Explainable AI: the basics

POLICY BRIEFING

THE ROYAL SOCIETY

Explainable AI: the basics
Policy briefing
Issued: November 2019 DES6051
ISBN: 978-1-78252-433-5

This report can be viewed online at:
royalsociety.org/ai-interpretability

Contents

Summary

Recent years have seen significant advances in the capabilities of Artificial Intelligence (AI) technologies. Many people now interact with AI-enabled systems on a daily basis: in image recognition systems, such as those used to tag photos on social media; in voice recognition systems, such as those used by virtual personal assistants; and in recommender systems, such as those used by online retailers.

As AI technologies become embedded in decision-making processes, there has been discussion in research and policy communities about the extent to which individuals developing AI, or subject to an AI-enabled decision, are able to understand how the resulting decision-making system works.

Some of today's AI tools are able to produce highly-accurate results, but are also highly complex. These so-called 'black box' models can be too complicated for even expert users to fully understand. As these systems are deployed at scale, researchers and policymakers are questioning whether accuracy at a specific task outweighs other criteria that are important in decision-making systems. Policy debates across the world increasingly see calls for some form of AI explainability, as part of efforts to embed ethical principles into the design and deployment of AI-enabled systems. This briefing therefore sets out to summarise some of the issues and considerations when developing explainable AI methods.

There are many reasons why some form of interpretability in AI systems might be desirable or necessary. These include: giving users confidence that an AI system works well; safeguarding against bias; adhering to regulatory standards or policy requirements; helping developers understand why a system works a certain way, assess its vulnerabilities, or verify its outputs; or meeting society's expectations about how individuals are afforded agency in a decision-making process.

Different AI methods are affected by concerns about explainability in different ways. Just as a range of AI methods exists, so too does a range of approaches to explainability. These approaches serve different functions, which may be more or less helpful, depending on the application at hand. For some applications, it may be possible to use a system which is interpretable by design, without sacrificing other qualities, such as accuracy.

There are also pitfalls associated with these different methods, and those using AI systems need to consider whether the explanations they provide are reliable, whether there is a risk that explanations might deceive their users, or whether they might contribute to gaming of the system or opportunities to exploit its vulnerabilities.

Different contexts give rise to different explainability needs, and system design often needs to balance competing demands – to optimise the accuracy of a system or ensure user privacy, for example. There are examples of AI systems that can be deployed without giving rise to concerns about explainability, generally in areas where there are no significant consequences from unacceptable results or the system is well-validated. In other cases, an explanation about how an AI system works is necessary but may not be sufficient to give users confidence or support effective mechanisms for accountability.

In many human decision-making systems, complex processes have developed over time to provide safeguards, audit functions, or other forms of accountability. Transparency and explainability of AI methods may therefore be only the first step in creating trustworthy systems and, in some circumstances, creating explainable systems may require both these technical approaches and other measures, such as assurance of certain properties. Those designing and implementing AI therefore need to consider how its use fits in the wider socio-technical context of its deployment.

AI and the 'black box'

AI's explainability issue

AI is an umbrella term. It refers to a suite of technologies in which computer systems are programmed to exhibit complex behaviour – behaviour that would typically require intelligence in humans or animals – when acting in challenging environments.

Recent years have seen significant advances in the capabilities of AI technologies, as a result of technical developments in the field, notably in machine learning[1]; increased availability of data; and increased computing power. As a result of these advances, systems which only a few years ago struggled to achieve accurate results can now outperform humans at some specific tasks[2].

Many people now interact with AI-enabled systems on a daily basis: in image recognition systems, such as those used to tag photos on social media; in voice recognition systems, such as those used by virtual personal assistants; and in recommender systems, such as those used by online retailers.

Further applications of machine learning are already in development in a diverse range of fields. In healthcare, machine learning is creating systems that can help doctors give more accurate or effective diagnoses for certain conditions. In transport, it is supporting the development of autonomous vehicles, and helping to make existing transport networks more efficient. For public services it has the potential to target support more effectively to those in need, or to tailor services to users[3]. At the same time, AI technologies are being deployed in highly-sensitive policy areas – facial recognition in policing or predicting recidivism in the criminal justice system, for example – and areas where complex social and political forces are at work. AI technologies are therefore being embedded in a range of decision-making processes.

There has, for some time, been growing discussion in research and policy communities about the extent to which individuals developing AI, or subject to an AI-enabled decision, are able to understand how AI works, and why a particular decision was reached[4]. These discussions were brought into sharp relief following adoption of the European General Data Protection Regulation, which prompted debate about whether or not individuals had a 'right to an explanation'.

This briefing sets out to summarise the issues and questions that arise when developers and policymakers set out to create explainable AI systems.

1. Machine learning is the technology that allows computer systems to learn directly from data.
2. It should be noted, however, that these benchmark tasks tend to be constrained in nature. In 2015, for example, researchers created a system that surpassed human capabilities in a narrow range of vision-related tasks, which focused on recognising individual handwritten digits. See: Markoff J. (2015) A learning advance in artificial intelligence rivals human abilities. New York Times. 10 December 2015.
3. Royal Society (2017) Machine learning: the power and promise of computers that learn by example, available at www.royalsociety.org/machine-learning
4. Pasquale, F. (2015) The Black Box Society: The Secret Algorithms that Control Money and Information, Harvard University Press, Cambridge, Massachusetts

BOX 1

AI, machine learning, and statistics: connections between these fields

The label 'artificial intelligence' describes a suite of technologies that seek to perform tasks usually associated with human or animal intelligence. John McCarthy, who coined the term in 1955, defined it as "the science and engineering of making intelligent machines"; in the time since, many different definitions have been proposed.

Machine learning is a branch of AI that enables computer systems to perform specific tasks intelligently. Traditional approaches to programming rely on hardcoded rules, which set out how to solve a problem, step-by-step. In contrast, machine learning systems are set a task, and given a large amount of data to use as examples (and non-examples) of how this task can be achieved, or from which to detect patterns. The system then learns how best to achieve the desired output. There are three key branches of machine learning:

- In supervised machine learning, a system is trained with data that has been labelled. The labels categorise each data point into one or more groups, such as 'apples' or 'oranges'. The system learns how this data – known as training data – is structured, and uses this to predict the categories of new – or 'test' – data.
- Unsupervised learning is learning without labels. It aims to detect the characteristics that make data points more or less similar to each other, for example by creating clusters and assigning data to these clusters.
- Reinforcement learning focuses on learning from experience. In a typical reinforcement learning setting, an agent interacts with its environment, and is given a reward function that it tries to optimise, for example the system might be rewarded for winning a game. The goal of the agent is to learn the consequences of its decisions, such as which moves were important in winning a game, and to use this learning to find strategies that maximise its rewards.

While not approaching the human-level general intelligence which is often associated with the term AI, the ability to learn from data increases the number and complexity of functions that machine learning systems can undertake. Rapid advances in machine learning are today supporting a wide range of applications, many of which people encounter on a daily basis, leading to current discussion and debate about the impact of AI on society.

Many of the ideas which frame today's machine learning systems are not new; the field's statistical underpinnings date back many decades, and researchers have been creating machine learning algorithms with various levels of sophistication since the 1950s.

Machine learning involves computers processing a large amount of data to predict outcomes. Statistical approaches can inform how machine learning systems deal with probabilities or uncertainty in decision-making.

However, statistics also includes areas of study which are not concerned with creating algorithms that can learn from data to make predictions or decisions. While many core concepts in machine learning have their roots in data science and statistics, some of its advanced analytical capabilities do not naturally overlap with these disciplines.

Other approaches to AI use symbolic, rather than statistical, approaches. These approaches use logic and inference to create representations of a challenge and to work through to a solution.

This document employs the umbrella term 'AI', whilst recognising that this encompasses a wide range of research fields, and much of the recent interest in AI has been driven by advances in machine learning.

The 'black box' in policy and research debates

Some of today's AI tools are able to produce highly-accurate results, but are also highly complex if not outright opaque, rendering their workings difficult to interpret. These so-called 'black box' models can be too complicated for even expert users to fully understand[5]. As these systems are deployed at scale, researchers and policymakers are questioning whether accuracy at a specific task outweighs other criteria that are important in decision-making[6].

Policy debates across the world increasingly feature calls for some form of AI explainability, as part of efforts to embed ethical principles into the design and deployment of AI-enabled systems[7]. In the UK, for example, such calls have come from the House of Lords AI Committee, which argued that "the development of intelligible AI systems is a fundamental necessity if AI is to become an integral and trusted tool in our society"[8]. The EU's High-Level Group on AI has called for further work to define pathways to achieving explainability[9]; and in the US, the Defence Advanced Research Projects Agency supports a major research programme seeking to create more explainable AI[10]. As AI methods are applied to address challenges in a wide range of complex policy areas, as professionals increasingly work alongside AI-enabled decision-making tools, for example in medicine, and as citizens more frequently encounter AI systems in domains where decisions have a significant impact, these debates will become more pressing.

AI research, meanwhile, continues to advance at pace. Explainable AI is a vibrant field of research, with many different methods emerging, and different approaches to AI are affected by these concerns in different ways.

Terminology

Across these research, public, and policy debates, a range of terms is used to describe some desired characteristics of an AI. These include:

- interpretable, implying some sense of understanding how the technology works;
- explainable, implying that a wider range of users can understand why or how a conclusion was reached;
- transparent, implying some level of accessibility to the data or algorithm;
- justifiable, implying there is an understanding of the case in support of a particular outcome; or
- contestable, implying users have the information they need to argue against a decision or classification.

5. As Rudin (2019) notes, this term also refers to proprietary models to which users are denied access. Rudin, C. (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, Nature Machine Intelligence, 1, 206-215
6. Doshi-Velez F. and Kim B. (2018) Considerations for Evaluation and Generalization in Interpretable Machine Learning. In: Escalante H. et al. (eds) Explainable and Interpretable Models in Computer Vision and Machine Learning. The Springer Series on Challenges in Machine Learning. Springer
7. See Annex 1 for a sketch of the policy landscape
8. House of Lords (2018) AI in the UK: ready, willing and able? Report of Session 2017 – 19. HL Paper 100.
9. EU High Level Group on AI (2019) Ethics guidelines for trustworthy AI. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai [accessed 2 August 2018]
10. DARPA Explainable Artificial Intelligence (XAI program), available at: https://www.darpa.mil/program/explainable-artificial-intelligence [accessed 2 August 2018]

While use of these terms is inconsistent[11], each tries to convey some sense of a system that can be explained or presented in terms that are understandable to a particular audience for a particular purpose.

Individuals might seek explanations for different reasons. Having an understanding of how a system works might be necessary to examine and learn about how well a model is functioning; to investigate the reasons for a particular outcome; or to manage social interactions[12]. The nature and type of explanation, transparency or justification that they require varies in different contexts.

This briefing maps the range of reasons why different forms of explainability might be desirable for different individuals or groups, and the challenges that can arise in bringing this into being.

The case for explainable AI: how and why interpretability matters

There is a range of reasons why some form of interpretability in AI systems might be desirable. These include:

Giving users confidence in the system: User trust and confidence in an AI system are frequently cited as reasons for pursuing explainable AI. People seek explanations for a variety of purposes: to support learning, to manage social interactions, to persuade, and to assign responsibility, amongst others[13]. However, the relationship between the trustworthiness of a system and its explainability is not a straightforward one, and the use of explainable AI to garner trust may need to be treated with caution: for example, plausible-seeming explanations could be used to mislead users about the effectiveness of a system[14].

Safeguarding against bias: In order to check or confirm that an AI system is not using data in ways that result in bias or discriminatory outcomes, some level of transparency is necessary.

Meeting regulatory standards or policy requirements: Transparency or explainability can be important in enforcing legal rights surrounding a system, in proving that a product or service meets regulatory standards, and in helping navigate questions about liability. A range of policy instruments already exist that seek to promote or enforce some form of explainability in the use of data and AI (outlined in Annex 1).

11. See, for example: Lipton, Z. (2016) The Mythos of Model Interpretability. ICML 2016 Workshop on Human Interpretability in Machine Learning (WHI 2016)
12. Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence. 267. 10.1016/j.artint.2018.07.007.
13. Discussed in more detail in Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence. 267. 10.1016/j.artint.2018.07.007.
14. See later discussion, and Weller, A. (2017). Challenges for Transparency. Workshop on Human Interpretability (ICML 2017).

Improving system design: Interpretability can allow developers to interrogate why a system has behaved in a certain way, and develop improvements. In self-driving cars, for example, it is important to understand why and how a system has malfunctioned, even if the error is only minor. In healthcare, interpretability can help explain seemingly anomalous results[15].

Engineers design interpretable systems in order to track system malfunctions. The types of explanations created to fulfil this function could take different forms to those required by user groups – though both might include investigating both the training data and the learning algorithm.

BOX 2

Bias in AI systems

Real-world data is messy: it contains missing entries, it can be skewed or subject to sampling errors, and it is often collected for purposes other than the analysis at hand.

Sampling errors or other issues in data collection can influence how well the resulting machine learning system works for different users. There have been a number of high profile instances of image recognition systems failing to work accurately for users from minority ethnic groups, for example.

The models created by a machine learning system can also generate issues of fairness or bias, even if trained on accurate data, and users need to be aware of the limitations of the systems they use. In recruitment, for example, systems that make predictions about the outcomes of job offers or training can be influenced by biases arising from social structures that are embedded in data at the point of collection. The resulting models can then reinforce these social biases, unless corrective actions are taken.

Concepts like fairness can have different meanings to different communities, and there can be trade-offs between these different interpretations. Questions about how to build 'fair' algorithms are the subject of increasing interest in technical communities and ideas about how to create technical 'fixes' to tackle these issues are evolving, but fairness remains a challenging issue. Fairness typically involves enforcing equality of some measure across individuals and/or groups, but many different notions of fairness are possible – these different notions can often be incompatible, requiring more discussions to negotiate inevitable trade-offs[16].

15. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N. (2015) Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721-1730

16. Kleinberg, J., Mullainathan, S. and Raghavan, M. (2016) Inherent trade offs in the fair determination of risk scores, Proceedings of the ACM International Conference on Measurement and Modelling of Computer Systems, p40; and Kleinbery, J., Ludwig, J., Mullainathan, S. and Rambachan, A. (2018) Algorithmic fairness, Advances in Big Data Research in Economics, AEA Papers and Proceedings 2018, 108, 22-27

Assessing risk, robustness, and vulnerability: Understanding how a system works can be important in assessing risk[17]. This can be particularly important if a system is deployed in a new environment, where the user cannot be sure of its effectiveness. Interpretability can also help developers understand how a system might be vulnerable to so-called adversarial attacks, in which actors seeking to disrupt a system identify a small number of carefully-chosen data points to alter in order to prompt an inaccurate output from the system. This can be especially important in safety-critical tasks[18].

Understanding and verifying the outputs from a system: Interpretability can be desirable in verifying the outputs from a system, by tracing how modelling choices, combined with the data used, affect the results. In some applications, this can be useful in helping developers understand cause-and-effect relationships in their analysis[19].

Autonomy, agency, and meeting social values: For some, transparency is a core social or constitutional value, and a core part of systems of accountability for powerful actors. This relates to dignity concerns about how an individual is treated in a decision-making process. An explanation can play a role in supporting individual autonomy, allowing an individual to contest a decision and helping provide a sense of agency in how they are treated[20].

17. In financial applications, for example, investors might be unwilling to deploy a system without understanding the risks involved or how it might fail, which requires an element of interpretability.

18. See, for example: S. Russell, D. Dewey, and M. Tegmark (2015) "Research priorities for robust and beneficial artificial intelligence," AI Magazine, vol. 36, no. 4, pp. 105–114.

19. For example, AI has a wide range of applications in scientific research. In some contexts, accuracy alone might be sufficient to make a system useful. This is discussed further in the Royal Society and The Alan Turing Institute's discussion paper on AI in science, available at: https://royalsociety.org/topics-policy/data-and-ai/artificial-intelligence/

20. Discussed in Burrell, J. (2016) How the Machine 'Thinks:' understanding opacity in machine learning algorithms, Big Data & Society; and Ananny, M. & K. Crawford (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society. doi: https://doi.org/10.1177/1461444816676645

Explainable AI: the current state of play

There are different approaches to AI, which present different types of explainability challenge.

Symbolic approaches to AI use techniques based on logic and inference. These approaches seek to create human-like representations of problems and the use of logic to tackle them; expert systems, which work from datasets codifying human knowledge and practice to automate decision-making, are one example of such an approach. While symbolic AI in some senses lends itself to interpretation – it being possible to follow the steps or logic that led to an outcome – these approaches still encounter issues with explainability, with some level of abstraction often being required to make sense of large-scale reasoning.

Much of the recent excitement about advances in AI has come as a result of advances in statistical techniques. These approaches – including machine learning – often leverage vast amounts of data and complex algorithms to identify patterns and make predictions. This complexity, coupled with the statistical nature of the relationships between inputs that the system constructs, renders them difficult to understand, even for expert users, including the system developers.

Reflecting the diversity of AI methods that fall within these two categories, there are many different explainable AI techniques in development. These fall – broadly – into two groups:

- The development of AI methods that are inherently interpretable, meaning the complexity or design of the system is restricted in order to allow a human user to understand how it works.
- The use of a second approach that examines how the first 'black box' system works, to provide useful information. This includes, for example, methods that re-run the initial model with some inputs changed or that provide information about the importance of different input features.

Table 1 gives a (non-exhaustive) overview of some of these approaches. These provide different types of explanation, which include: descriptions of the process by which a system works; overviews of the way that a system creates a representation; and parallel systems that generate an output and an explanation using different models.

Choices made in data selection and model design influence the type of explainability that a system can support, and different approaches have different strengths and limitations. Saliency maps, for example, can help an expert user understand what data (or inputs) is most relevant to how a model works, but gives limited insight into how that information is used[21]. This may be sufficient for some purposes, but also risks leaving out relevant information.

21. Rudin, C. (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, Nature Machine Intelligence, 1, 206-215

TABLE 1

Different approaches to explainable AI address different types of explainability needs and raise different concerns[22]. What forms of AI explainability are available?

What type of explanation is sought	What method might be appropriate?	What questions or concerns do these methods raise?
Transparent details of what algorithm is being used	**Publishing the algorithm**	What form of explanation is most useful to those affected by the outcome of the system? Is the form of explanation provided accessible to the community for which it is intended? What processes of stakeholder engagement are in place to negotiate these questions? What additional checks might be needed at other stages of the decision-making pipeline? For example, how are the objectives of the system set? In what ways are different types of data used? What are the wider societal implications of the use of the AI system? How accurate and faithful is the explanation provided? Is there a risk it might mislead users? Is the desired form of explanation technically possible in a given context?
How does the model work?	**Inherently interpretable models** Use models whose structure and function is easily understood by a human user, eg a short decision list. **Decomposable systems** Structure the analysis in stages, with interpretable focus on those steps that are most important in decision-making. **Proxy models** Use a second – interpretable – model which approximately matches a complex 'black box' system.	
Which inputs or features of the data are most influential in determining an output?	**Visualisation or saliency mapping** Illustrate how strongly different input features affect the output from a system, typically performed for a specific data input.	
In an individual case, what would need to change to achieve a different output?	**Counterfactual (or example-based) explanations** Generate explanations focused on a single case, which identify the characteristics of the input data that would need to change in order to produce an alternative output.	

22. Adapted from Lipton, Z. (2016) The Mythos of Model Interpretability. ICML 2016 Workshop on Human Interpretability in Machine Learning (WHI 2016) and Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M. and Kagal, L. (2018) Explaining explanations: an overview of interpretability of machine learning. IEEE 5th International Conference on Data Science. DOI:10.1109/dsaa.2018.00018

In this context, the question for those developing and deploying AI is not simply whether it is explainable – or whether one model is more explainable than another – but whether the system can provide the type of explainability that is necessary for a specific task or user group (lay or expert, for example). In considering this, users and developers have different needs:

- For users, often a 'local' approach, explaining a specific decision, is most helpful. Sometimes, enabling an individual to contest an output is important, for example challenging an unsuccessful loan application.
- Developers might need 'global' approaches that explain how a system works (for example, to understand situations when it will likely perform well or badly).

Insights from psychology and social sciences also point to how human cognitive processes and biases can influence the effectiveness of an explanation in different contexts:

- Individuals tend to seek contrastive explanations – asking why one decision was made instead of another – rather than only asking why a particular outcome came about;
- Explanations are selective, drawing from a sub-set of the total factors that influenced an outcome in order to explain why it happened;
- Explanations that refer to the causes of an outcome are often more accessible or convincing than those that refer to probabilities, even if – in the context of AI – the mechanism is statistical rather than causal;
- The process of explaining something is often a social interaction – an exchange of information between two actors – which influences how they are delivered and received[23].

Boxes 3, 4 and 5 explore how some of these issues play out in different contexts.

23. Each of these reasons above is explored in detail in Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence. 267. 10.1016/j.artint.2018.07.007.

BOX 3

Science

Data collection and analysis is a core element of the scientific method, and scientists have long used statistical techniques to aid their work. In the early 1900s, for example, the development of the t-test gave researchers a new tool to extract insights from data in order to test the veracity of their hypotheses.

Today, machine learning has become a key tool for researchers across domains to analyse large datasets, detecting previously unforeseen patterns or extracting unexpected insights. Current application areas include:

- Analysing genomic data to predict protein structures, using machine learning approaches that can predict the three-dimensional structure of proteins from DNA sequences;
- Understanding the effects of climate change on cities and regions, combining local observational data and large-scale climate models to provide a more detailed picture of the local impacts of climate change; and
- Finding patterns in astronomical data, detecting interesting features or signals from vast amounts of data that might include large amounts of noise, and classifying these features to understand the different objects or patterns being detected[24].

In some contexts, the accuracy of these methods alone is sufficient to make AI useful – filtering telescope observations to identify likely targets for further study, for example. However, the goal of scientific discovery is to understand. Researchers want to know not just what the answer is but why.

Explainable AI can help researchers to understand the insights that come from research data, by providing accessible interpretations of how AI systems conduct their analysis. The Automated Statistician project, for example, has created a system which can generate an explanation of its forecasts or predictions, by breaking complicated datasets into interpretable sections and explaining its findings to the user in accessible language[25]. This both helps researchers analyse large amounts of data, and helps enhance their understanding of the features of that data.

24. Discussed further in Royal Society and Alan Turing Institute (2019) The AI revolution in science, available at https://royalsociety.org/topics-policy/data-and-ai/artificial-intelligence/
25. Further details available at: https://www.automaticstatistician.com/about/

BOX 4

Criminal justice

Criminal justice risk assessment tools analyse the relationship between an individual's characteristics (demographics, record of offences, and so on) and their likelihood of committing a crime or being rehabilitated. Risk assessment tools have a long history of use in criminal justice, often in the context of making predictions about the likely future behaviour of repeat offenders. For some, such tools offer the hope of a fairer system, in which human bias or socially-influenced perceptions about who is a 'risk' are less likely to influence how an individual is treated by the justice system[26]. The use of AI-enabled risk assessment tools therefore offers the possibility of increasing the accuracy and consistency of these predictive systems.

However, the opacity of such tools has raised concerns in recent years, particularly in relation to fairness and the ability to contest a decision.

In some jurisdictions, there already exists legislation against the use of protected characteristics – such as race or gender – when making decisions about an individual's likelihood of reoffending. These features can be excluded from analysis in an AI-enabled system. However, even when these features are excluded, their association with other features can 'bake in' unfairness in the system[27]; for example, excluding information about ethnicity but including postcode data that might correlate with districts with high populations from minority communities. Without some form of transparency, it can be difficult to assess how such biases might influence an individual's risk score.

26. Walklate, S. (2019) What would a Just Justice System Look Like? In Dekeseredy, W. and Currie, E. (Eds) Progressive Justice in an Age of Repression, Routledge, London.

27. Berk, R., Heidari, H., Jabbari, S., Kearns, M. and Roth, A. (2017) Fairness in Criminal Justice Risk Assessments: The State of the Art, Sociological Methods and Research, first published online 2 July, 2018: http://journals.sagepub.com/doi/10.1177/0049124118782533

In the US, there have already been examples of AI-enabled systems being associated with unfair judicial outcomes[28], and of those affected by its outputs seeking to contest its results[29]. In the debates that followed, the lack of transparency surrounding the use of AI – due to IP protections and trade secrets – were front and centre. This raised questions about whether a 'black box' algorithm violates a right to due process; what provisions for explainability or other forms of public scrutiny are necessary when developing AI tools for deployment in public policy domains; about how more explainable AI tools could balance the desire for transparency with the risk of revealing sensitive personal information about an individual[30]; and about the ways in which technological tools that appear neutral or authoritative could unduly influence their users[31]. These are important areas for more research.

Proposals to address these concerns have included:

- provision of additional information to judges and those working in the justice system to help them interpret the results of a system, and additional training for those individuals;
- the use of confidence estimates to help users interpret the results of a system;
- systems to evaluate, monitor, and audit algorithmic tools[32].

28. Partnership on AI (2018) report on algorithmic risk assessment tools in the US criminal justice system. Available at: https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/
29. For example: State v. Loomis (Wis 2016). Further information available at: https://harvardlawreview.org/2017/03/state-v-loomis/
30. Partnership on AI (2018) report on algorithmic risk assessment tools in the US criminal justice system. Available at: https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/
31. Hannah-Moffat, K. (2018) Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates', Theoretical Criminology. https://doi.org/10.1177/1362480618763582
32. Partnership on AI (2018) report on algorithmic risk assessment tools in the US criminal justice system. Available at: https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/

BOX 5

Health

Medical imaging is an important tool for physicians in diagnosing a range of diseases, and informing decisions about treatment pathways. The images used in these analyses – scans of tissue samples, for example – require expertise to analyse and interpret. As the use of such imaging increases across different medical domains, this expertise is in increasingly high demand.

The use of AI to analyse patterns in medical images, and to make predictions about the likely presence or absence of disease, is a promising area of research. To be truly useful in clinical settings, however, these AI systems will need to work well in clinical practice – and clinicians and patients may both want to understand the reasoning behind a decision. If doctors or patients are unable to understand why an AI has made a specific prediction, there may be dilemmas about how much confidence to have in that system, especially when the treatments that follow can have life-altering effects[33].

A recent research project by DeepMind and Moorfield's Eye Hospital points to new methods that can allow doctors to better understand AI systems in the context of medical imaging. This project looked at over 14,000 retinal scans, creating an AI system that analysed these images to detect retinal disease. Despite using deep learning techniques that would usually be considered 'black box', researchers built the system so that human users were able to understand why it had made a recommendation about the presence or absence of disease. This explainability was built into the system by making it decomposable.

The system itself consists of two neural networks, each performing different functions:

- The first analyses a scan, using deep learning to detect features in the image that are illustrative of the presence (or absence) of disease – haemorrhages in the tissue, for example. This creates a map of the features in the image.
- The second analyses this map, using the features identified by the first to present clinicians with a diagnosis, while also presenting a percentage to illustrate confidence in the analysis.

At the interface of these two systems, clinicians are able to access an intermediate representation that illustrates which areas of an image might suggest the presence of eye disease. This can be integrated into clinical workflows and interrogated by human experts wishing to understand the patterns in a scan and why a recommendation has been made, before confirming which treatment process is suitable. Clinicians therefore remain in the loop of making a diagnosis and can work with patients to confirm treatment pathways[34].

33. Castelvecchi, D. (2016) Can we open the black box of AI? Nature, 538, 20-23, doi:10.1038/538020a

34. De Fauw, J., Ledsam, J., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C., Raine, R., Hughes, J., Sim, D., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P., Suleyman, M., Cornebise, J., Keane, P., Ronneberger, O. (2018) Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24, 1342 – 1350 https://doi.org/10.1038/s41591-018-0107-6

Challenges and considerations when implementing explainable AI

While increasing the explainability of AI systems can be beneficial for many reasons, there are also challenges in implementing explainable AI. These include the following:

Different users require different forms of explanation in different contexts

Different contexts give rise to different explainability needs. As noted previously, system developers might require technical details about how an AI system functions, while regulators might require assurance about how data is processed, and those subject to a decision might want to understand which factors led to an output that affected them. A single decision or recommendation might therefore need to be explained in multiple ways, reflecting the needs of different audiences and the issues at play in different situations.

This can require different types of content; for example, technical information for a developer; accessible information for a lay-user. It can also put different types of demand on system design at each stage of the analytics pathway. To understand how a system works, users might (variably) wish to interrogate: which data the system used, the provenance of that data, and why that data was selected; how the model works, and which factors influence a decision; why a particular output was obtained. In order to understand what type of explanation is necessary, careful stakeholder engagement and system design are both necessary.

BOX 6

What do the results of public dialogue exercises say about the need for explainable AI?

Citizens juries in 2018 explored public views about explainability in AI across different application areas. These were commissioned by the Greater Manchester Patient safety Translational Research Centre and the Information Commissioner's Office (ICO), in partnership with the Alan Turing Institute and facilitated by Citizens' Juries c.i.c. and the Jefferson Centre. The juries deliberated over the importance of having an explanation in AI-enabled systems to: diagnose strokes; shortlist CVs for recruitment in a company; match kidney transplant donors and recipients; and select offenders for rehabilitation programmes in the criminal justice system. In each scenario, participants were asked to consider the relative importance of having an explanation for how the AI system worked, and the overall accuracy of the system.

Results from these juries showed that context is important when individuals evaluate the need for an explanation. Participants put a high priority on explanations being available in some contexts, while in others they indicated that other factors – in this case, accuracy – were more important. Discussions in the juries indicated that this variation was linked to different reasons for wanting an explanation: if it would be needed to challenge a decision or give feedback so an individual could change their behaviour, then an explanation became more important.

These results show that individuals make different trade-offs when evaluating whether an AI system is trustworthy. They suggest that, in some cases, it may be acceptable to deploy a 'black box' system if it can be verified as being more accurate than the available explainable method.

For example, in healthcare situations, jurors indicated they would prioritise the accuracy of a system, while in criminal justice they placed a high value on having an explanation in order to challenge a decision[35].

The results from the ICO's citizens jury echo the findings of the Royal Society's 2016 and 2017 public dialogues on machine learning.

In these dialogues, most people had not heard the term 'machine learning' – only 9% of those surveyed recognised it – but the majority had come across at least some of its applications in their day-to-day life. For example, 76% of respondents had heard about computers that can recognise speech and answer questions, as found in the virtual personal assistants available on many smartphones.

Attitudes towards machine learning – whether positive or negative – depended on the circumstances in which it is being used. The nature or extent of public concerns, and the perception of potential opportunities, were linked to the application being considered. People's views on particular applications of machine learning were often affected by their perception of who was developing the technology, and who would benefit.

35. ICO (2019) Project explain: interim report. Available at: https://ico.org.uk/media/2615039/project-explain-20190603.pdf

System design often needs to balance competing demands

AI technologies draw from a range of methods and approaches, each with different benefits and limitations. The AI method used in any application will influence the performance of the system on multiple levels, including accuracy, interpretability, and privacy.

Accuracy

There are different approaches to creating interpretable systems. Some AI is interpretable by design; these tend to be kept relatively simple. An issue with these systems is that they cannot get as much customisation from vast amounts of data that more complex techniques, such as deep learning, allow. This creates a performance-accuracy trade-off when using these systems in some settings, meaning they might not be desirable for those applications where high accuracy is prized over other factors. The nature of these requirements will vary across application area: members of the public have different expectations of systems used in healthcare versus those used in recruitment, for example[36].

Different models suit different tasks – for some structured problems, there can be little difference in performance between models that are inherently interpretable and 'black box' systems. Other problems require different forms of data analysis, drawing from 'black box' methods.

Privacy

In some AI systems – in particular, those using personal data or those where proprietary information is at stake – the demand for explainability may interact with concerns about privacy.

In areas including healthcare and finance, for example, an AI system might be analysing sensitive personal data in order to make a decision or recommendation. In considering the type of explainability that might be desirable in these cases, organisations using AI will need to take into account the extent to which different forms of transparency might result in the release of sensitive insights about individuals[37], or potentially expose vulnerable groups to harm[38].

There may also be cases in which the algorithm or data upon which it is trained are proprietary, with organisations reluctant to disclose either for business reasons. This raises questions about whether such systems should be deployed in areas where understanding why a decision was reached is important for public confidence or for ensuring accountability. In recent reviews of the application of AI in criminal justice, for example, there have been calls to rule-out the use of unintelligible systems[39].

36. ICO (2019) Project explain: interim report. Available at: https://ico.org.uk/media/2615039/project-explain-20190603.pdf
37. Weller, A. (2017) Challenges for transparency, from Workshop on Human Interpretability in machine learning (WHI), ICML 2017.
38. Schudson M (2015) The Rise of the Right to Know. Cambridge, MA: Belknap Publishing.
39. Partnership on AI (2018) report on algorithmic risk assessment tools in the US criminal justice system. Available at: https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/

Data quality and provenance is part of the explainability pipeline

Many of today's most successful AI methods rely on the availability of large amounts of data in order to make predictions or provide analysis to inform decisions. Understanding the quality and provenance of the data used in AI systems is therefore an important part of ensuring that a system is explainable. Those working with data need to understand how it was collected, and the limitations it may be subject to. For example, data used to create image recognition systems might not work well for minority groups (see Box 2), data from social media might reflect only certain communities of users, or sensor data from cities might only reflect certain types of neighbourhood. Explaining what data has been used by an AI system and how can therefore be an important part of ensuring that a system is explainable.

Explainability can have downsides

Explanations are not necessarily reliable

One of the proposed benefits of increasing the explainability of AI systems is increased trust in the system: if users understand what led to an AI-generated decision or recommendation, they will be more confident in its outputs[40].

However, not only is the link between explanations and trust complex[41], but trust in a system may not always be a desirable outcome[42]. There is a risk that, if a system produces convincing but misleading explanations, users might develop a false sense of confidence or understanding, mistakenly believing it is trustworthy as a result. Such misplaced trust might also encourage users to invest too much confidence in the effectiveness or safety of systems, without such confidence being justified[43]. Explanations might help increase trust in the short term, but they do not necessarily help create systems that generate trustworthy outputs or ensure that those deploying the system make trustworthy claims about its capabilities.

There can also be limitations on the reliability of some approaches to explaining why an AI has reached a particular output. In the case of creating post-hoc explanations, for example, in which one AI system is used to analyse the outputs of another (uninterpretable) system, there is the potential to generate explanations that are plausible – the second system appears to perform identically to the first – but inaccurate.

40. Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence. 267. 10.1016/j.artint.2018.07.007.
41. Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence. 267. 10.1016/j.artint.2018.07.007.
42. O'Neil, O. (2018) Linking trust to trustworthiness. International Journal of Philosophical Studies, 26, 2, 293 – 300 https://doi.org/10.1080/09672559.2018.1454637
43. Kroll, J. (2018) The fallacy of inscrutability, Phil. Trans. R. Soc. A 376: 20180084 https://doi.org/10.1098/rsta.2018.0084

The user has an explanation about how the system works, but one that is detached from the actual workings of the AI, making use of different features of the data or leaving out important aspects of information.

A further risk is deception: the use of tools that generate plausible interpretations that – intentionally or unintentionally – fool or manipulate people. For example, there is growing pressure to increase transparency around targeted advertising on social media, so that users are better able to understand how data about them is used by social media companies. In response, several companies are developing tools that explain to their users why they have seen particular content. However, it is not clear that these tools are effective in helping users understand how their online interactions influence the content they see. One study of the rationale given for why users received certain adverts found that the explanations provided were consistently misleading, missing key details that would have allowed users to better understand and potentially influence the ways in which they were targeted[44].

Gaming

Transparency can play an important role in supporting individual autonomy: if users have access to information about how a decision or recommendation has been made, they may be able to alter their behaviour to gain a more favourable outcome in future (discussed below).

However, this link between transparency and the ability to influence system outputs can have undesirable consequences in some contexts. For example, authorities investigating patterns of tax evasion may search for characteristics of a tax return that are correlated with evasion. If these indicators are widely known, those seeking to evade tax could adjust their behaviour in order to more effectively avoid detection[45]. Again, the extent to which this is an issue will likely vary across applications, and be influenced by the over-arching policy goal of the system[46].

Explainability alone cannot answer questions about accountability

The ability to investigate and appeal decisions that have a significant impact on an individual is central to systems of accountability, and the goal of some current regulatory approaches. In this context, explainable AI can contribute to systems of accountability by providing users with access to information and insights that allow them to appeal a decision or alter their behaviour to achieve a different outcome in future.

A variety of factors influence the extent to which an individual is effectively able to contest a decision. While explainability may be one, organisational structures, appeal processes, and other factors will also play a role in shaping how an individual can interact with the system. To create an environment that supports individual autonomy and creates a system of accountability, further steps are therefore likely to be needed. These might include, for example, processes by which individuals can contest the output of a system, or help shape its design and operation[47].

44. Andreou, A., Venkatadri, G., Goga, O., Gummadi, K. and Weller, A. (2018) Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations, Proceedings of the 24th Network and Distributed System Security Symposium (NDSS), San Diego, California, February 2018

45. Kroll, J., Huey, J., Barocas , S., Felten, E., Reidenberg, J., Robinson, D., and Yu, H. (2017) Accountable Algorithms, University of Pennsylvania Law Review, 165, 633

46. Edwards L., & Veale M. (2017). Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not the Remedy You Are Looking For, Duke Law & Technology Review, 16(1), 18–84, doi:10.2139/ssrn.2972855.

47. See, for example: https://afog.berkeley.edu/files/2018/08/AFOG_workshop_panel3_report.pdf

Explaining AI: where next?

Stakeholder engagement is important in defining what form of explainability is useful

As AI technologies are applied at scale and in spheres of life where the consequences of decisions can have significant impacts, pressures to develop AI methods whose results can be understood by different communities of users will grow. Research in explainable AI is advancing, with a diversity of approaches emerging. The extent to which these approaches are useful will depend on the nature and type of explanation required.

Different types of explanations will be more or less useful for different groups of people developing, deploying, affected by, or regulating decisions or predictions from AI, and these will vary across application areas. It is unlikely that there would be one method or form of explanation that would work across these diverse user groups. Collaborations across research disciplines and with stakeholder groups affected by an AI system will be important in helping define what type of explanation is useful or necessary in a given context, and in designing systems to deliver these. This requires working across research and organisational boundaries to bring to the fore differing perspectives or expectations before a system is deployed.

Explainability might not always be the priority in designing an AI system; or it may only be the starting point

There are already examples of AI systems that are not easily explainable, but can be deployed without giving rise to concerns about explainability – postal code sorting mechanisms, for example, or recommender systems in online shopping. These cases can generally be found in areas where there are no significant consequences from unacceptable results, and the accuracy of the system has been well-validated[48]. Even in areas where the system might have significant impacts, if the quality of results is high, the system may still enjoy high levels of user confidence[49].

The need for explainability must be considered in the context of the broader goals or intentions for the system, taking into account questions about privacy, accuracy of a system's outputs, the security of a system and how it might be exploited by malicious users if its workings are well-known, and the extent to which making a system explainable might raise concerns about intellectual property or privacy. This is not a case of there being linear trade-offs – increased explainability leading to reduced accuracy, for example – but instead of designing a system that is suitable for the demands placed on it.

48. Doshi-Velez F., Kim B. (2018) Considerations for Evaluation and Generalization in Interpretable Machine Learning. In: Escalante H. et al. (eds) Explainable and Interpretable Models in Computer Vision and Machine Learning. The Springer Series on Challenges in Machine Learning. Springer, Cham

49. See, for example, indications from public dialogue in: ICO (2019) Project explain: interim report. Available at: https://ico.org.uk/media/2615039/project-explain-20190603.pdf

In other cases, explainability might be a necessary pre-condition, but one that needs to be accompanied by deeper structures to build user confidence or systems of accountability. If the desired goal is to empower individuals in their interactions with AI, for example, then there may need to be broader mechanisms of feedback that allow those interacting with the system to interrogate its results, contest them, or be able to alter their outcomes by other means. Those designing a system will therefore need to consider how AI fits into the wider socio-technical context of its deployment. Given the range of AI methods that exist today, it is also the case that there are often less complex approaches – whose properties are well-understood – which can demonstrate performance as strong as 'black box' methods. The method selected need to be suitable for the challenge at hand.

Complex processes often surround human decision-making in critical domains, and a wider environment of accountability may need to develop around the use of AI

One line of argument against the deployment of explainable AI points to the lack of interpretability in many human decisions: human actions can be difficult to explain, so why should individuals expect AI to be different? However, in many critical areas – including healthcare, justice, and other public services – decision-making processes have developed over time to put in place procedures or safeguards to provide different forms of accountability or audit. Important human decisions often require explanation, conferring, or second opinions, and are subject to appeals mechanisms, audits, and other accountability structures. These reflect complex interactions between technical, political, legal and economic concerns, relying on ways of scrutinising the workings of institutions that include both explanatory and non-explanatory mechanisms[50]. Transparency and explainability of AI methods may therefore be only the first step in creating trustworthy systems.

50. Kroll, J. (2018) The fallacy of inscrutability, Phil. Trans. R. Soc. A 376: 20180084 https://doi.org/10.1098/rsta.2018.0084

The Royal Society's report *Science as an open enterprise* called for an environment of intelligent openness in the conduct of science[51]. In such an environment, information would be:

- Accessible: information should be located in such a manner that it can readily be found and in a form that can be used.
- Assessable: information should be held in a state in which judgments can be made as to the its reliability. For example, data should be differentiated for different audiences, or it may be necessary to disclose other information about the data that could influence an individual's assessment of its trustworthiness.
- Intelligible: Audiences need to be able to make some judgment or assessment of what is communicated.
- Useable: information should be in a format where others can use it, with proper background data supplied.

These criteria remain relevant to today's debates about the development of explainable AI. Further action to create an environment of intelligent interpretability – including both explainable AI methods and wider systems of accountability – could contribute to careful stewardship of AI technologies. These systems would need to take into account the full pipeline of AI development and implementation, which includes consideration of how objectives for the system are set, how the model is trained, what data is used, and the implications for the end user and society. Further engagement between policymakers, researchers, publics, and those implementing AI-enabled systems will be necessary to create such a stewardship environment.

51. Royal Society (2012) Science as an open enterprise. Available at: https://royalsociety.org/topics-policy/projects/science-public-enterprise/#targetText=The%20final%20report%2C%20Science%20as,research%20that%20reflects%20public%20values.

Annex 1: The policy environment for explainable AI: an outline

Policy and governance structures surrounding data use and AI are made up of a configuration of legal and regulatory instruments, standards, and professional and behavioural norms of conduct. In the UK, these structures include regulatory mechanisms to govern the use of data, emerging legal and policy frameworks governing the application of AI in certain sectors, and a range of codes of conduct that seek to influence the ways in which developers design AI systems. A sketch of this environment is below.

Regulation and data protection law

The EU's General Data Protection Regulation (GDPR) is the data protection and privacy regulation that governs the use of personal data in EU countries. Recent years have seen much debate about whether this regulation, which came into force in the UK in 2018, contains a 'right to an explanation'.

Article 22 of the Regulation focusses on "automated individual decision making, including profiling". It states that the person to whom data relates "shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her".

The Article states that, if such processing is necessary, the person will have "at least the right to obtain human intervention [...] to express his or her point of view and to contest the decision". Article 15 of the Regulation further requires that, if automated decision-making is used, the person to whom that relates should be able to access "meaningful information about the logic involved", upon request, while Articles 13 and 14 require such information be provided proactively[52].

Recital 71 also states that an individual subject to automated decision-making "should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision"[53].

Guidance that accompanies the GDPR text gives further insights into the nature of the explanation that might be relevant, for example stating: "the GDPR requires the controller to provide meaningful information about the logic involved, not necessarily a complex explanation of the algorithms used or disclosure of the full algorithm. The information provided should, however, be sufficiently comprehensive for the data subject to understand the reasons for the decision."

52. European Parliament and the Council of the European Union. 2016 EU General Data Protection Regulation – Article 22. Official Journal of the European Union 59, L119/1–L119/149.

53. In this context, the guidance implies that the term 'logic' is being employed to describe the rationale behind a decision, rather than a process of formal mathematical deduction, though this is a point of discussion in both AI and legal communities. This guidance does not have the same legal weight as the GDPR text. Article 20 Working Party (2018) Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, available at http://ec.europa.eu/justice/data-protection/index_en.htm

Early interpretations of these provisions suggested that they constituted a 'right to an explanation' of how an AI-enabled decision was made. However, debates since have noted that:

- Article 22 only applies to limited types of Automated Decision-Making (ADM) – that which is solely automated and has a legal or significant effect – and does not cover all AI-assisted decisions.
- any implied 'right to an explanation', in relation to Article 22 is not legally binding[54], raising questions about the nature and extent of any legal requirement to provide an explanation[55].
- the GDPR's provisions apply only to cases involving the processing of personal data.
- the GDPR includes specific exemptions from this provision in cases when an automated decision is necessary for a contract; if the decision is authorised by Member State law; and if an individual explicitly consents to a decision. There are also additional protections in relation to trade secrets and IP protection[56].

Debates continue about the nature of explanations required under the GDPR[57] – and the different ways in which its other provisions might require transparency from those processing data and using AI systems beyond those specifically covered by Article 22. Further interpretation by national and European courts may be necessary to better understand the limits of these provisions[58].

In the UK, the Information Commissioner's Office (ICO) is developing guidance for organisations, to support them in creating processes or systems that can assist in explaining AI decisions to the people affected by them[59]. As part of this work, the ICO and The Alan Turing Institute have carried out a series of citizen's juries in partnership with the National Institute for Health Research (NIHR) Greater Manchester Patient Safety Translational Research Centre (GM PSTRC). In an interim report from these juries, the ICO notes three key insights:

- the importance of context in shaping people's views;
- the need for improved education; and
- the challenges of deploying explainable AI[60].

54. Recitals are a source of guidance about the law, intended to help its interpretation.
55. 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, 7, 2, 76–99, https://doi.org/10.1093/idpl/ipx005
56. Edwards L., & Veale M. (2017). Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not the Remedy You Are Looking For Duke Law & Technology Review, 16(1), 18–84, doi:10.2139/ssrn.2972855
57. See, for example: Kaminski, M. (2018) The Right to Explanation, Explained. U of Colorado Law Legal Studies Research Paper No. 18-24; Berkeley Technology Law Journal, Vol. 34, No. 1, 2019 http://dx.doi.org/10.2139/ssrn.3196985; Edwards L., & Veale M. (2017). Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not the Remedy You Are Looking For Duke Law & Technology Review, 16(1), 18–84, doi:10.2139/ssrn.2972855; and 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, 7, 2, 76–99, https://doi.org/10.1093/idpl/ipx005
58. Royal Society and British Academy (2017) Data management and use: governance in the 21st century. Available at https://royalsociety.org/topics-policy/data-and-ai/
59. ICO (2019) Project explain: interim report. Available at: https://ico.org.uk/media/2615039/project-explain-20190603.pdf
60. ICO (2019) Project explain: interim report. Available at: https://ico.org.uk/media/2615039/project-explain-20190603.pdf

Emerging policy approaches across sectors

In several sectors, regulatory bodies are investigating whether AI-enabled products or services might raise questions about explainability, and whether existing frameworks are sufficient to manage any concerns that might follow:

- In transport, the UK's Automated and Electric Vehicles Act (2018) makes provisions for the liability on insurers or owners of autonomous vehicles. This answers some questions about whether an explanation would be required to allocate responsibility for an accident[61].
- The Financial Conduct Authority is commissioning research to gain a better understanding of the explainability challenges that arise when applying AI in the finance sector, and working with the International Organization of Securities Commissions to develop a framework for the application of ethical AI in financial services[62].
- The Medicines and Healthcare Regulatory Authority is working with the British Standards Institute to examine the extent to which existing frameworks for regulating medical devices are able to address the challenges posed by AI[63], and whether new standards in areas such as transparency and explainability might be necessary in validating the efficacy of such devices[64].

International bodies are also developing technical standards for areas including: data governance for children and students, and the transparency and accountability mechanisms that such governance should include; employer data collection practices, and the transparent and ethical handling of data about employees; and transparency in machine-machine decision systems[65].

Codes of conduct and ethical design

Across the world, companies, governments, and research institutes have published principles for the development and deployment of AI. Many of these include calls for forms of explainability or transparency, often as a means of creating accountability mechanisms surrounding AI, and ensuring that those subjected to decisions supported by AI systems have the information they may need to contest those decisions. A (non-exhaustive) list of such principles is in Table 2.

61. Reed C. 2018 How should we regulate artificial intelligence. Phil. Trans. R. Soc. A 376: 20170360. http://dx.doi.org/10.1098/rsta.2017.0360
62. See, for example: https://www.fca.org.uk/news/speeches/future-regulation-ai-consumer-good
63. Some further details available at: https://content.yudu.com/web/43fqt/0A43ghs/IssueTwoMarch2019/html/index.html?page=20&origin=reader
64. Some further details available at: https://www.bsigroup.com/globalassets/localfiles/en-gb/about-bsi/nsb/innovation/mhra-ai-paper-2019.pdf
65. IEEE Global Initiative for Ethical Considerations in AI and Autonomous Systems. Further details at: https://standards.ieee.org/news/2017/ieee_p7004.html

TABLE 2

Explainability in AI principles.

Author	Principle	Statement
UK Government – Data Science Ethical Framework	Make your work transparent and be accountable	"The more complex data science tools become, the more difficult it may be to understand or explain the decision-making process. This is a critical issue to consider when carrying out data science or any analysis in government. It is essential that government policy be based on interpretable evidence in order to provide accountability for a policy outcome[66]."
Google – Our Principles	Be accountable to people	"We will design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal. Our AI technologies will be subject to appropriate human direction and control[67]."
Microsoft – Our Approach to AI	Transparency	"Designing AI to be trustworthy requires creating solutions that reflect ethical principles that are deeply rooted in important and timeless values. [...] AI systems should be understandable[68]."
OECD and G20[69] – AI principles	Transparency and explainability	"Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art: i. to foster a general understanding of AI systems, ii. to make stakeholders aware of their interactions with AI systems, including in the workplace, iii. to enable those affected by an AI system to understand the outcome, and, iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision[70]."
Beijing Academy of Artificial Intelligence – Beijing AI principles	Be Ethical	"AI R&D should take ethical design approaches to make the system trustworthy. This may include, but not limited to: making the system as fair as possible, reducing possible discrimination and biases, improving its transparency, explainability, and predictability, and making the system more traceable, auditable and accountable[71]."
EU High Level Expert Group on AI – Ethics guidelines for trustworthy AI	Transparency	"[T]he data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations[72]."

66. Cabinet Office Data Science Ethics Framework, available at https://www.gov.uk/guidance/6-make-your-work-transparent-and-be-accountable
67. Google AI principles, available at: https://www.blog.google/technology/ai/ai-principles/
68. Microsoft AI principles, available at https://www.microsoft.com/en-us/ai/our-approach-to-ai
69. G20 statement on AI, available at: https://g20trade-digital.go.jp/dl/Ministerial_Statement_on_Trade_and_Digital_Economy.pdf
70. OECD AI principles, available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
71. Beijing AI principles, available at: https://www.baai.ac.cn/blog/beijing-ai-principles
72. EU High Level Group AI principles, available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Annex 2: Steering Group

Steering group

Members of a Steering Group that developed content for this briefing are listed below. Members acted in an individual and not a representative capacity, contributing to the project on the basis of their own expertise.

Steering group
Alan Bundy CBE FRS FRSE FREng, Professor of Automated Reasoning, University of Edinburgh
Jon Crowcroft FRS FREng, Marconi Professor of Communications Systems, University of Cambridge
Zoubin Ghahramani FRS, Professor of Information Engineering, University of Cambridge, and Chief Scientist, Uber
Nancy Reid OC FRS FRSC, Canada Research Chair in Statistical Theory and Applications, University of Toronto
Adrian Weller, Programme Director for AI, The Alan Turing Institute

Royal Society staff

Royal Society staff
Dr Natasha McCarthy, Head of Policy, Data
Jessica Montgomery, Senior Policy Adviser

Workshop participants

In addition to the literature cited, this briefing draws from discussions at a workshop held with the American Academy of Arts and Sciences in 2018, held under the Chatham House Rule. The Society would like to express its thanks to all those who presented and participated at this workshop.

The Royal Society is a self-governing Fellowship of many of the world's most distinguished scientists drawn from all areas of science, engineering, and medicine. The Society's fundamental purpose, as it has been since its foundation in 1660, is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity.

The Society's strategic priorities emphasise its commitment to the highest quality science, to curiosity-driven research, and to the development and use of science for the benefit of society. These priorities are:

- Promoting excellence in science
- Supporting international collaboration
- Demonstrating the importance of science to everyone

For further information
The Royal Society
6 – 9 Carlton House Terrace
London SW1Y 5AG

T +44 20 7451 2500
E science.policy@royalsociety.org
W royalsociety.org

Registered Charity No 207043

ISBN: 978-1-78252-433-5
Issued: November 2019 DES6051