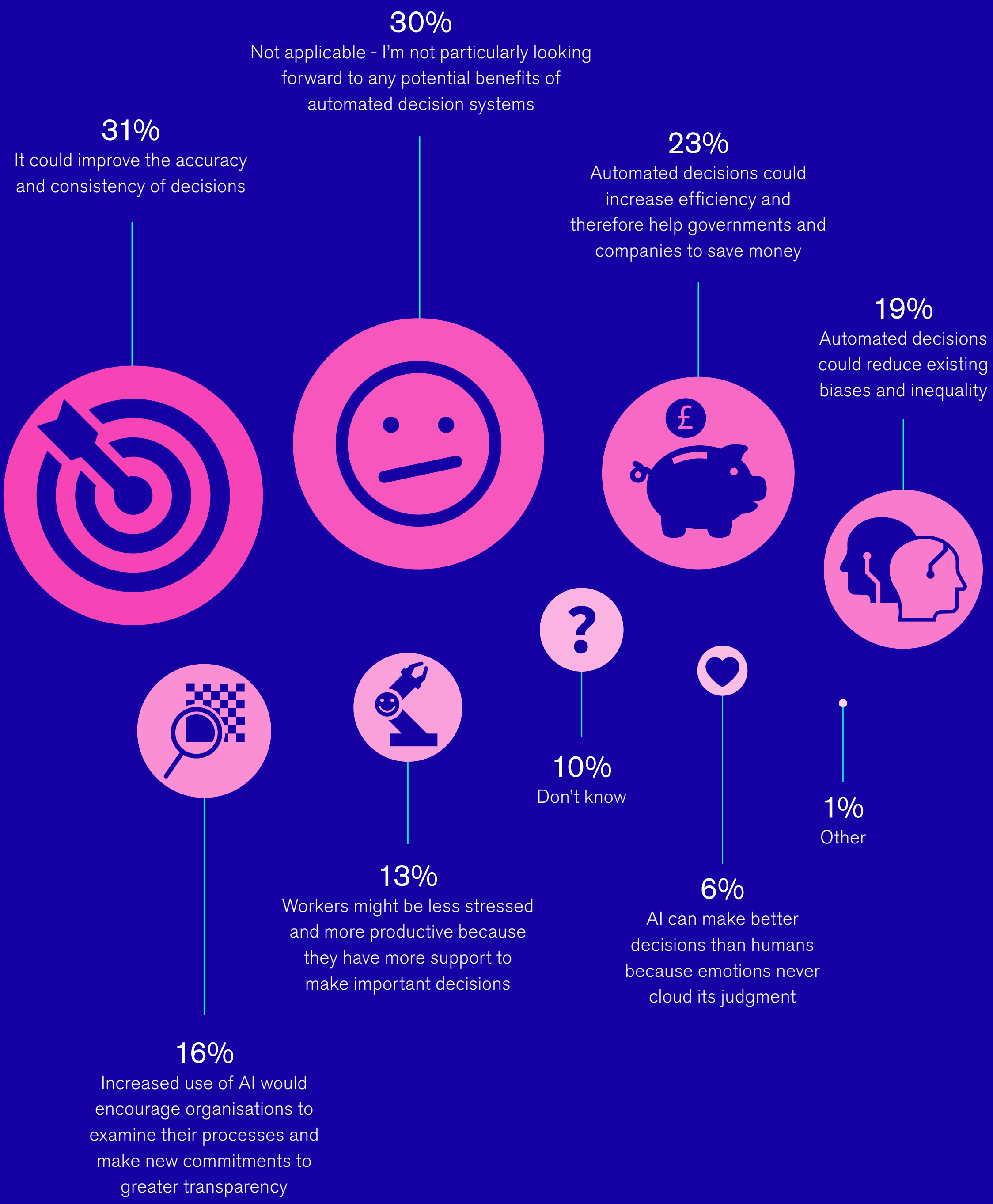
6%
Not concerned

5%
Don't know

1%
Other

Figure 9

Q: Which TWO, if any, of the following potential benefits of using automated decision systems would you MOST look forward to?



30%
Not applicable - I'm not particularly looking forward to any potential benefits of automated decision systems

31%
It could improve the accuracy and consistency of decisions

23%
Automated decisions could increase efficiency and therefore help governments and companies to save money

19%
Automated decisions could reduce existing biases and inequality

10%
Don't know

1%
Other

13%
Workers might be less stressed and more productive because they have more support to make important decisions

6%
AI can make better decisions than humans because emotions never cloud its judgment

16%
Increased use of AI would encourage organisations to examine their processes and make new commitments to greater transparency

Figure 10

Q: Generally speaking, which, if any, of the following would increase your overall support for automated decision systems?

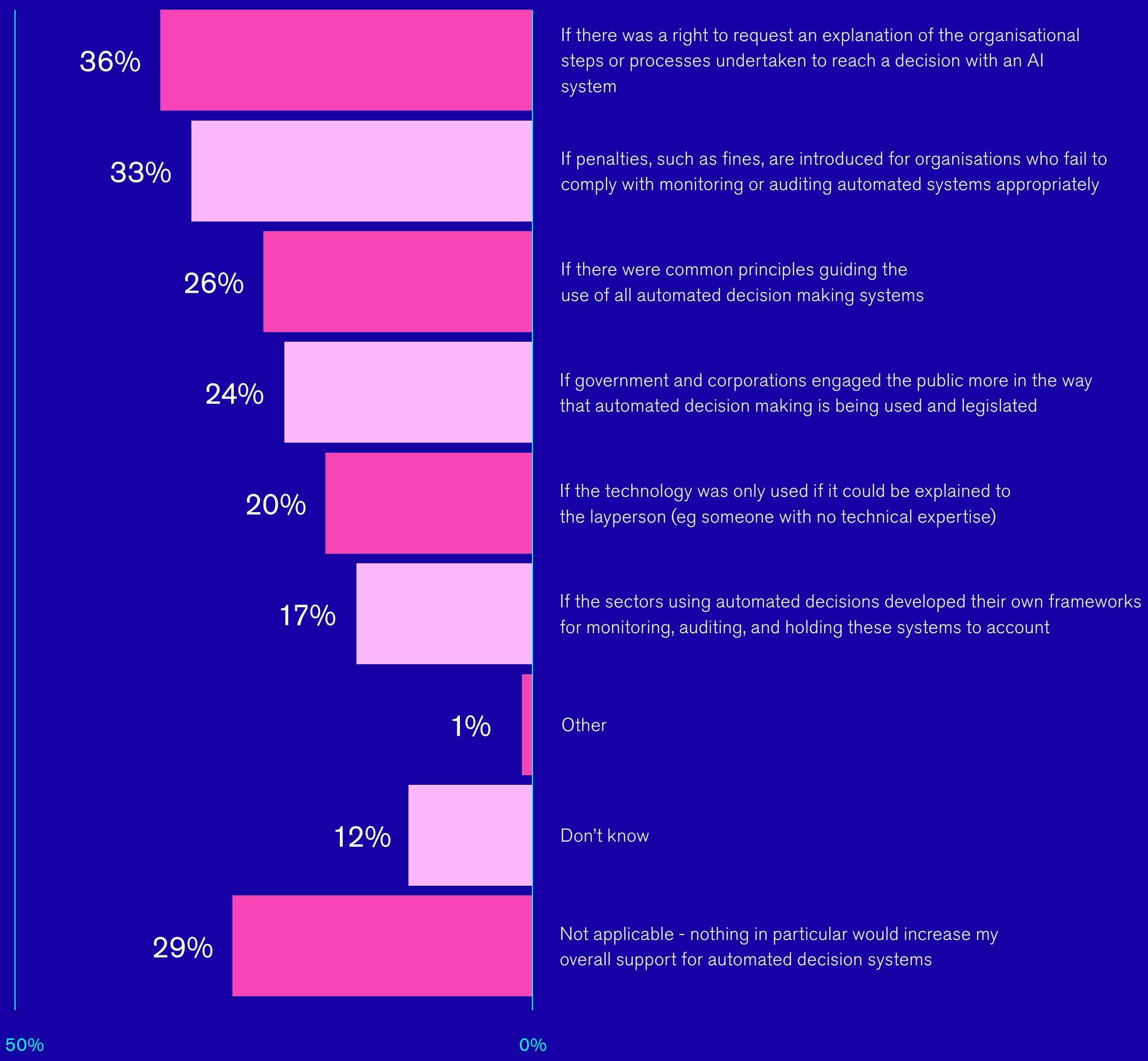
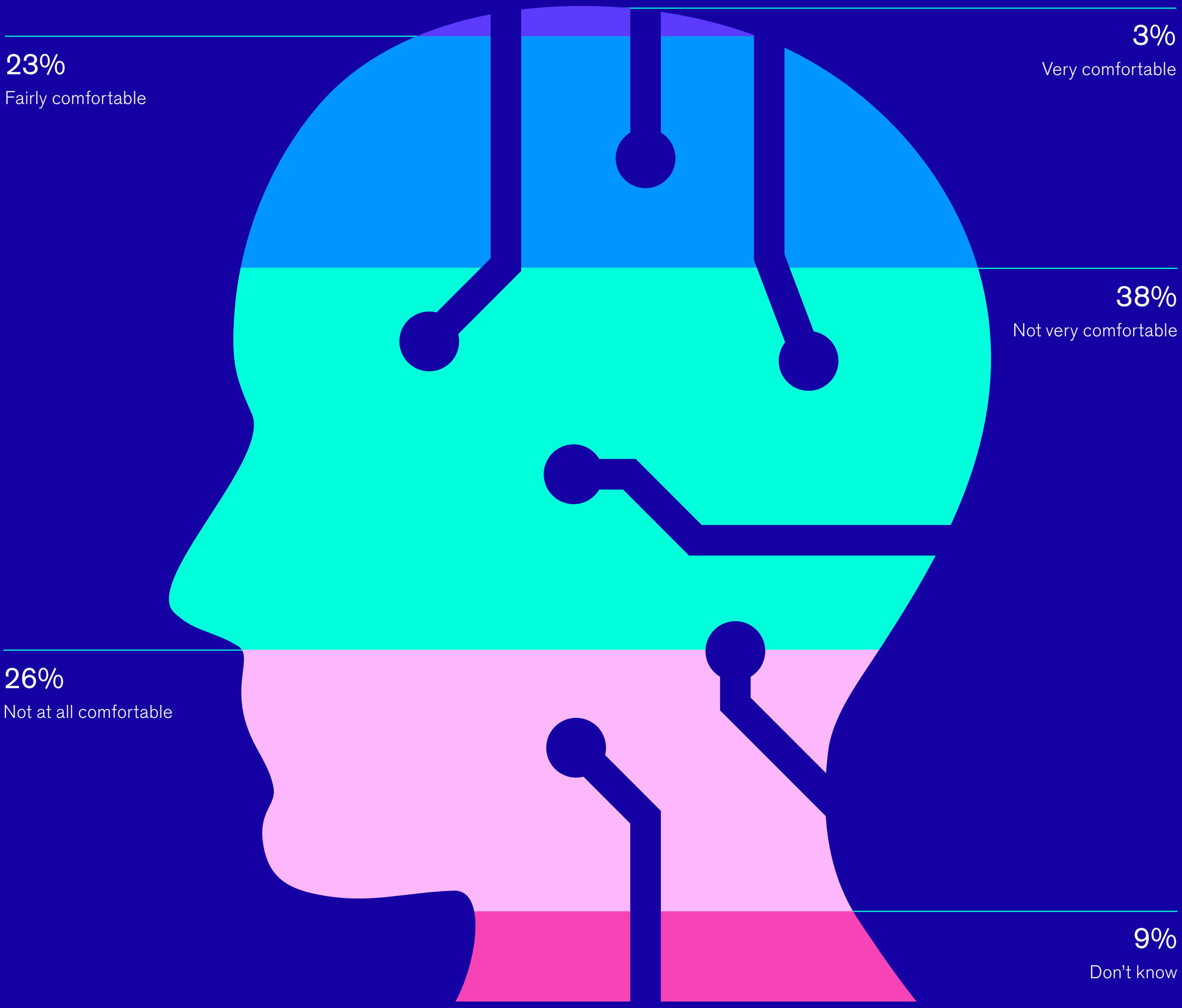


Figure 11

Q: How comfortable, if at all, are you with the following idea?

As the accuracy and consistency of automated systems improve over time, more decisions can be fully automated without human intervention required.



Differences in responses

Younger people (ages 18 – 34) were slightly more likely to be familiar with AI and different uses of automated decision systems. In particular, they were much more aware of its use for decisions about the content and advertisements displayed by search engines and on social media (71 percent of 18 – 24 year olds and 60 percent of 25 – 34 year olds, in contrast with 55 percent of 35 – 44 year olds and 36 percent of 55+ year olds who were familiar). Accordingly, they were also slightly more likely to be supportive of these systems than older people (ages 35 – 55+). For example, 45 percent of 18 – 24 year olds supported the use of AI in making decisions about the content or advertisements displayed by search engines and on social media, compared to only 20 percent of 55+ year olds.

People from more affluent backgrounds were slightly more likely to be familiar with the use of automated decision systems and, correspondingly, slightly more supportive.[70] It is possible that those from more affluent groups believe they are most likely to see the benefits of technological advances, and so are inherently better disposed towards them. This is something to investigate as, if true, it would suggest that societies with greater economic equality could enjoy a competitive advantage in reaping the benefits of AI.

These responses indicate a baseline in the awareness, engagement and support of citizens with regard to automated decision systems. However, we will also be surveying the participants of our citizens' jury specifically in order to evaluate the difference that an informed dialogue can make.

Younger people (ages 18 – 34) were slightly more likely to be familiar with AI. They were also slightly more likely to be supportive of these systems than older people (ages 35 – 55+).

People from more affluent backgrounds were slightly more likely to be familiar with the use of automated decision systems and, correspondingly, slightly more supportive.

Key issues for public deliberation

Based on the survey results and our own research into the growing use of automated decision systems, we propose three key issues that are particularly appropriate for public deliberation and will raise a number of ethical questions including, but not limited to, ownership over data and intellectual property, privacy, agency, accountability and fairness. This is a preliminary set that we expect will evolve during the course of the project.

1. Transparency and explainability

As we've noted, automated decision systems refer to the computer systems that either inform or make a decision about a course of action to pursue about an individual or business. Some have characterised these as systems that limit human judgment,[71] although it's important to recognise that humans usually use the information generated by these systems in order to make a decision. As our colleague Jasmine Leonard explains, this information is typically a prediction about the likelihood of something occurring; for example, the likelihood that a defendant will reoffend, or that an individual will default on a loan.[72] A human will then use the prediction to make a decision about whether or not to grant a defend bail or provide an individual with a credit card. She suggests thinking of automated decision systems as 'prediction engines', which can help us clearly distinguish their role as part of a wider process of decision-making.

In common with the broader approach to understanding decision-making contexts that we set out in the previous chapter, some researchers have stressed the need to clarify the 'constitution' of this wider process (of decision-making); specifically, the "nature of its technical elements, human participation, governing rules, and how they all interact."[73] They argue that in order to understand a lending decision, for example, it should be known that credit scores are generated by software programmes and that human analysts review those numbers as part of a final determination.

However, even when the constitution of the process is mapped out it can be challenging to explain an automated decision to citizens. Some of these computer systems employ machine learning methods which complicate the ability to communicate why a certain prediction was made. Machine learning enables computer systems to "learn directly from examples, data, and experience" rather than following pre-programmed rules.[74] While more advanced methods of machine learning, such as deep learning, are proving to be the most effective at recognising patterns in data, it is currently not possible for us to make sense of what those patterns are. In other words, the complexity of the machine's learning process is an obstacle for humans trying to interrogate its conclusions.

These computer systems that defy explanation are referred to as 'black boxes'.[75] Some experts have called for public bodies to end their use of black box systems. For example, the AI Now Institute recommended in 2017 that core public agencies in 'high-stakes' domains, such as those responsible for criminal justice, healthcare, welfare and education, should no longer use black box systems, especially if they cannot be publicly audited and subject to accountability standards.[76] More recently in the UK, the House of Lords Select Committee on AI expressed that it was unacceptable to deploy any AI system that could have a substantial impact on an individuals' life, unless it can generate "a full and satisfactory explanation" for the decisions it will take.[77] The Committee added that this may mean delaying the deployment of some systems, such as those which are based on deep neural networks, because it is impossible to generate thorough explanations for the decisions that are made.[78]

In our view, there is a difference between 'black box systems' and 'black box processes' (or constitutions). The former refers to opaque computer systems that are beyond scrutiny, whereas the latter refers to opaque organisational processes, such as those that relate to decision-making, that are not made transparent. This distinction encourages reflection on whether a technical explanation is needed for how an algorithm arrived at its prediction, or whether an explanation of the process by which a decision is made would suffice as 'satisfactory'. For instance, it is still possible to audit inputs (such as training data) and outputs (such as the accuracy of the predictions) without knowledge of how the algorithm itself works. Moreover, even if it were possible to provide a technical explanation, this would not indicate to us how a human factors this prediction into the decision they ultimately make or how much weight they give the prediction.

There is also a question of how accessible technical explanations are and whether they are necessary to justify a decision. Our own survey results reveal that technical explanations are less desirable to citizens than explanations of the relevant organisational processes for making a decision and holding the decision-maker accountable. [79] However, we want the citizens of our jury to help clarify what a 'full and satisfactory' explanation means to them and the extent to which it matters to them whether they are able to be informed about the inner workings of an automated decision system.

In some cases, companies designing these systems may be able to provide an explanation for the outcomes, but would prefer not to disclose this information, citing intellectual property rights. Citizens may be able to consider if there are some circumstances in which commercial and competitive interests can supersede individuals' rights (eg when making financial decisions, in recognition that providing a detailed explanation could backfire by helping fraudsters to outwit the system), and when, if at all, such interests should be overruled.

2. Agency and accountability

Even if it is possible to be entirely transparent about an automated decision system and how it arrives at its outcomes, the question remains as to whether information alone enable individual agency. Specifically, what sort of power can citizens exercise over the use of these systems and how they are applied to them? These systems may raise concerns about data privacy, security, and ownership, which has been recognised to a certain extent by new EU General Data Protection Regulation (GDPR). But does this regulation provide the right level of protection to adequately address the level of concern?

Regarding these questions, the citizens will consider how GDPR guidelines should be interpreted and put into practice. For example, GDPR grants individuals the right to not be subjected to a decision based solely on automated processing, including profiling, if it would 'significantly' affect them. But does this go far enough for citizens? As some researchers have noted, this clause may not amount to much in practice because it is rare that decisions are made without any human intervention nor is it clear what constitutes significance.[80] How might citizens distinguish between what is a significant decision and what is not?

Additionally, GDPR indicates the 'right to an explanation' (or specifically, when profiling as part of an automated decision takes place, a data subject has the right to "meaningful information about the logic involved"). But this raises questions about what that would entail in practice, such as whether citizens would be entitled to an explanation of how the system functions technically or of the organisational rationale for a decision.

Although some researchers have challenged whether GDPR is genuinely extending a legally-binding right to an explanation, the Article 29 Working Party guidance states that "controllers must ensure they explain clearly and simply to individuals how the profiling or automated decision-making process works".[81] The Department for Digital, Culture, Media and Sport (DCMS) have also recognised that there may be value in this provision; following a recommendation made to government by an independent review on growing the AI industry in the UK,[82] DCMS have commissioned the ICO and the Alan Turing Institute to produce an ethical framework for explaining automated decision-making.[83] In our view, guidelines on what sort of explanation should be given to the public should also be informed by the public. Engagement with citizens on the nature of an explanation for an automated decision could also address urgent questions about whether it is necessary to ban the use of 'black box' systems by public agencies as some researchers have recently called for.

However, we are also interested in whether an explanation would give individuals sufficient grounds to challenge a decision and/or enable them to hold a person or organisation to account if they believed it was wrong. Should there be mechanisms beyond legislation and regulation to assure citizens that these systems are accountable? What role do companies and civil society play alongside government?

3. Fairness

GDPR may be the most substantive attempt to set out individuals' rights in relation to automated decision systems, but they do not regulate the overall use of these systems. There is no guidance for organisations on whether the use of these systems is appropriate at all in certain contexts, or on the sort of oversight there should be in order to ensure that they meet acceptable standards (eg of accuracy).

According to our survey, citizens have the greatest reservations about the use of automated decision systems in the criminal justice system and in the workplace (60 percent are opposed or strongly opposed to their use in these areas). From follow-up questions about their concerns, it appears that they believe these systems do not have the empathy or compassion required to make decisions that would typically require human judgment, and with it, emotional intelligence. However, not all decisions that can be taken in these areas require emotional engagement, and in some instances individuals may be better served if there was no emotional engagement whatsoever. For example, some argue that hiring and promotion would be less biased if machines were used to either help make, or make, these decisions in the workplace.[84]

Fairness can be subjective, as there are different moral judgments about what fairness is. To better understand what informs these moral judgments, researchers surveyed users on how they perceive and reason about fairness in algorithmic decision-making. They identified eight properties of features that inform judgments about fairness, including reliability, relevance, and privacy.[85] The researchers found that there was a lack of a clear consensus in respondents' judgments about the fairness of using a number of features, but that respondents mainly differed in their objective, rather than subjective, assessments of these properties. We would be keen to explore whether it is possible to reach some sort of consensus or compromise if people are brought together in a dialogue.

In addition to considering how people make trade-offs between different notions of fairness when it comes to the decision-making system and how it is deployed, people may have differing views on what conditions, if any, it is fair to deploy these systems overall. This may mean that people hold different positions on the contexts in which these systems should be used (eg as we know from our survey, people feel differently about these systems depending on the sector or type of use). It may mean that people weigh the benefits (such as potential to improve accuracy, reduce biases, save on costs, and reduce inefficiencies) against the risks (such as the potential to reinforce bias, shift personal responsibility, create ethical distance, and destroy jobs), and determine fairness based on how they are impacted individually or collectively. A question worth asking citizens might be whether it matters who benefits the most from the use of automated decision systems – the organisations making the decisions vs the individuals subject to those decisions.

Based on the survey results and our own research into the growing use of automated decision systems, we propose three key issues that are particularly appropriate for public deliberation and will raise a number of ethical questions including, but not limited to, ownership over data and intellectual property, privacy, agency, accountability and fairness.

Our public dialogue in practice

5

In chapter four, we provided an overview of public attitudes towards automated decision systems, based on survey data, and set out three of the key issues we would like to engage citizens on more in-depth. In this final chapter, we share what this might this look like in practice.

Designing public dialogue on ethical AI

Considering first of all the adoption of automated decision systems by public bodies in the UK, we propose that the comment period included in 'Impact Statement' or 'Impact Assessment' frameworks could enable a deeper level of engagement with citizens than the usual consultation.

For the consideration of significant or controversial systems (eg that are high-impact), this engagement should draw on long-form deliberative processes, such as the use of citizens' juries or citizens' reference panels, which are on the top two rungs of the ladder. The conclusions of the deliberation could be summed up in a statement released by the citizens on either why they accept the use of the automated decision system or under what conditions, if any, they would accept the use of such a system.

The RSA's Forum for Ethical AI is testing whether this would be an effective and meaningful way to engage the public in the ethical use of AI (in this case, for automated decision-making), and would therefore improve the governance of the system and/or increase the extent of public consent or active support. Our preferred methodology is a citizens' jury which is described in Figure 12 below. To provide independent expert scrutiny and advice we have convened an Advisory Panel which will meet a number of times through the duration of the project. The members of the Panel are listed in the Appendix.

While our citizens' jury will deliberate on different uses of automated decision-making systems by various public bodies and private companies, they will not be weighing in on whether these systems should be used. Rather, they will consider what government agencies like the new Centre for Data Ethics and Innovation and the Information Commissioner's Office (ICO) could do to ensure transparency, fairness, and accountability of these systems from the public's perspective.

If this proves successful, there could be stronger support for public agencies to carry out long-form deliberative processes when introducing new AI systems or other technology of significance, however that is defined by the agency (eg as part of an Algorithmic Impact Assessment). In our view, this would include the introduction of some new automated decision systems; for example, in the future there may be more systems developed for use in the criminal justice system or for welfare. A number of organisations, such as Involve and the Ada Lovelace Institute have signalled an interest and commitment to deliberation on ethical AI, and appear well-placed to work with public bodies to embed and progress long-form deliberative processes for this purpose.

The insights generated from these deliberations would be publicly available and would ideally be used to influence a wider range of stakeholders, including tech entrepreneurs and business leaders, investors, regulators, researchers, and campaigners.

Next steps

The citizens' jury will reach their conclusions in June 2018. These conclusions will then be tested during two workshops with citizens who may be disproportionately impacted by the use of these systems. There will be a final event in October 2018, and the programme will culminate with a report.

If you would like to learn more about the project then please visit the project pages at www.thersa.org/action-and-research/rsa-projects/economy-enterprise-manufacturing-folder/forum-for-ethical-AI

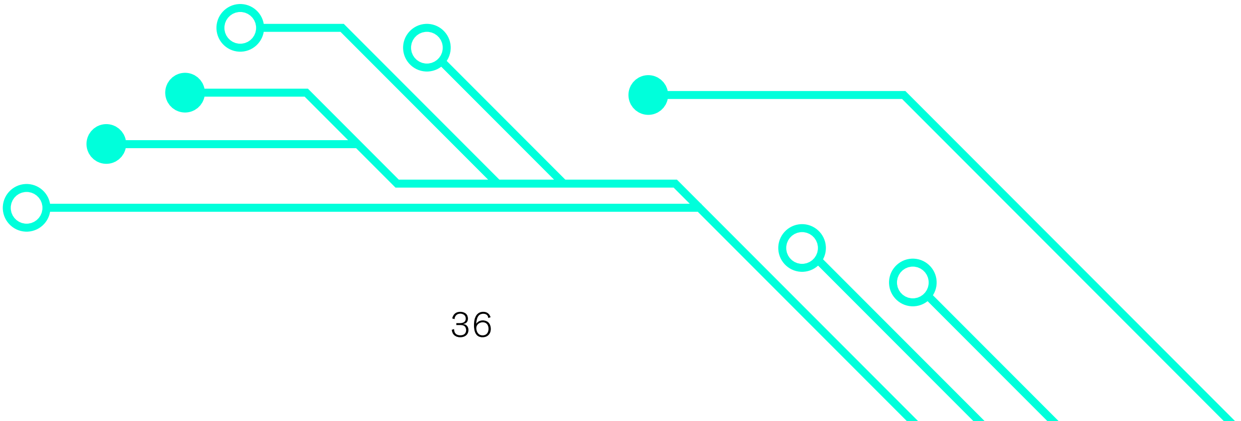
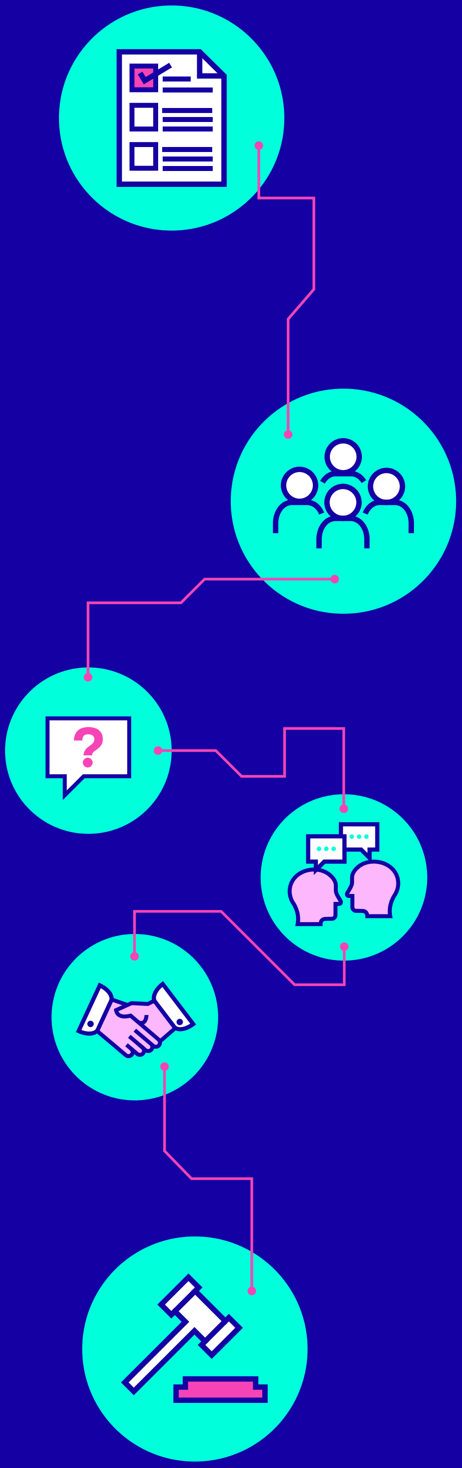


Figure 12
The RSA Forum for Ethical AI's Citizens' Jury – Our Journey



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 – 30 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. [86]

3. Deliberating as a jury

a. 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.

b. 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.

c. 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

Appendix

The Independent Advisory Panel

Dr. Beth Singler
Anthropologist, Faraday Institute for Science and Religion, University of Cambridge

Catherine Miller
Policy Director, Doteveryone

Dr. David Edmonds
Philosopher and documentary maker, BBC World Service

Professor Ian Walden
Professor of Information and Communications Law, Queen Mary University, and Solicitor at Baker McKenzie

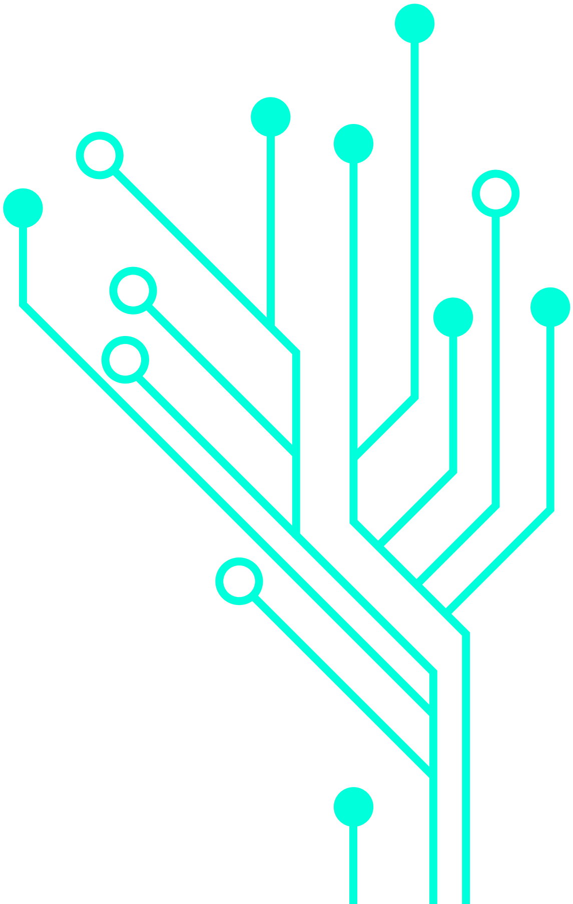
Professor Maja Pantic
Professor of Affective and Behavioural Computing, Imperial College London

Paul Mason
Director of Emerging and Enabling Technologies, Innovate UK

Dr. Rumman Chowdhury
Head of Responsible AI, Accenture

Simon Burrall (Chair)
Programme Director, Sciencewise

Wendy Tan White MBE
Partner, BGF Ventures



Appendix

RSA/YouGov 2018 survey on AI and automated decision systems – Questionnaire

1. **For the following question, by "artificial intelligence (AI)", we mean use of machines that behave intelligently. Before taking this survey, which, if any, of the following uses of AI were you aware of? (Please select all that apply).**

- Digital assistants (eg Siri, Alexa etc.)
- Self-driving cars
- Automated decision systems
- Online 'chatbots'
- Detecting fraudulent transactions (eg when shopping online, making insurance claims etc.)
- Optimising energy usage (eg by helping to reduce electricity usage)
- Discovering new medicines
- Identifying people in photos or videos
- None of these

2. **Artificial intelligence (AI) refers to the use of machines that behave intelligently. AI is increasingly being used to automate human decision-making by analysing data and generating predictions. Currently it is rare for decision-making to be fully automated; most automated decision systems are used to inform human decisions. However, AI has the potential to increase the scale of automated decision-making and further reduce the need for human judgement.**

   Before taking this survey, how familiar, if at all, were you with the idea of automated decision systems being used to aid each of the following decisions? (Please select one option on each row)

- Decisions in the criminal justice system (eg whether to grant a defendant bail or recommend rehabilitation)
- Decisions in the workplace (eg whom to hire and promote)
- Decisions about the content or advertisements displayed by search engines and on social media (eg what you see in your Facebook newsfeed)
- Decisions about immigration (eg whether to permit someone entry into a country)
- Decisions about access to financial services (eg whether to provide someone with a loan or insurance)
- Decisions about healthcare (eg what treatments to prescribe)
- Decisions about claims for social support (eg whether to grant unemployment, disability or housing benefits)

3. **To what extent, if at all, would you support or oppose the use of automated decision systems to aid each of the following decisions? (Please select one option on each row)**

- Decisions in the criminal justice system (eg whether to grant a defendant bail or recommend rehabilitation)
- Decisions in the workplace (eg whom to hire and promote)
- Decisions about the content or advertisements displayed by search engines and on social media (eg what you see in your Facebook newsfeed)
- Decisions about immigration (eg whether to permit someone entry into a country)
- Decisions about access to financial services (eg whether to provide someone with a loan or insurance)
- Decisions about healthcare (eg what treatments to prescribe)
- Decisions about claims for social support (eg whether to grant unemployment, disability or housing benefits)

4. **As a reminder, artificial intelligence (AI) refers to the use of machines that behave intelligently. AI is increasingly being used to automate human decision-making by analysing data and generating predictions. Currently it is rare for decision-making to be fully automated; most automated decision systems are used to inform human decisions. However, AI has the potential to increase the scale of automated decision-making and further reduce the need for human judgement.**

   Which TWO, if any, of the following potential problems of using automated decision systems would you be MOST concerned about? (Please select up to two options)

- AI does not have the empathy or compassion required to make important decisions that affect individuals and communities
- Automated decisions could reinforce or deepen existing biases and inequality in systems
- There is a lack of adequate oversight or government regulation of automated decisions to protect people if a decision made is unfair
- Jobs could be lost to automated decision-making
- Automated decision-making reduces peoples responsibility and accountability for the decisions they implement
- It's not always clear how AI reaches a decision
- Other
- Don't know
- Not applicable - I'm not particularly concerned about any potential problems of automated decision systems

5 . Which TWO, if any, of the following potential benefits of using automated decision systems would you MOST look forward to? (Please select up to two options)

- AI can make better decisions than humans because emotions never cloud its judgment
- Automated decisions could reduce existing biases and inequality
- Automated decisions could increase efficiency and therefore help governments and companies to save money
- Workers might be less stressed and more productive because they have more support to make important decisions
- Increased use of AI would encourage organisations to examine their processes and make new commitments to greater transparency and accountability
- It could improve the accuracy and consistency of decisions
- Other
- Don't know
- Not applicable - I'm not particularly looking forward to any potential benefits of automated decision systems

6. Generally speaking, which, if any, of the following would increase your overall support for automated decision systems? (Please select all that apply. If nothing in particular would increase your support, please select the "Not applicable" option)

- If there was a right to request an explanation of the organisational steps or processes undertaken to reach a decision using an AI system
- If the technology was only used if it could be explained to the layperson (i.e. someone with no technical expertise)
- If government and corporations engaged the public more in the way that automated decision making is being used and legislated
- If there were common principles guiding the use of all automated decision making systems
- If the sectors using automated decisions developed their own frameworks for monitoring, auditing, and holding these systems to account
- If penalties, such as fines, are introduced for organisations who fail to comply with monitoring or auditing automated systems appropriately
- Other
- Don't know
- Not applicable - nothing in particular would increase my overall support for automated decision systems

7. How comfortable, if at all, are you with the following idea? As the accuracy and consistency of automated systems improve over time, more decisions can be fully automated without human intervention required.

- Very comfortable
- Fairly comfortable
- Not very comfortable
- Not at all comfortable
- Don't know

References

1. For a few examples, see: FAT/ML, AI Now Institute, USACM, Future of Life Institute, Future of Humanity Institute, and House of Lords Select Committee on AI

2. 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

3. 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

4. RSA/YouGov Survey 2018 on AI and automated decision systems

5. Blackwell, J., Fowler, B., and Fox, R. (2018) Audit of Public Engagement 15: The 2018 Report. London: Hansard Society

6. Ito, J. (2016) 'Society in the Loop Artificial Intelligence'. Joi Ito, 7 May; Rahwan, I. (2017) 'Society-in-the-Loop: Programming the Algorithmic Social Contract'. Ethics of Information Technology

7. Rahwan (2017) op cit.

8. (2017) Artificial Intelligence Index: 2017 Annual Report

9. See Artificial Intelligence Index: 2017 Annual Report

10. Delaney, J.K. (2018) 'France, China, and the EU all have an AI strategy. Shouldn't the US?' Wired, 20 May

11. O'neil, C. (2016) Weapons of Math Destruction: How big data increases inequality and threatens democracy. New York: Crown; Eubanks, V. (2017) Automating Inequality: How high-tech tools profile, police, and punish the poor. New York: St. Martin's Press

12. Sample, I. (2017) 'Computer says no: why making AIs fair, accountable and transparent is crucial'. The Guardian, 5 November

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

14. Brundage, M. et al. (2018) *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*

15. See Brundage, M. et al. (2018)

16. FAT/ML (2016) 'Principles for Accountable Algorithms and a Social Impact Statement for Algorithms'
-
17. USACM (2017) 'Statement for Algorithmic Transparency and Accountability'

18. Reisman, D., Schultz, J., Crawford, K., and Whittaker, M. (2018) *Algorithmic Impact Assessments: A practical framework for public agency accountability.* New York: AI Now Institute

19. See Partnership on AI: https://www.partnershiponai.org/tenets

20. Harris, J.G. and Davenport, T.H. (2005) Automated Decision Making-making Comes of Age. Wellesley: Accenture Institute for High Performance Business

21. Eubanks, V. (2017) op cit.

22. Ibid.

23. Eubanks, V. (2018) 'The Digital Poorhouse'. Harper's Magazine

24. See note 3 above

25. The boundaries to public diaglouе are set by human rights and rule of law

26. Dietz, T. (2013) 'Bringing values and deliberation to science communication', PNAS, 110(3), pp. 14080 – 14087

27. Ibid.

28. Ibid.

29. Ibid.

30. Brundage et al. (2018) op cit.

31. Mistreanu, S. (2018) 'Life Inside China's Social Credit Laboratory'. Foreign Policy, 3 April
32. Martinho-Truswell, E. (2018) 'How AI Could Help the Public Sector.' Harvard Business Review, 29 January

33. See note 2 above

34. Craig, C. et al. (2017) op cit.

35. Rahwan (2017) op cit.

36. Patel, R. and Greenham, T. (2017) 'The Governance Challenge', in *In our interests: building an economy for all*. London: The Co-operative Party

37. Balaram, B. (2016) *Fair Share: reclaiming power in the sharing economy*. London: RSA

38. Patel, R., Gibbon, K. and Greenham, T. (2018) *Building a public culture of economics*. London: RSA

39. Conway, R., Masters, J., and Thorold, J. (2017) *From design thinking to systems change: how to invest in innovation for social impact*. London: RSA

40. Note that for simplicity we are not distinguishing between those who take a decision and those who implement a decision

41. For the purposes of this project, we are concerned with decisions that are made by individuals on behalf of institutions rather than on their own account in their personal lives

42. We are not aware of any proposals as yet to allow boards of directors, or equivalent organisational governing bodies, to entirely comprise of AI

43. FAT/ML (2016) op cit.

44. Reisman et al. (2018) op cit.

45. Karlin, M. (2018) 'A Canadian Algorithmic Impact Assessment'. Medium, 18 March

46. Signatories to the Convention are committing to ensuring public participation in decision-making on environmental matters

47. The Sciencewise programme was run by the Department of Energy and Industrial Strategy and helped policy-makers to carry out public dialogue to inform their decision-making on science and technology issues

48. For example, see 'Health expectations: an international journal of public participation in health care and health policy' published by Wiley

[49] For a good overview, see Hajer and Wagenaar (eds) (2003) 'Deliberative Policy Analysis: Understanding Governance in the Network Society'. Cambridge: University Press

[50] Escobar, O. (2011) *Public Dialogue and Deliberation: A communication perspective for public engagement practitioners.* Edinburgh: Edinburgh Beltane

[51] Bussu, S., Davis, H., and Pollard, A. (2014) *The best of Sciencewise reflections on public dialogue.* London: Sciencewise

[52] Ibid.

[53] Arnstein, S. (1969) 'A Ladder of Citizen Participation'. Journal of the American Planning Association, 35 (4), pp. 216 - 224

[54] International Association of Public Participation "Spectrum of Public Participation"

[55] Chwalisz, C. (2017) *The People's Verdict: Adding Informed Citizen Voices to Public Decision-making.* London: Policy Network

[56] Prikken, I. and Burall, S. (2012) *Doing Public Dialogue.* London: Involve

[57] See MASS LBP: https://www.masslbp.com/work

[58] Worthen, M. (2017) 'Where in the world can we find hope?' The New York Times, 18 February

[59] Ibid.

[60] Blackwell et al. (2018) op cit.

[61] Bussu et al. (2014) op cit.

[62] Mohr, A., Raman, S., and Gibbs, B. (2013) *Which Publics? When?: Exploring the potential of involving different publics in dialogue around science and technology.* London: Sciencewise

[63] Andersson, E., McLean, S., Parlak, M., and Melvin, G. (2013) *From Fairy Tale to Reality: Dispelling the myths around citizen engagement.* London: Involve

[64] Ibid.

[65] Ibid.

[66] House of Lords Select Committee on Artificial Intelligence (2018) AI in the UK: ready, willing and able? Report of Session 2017-19 - published 16 April 2017 - HL Paper 100

[67] All figures, unless otherwise stated, are from YouGov Plc. Total sample size was 2,074 adults. Fieldwork was undertaken between 16th and 17th April 2018. The survey was carried out online. The figures have been weighted and are representative of all UK adults (aged 18+)

[68] Only 46 percent of people are UK adults are aware of chatbots. Older people over the age of 55 were the least likely to be aware of chatbots; 29 percent of 55+ contrasted with 72 percent of young people between the ages of 18 - 24

[69] 40 percent of people were either fairly or very familiar with the use of AI in financial services, and 49 percent were either fairly or very familiar with the curation of content and advertisements by social media companies

[70] For example, 45 percent of UK adults with social grade ABC1 were familiar with AI being used in the financial services, compared to 33 percent of UK adults with social grade C2DE. 31 percent with ABC1 social grade support AI use in the financial services, compared to 22 percent with C2DE social grade

[71] Rieke, A., Bogen, M., and Robinson, D.G. (2018) *Public Scrutiny of Automated Decisions: Early lessons and emerging methods.* Washington/London: Upturn and Omidyar Network

[72] Leonard, J. (2018) 'Computer says no: Part 1 – Algorithmic bias'. RSA Blog, 14 March

[73] Rieke et al. (2018) op cit.

[74] Royal Society (2017) op cit.

[75] Ibid.

[76] Campolo, A., Sanfilippo, M., Whittaker, M., and Crawford, K. (2017) *AI Now 2017 Report.* New York: AI Now Institute

[77] House of Lords Select Committee on Artificial Intelligence (2018) op cit.

[78] Ibid.

[79] Wachter, S., Mittelstadt, B., and Floridi, L. (2017) 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation'. International Data Privacy Law

[80] Wachter, S., Mittelstadt, B. and Floridi, L. (2017) 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.' *International Data Privacy Law*

[81] Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, wp251rev.01. The Working Party was set up under Article 29 of Directive 95/46/EC. It is an independent European advisory body on data protection and privacy

[82] Hall, W. and Pesenti, J. (2018) Growing the artificial intelligence industry in the UK. London: Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy

[83] See Alan Turing Institute: https://www.turing.ac.uk/data-ethics

[84] O'Connor, S. (2016) 'When your boss is an algorithm'. Financial Times, 8 September

[85] Grgic-Hlaca, N., Redmiles, E.M., Gummadi, K.P. and Weller, A. (2018) 'Human perceptions in fairness in algorithmic decision-making: A case study of criminal risk prediction'

[86] As observed by Claudia Chwalisz, these groups are usually between 24 – 48 people in size and a 'community' can be national, regional or local depending on the scope of the question

RSA

21st century enlightenment

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

Registered as a charity in
England and Wales
no. 212424

www.thersa.org

ISBN 978-1-911532-20-0