THE ROYAL SOCIETY

AMERICAN ACADEMY OF ARTS & SCIENCES

AI, society and social good

Note of discussions at a Royal Society and American Academy workshop

8 November 2018, Centre for Advanced Study in the Behavioural Sciences, Stanford University

The Royal Society and American Academy of Arts and Sciences

The Royal Society is the UK's national academy of sciences. The Society's fundamental purpose, reflected in its founding Charters of the 1660s, is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity. In April 2017, the Society published the results of a major policy study on machine learning. This report considered the potential of machine learning in the next 5 – 10 years, and the actions required to build an environment of careful stewardship that can help realise its potential.

The American Academy of Arts and Sciences is one of the USA's oldest learned societies and independent policy research centres. Since its founding in 1780, the Academy has worked to champion scholarship, civil dialogue, and useful knowledge. It supports studies, publications, and programmes on science, engineering, and technology; global security and international affairs; education and the development of knowledge; humanities, arts, and culture; and American institutions, society, and the public good.

AI technologies are advancing at pace. New applications promise benefits in healthcare, transport, education, and more, bringing the possibility of more effective public services and economic growth. At the same time, these technologies pose new questions for society, and are encouraging new conversations about the ways in which AI technologies are shaping the world.

Experience of previous waves of transformative technologies shows that, even when societies well-understand the issues at hand, it is challenging to create responses that align the present with a desired future, engage collective action, and prepare mechanisms to support those who might be anxious, frustrated, or concerned at the pace or nature of technology-enabled change.

In this context, interdisciplinary discussions about the impact of AI on areas of societal interest are important in contributing to an environment in which the benefits of AI are brought into being safely and rapidly, and where these benefits are shared across society equitably and inclusively. To contribute to continuing science and policy debates about the impact of AI on society, on 8 November 2018 the Royal Society and American Academy convened a workshop to consider current understandings and future research directions in key areas of interest. This note summarises discussions at the workshop. It is not intended as a verbatim record, and does not reflect an agreed position by workshop participants or the Royal Society and American Academy.

Setting the scene: AI today and in the near-term

AI is an umbrella term. It refers to a suite of technologies in which computer systems are programmed to exhibit complex behavior, when acting in conditions of uncertainty.

Software engineering also pursues this goal, but focusses on comprehensively-representing the complexity of a computer system in order to ensure that it produces the desired outcome. Programmers set out the steps a system should take to achieve a goal, while minimising uncertainty, and following mathematical principles and Boolean approaches to logic. In contrast, AI development is more of an empirical science: machine learning systems make hypotheses about the world, and test them using data. These models are not provably correct, yet, but do allow the system to answer new queries, based on these theories.

There have been two paths to developing AI:

- In the 1970s and 1980s, programmers created expert systems. These codified how human experts would approach a question, and created computer systems that would follow these steps. While such systems achieved some notable successes, they were brittle: it is hard to account for all the possibilities or influences that arise when making a decision, and sometimes even human experts are not fully aware of their reasoning behind an action.
- Recent years have seen notable successes in machine learning, in which programmers feed machine learning systems a large amount of data and specify a goal or factor to optimise. Data could come from examples of inputs and outputs, or from broader sources of knowledge, such as tutorial instruction, analogy, or common sense. Deep learning is an approach to machine learning in which the system creates a map from input to output that has a large number of intermediate steps.

Machine learning methods are now employed across a range of sectors:

- In science, machine learning is helping analyse vast quantities of historical data from the Keppler satellite, contributing to the search for new exoplanets. This highly-complex dataset contains signals from a range of phenomena, which can cause variation in the levels of light that reaches Keppler's detectors. While statistical approaches have helped find many exoplanets amongst this noise, the sensitivity of machine learning techniques is allowing scientists to find further signals and produce new results.
- In medicine, machine learning applications have been developed to help diagnose eye disease by analysing retinal scans to identify indicators of disease. These systems are also being applied in new ways, for example in detecting high blood pressure.

AI today can help create highly accurate systems, which are able to automate increasingly sophisticated tasks, and which are becoming more available for use by those outside the AI research community. As the field progresses, it is increasingly grappling with questions that are not about accuracy, but about optimisation: namely, for what outcomes are AI systems optimised, and why? Answering these requires further examination of the systems into which AI technologies are developed and deployed.

At present, many markets favour systems that optimise for engagement, with an 'attention economy' emerging in which the number of 'clicks' received by a system is a marker of its success. In the short term, this might give people what they want, but how well do such systems serve societal needs? A range of organisations have produced guidelines or principles for the development of AI – and in computing such guidelines have existed for a long time – but at present the means for mechanising these goals require further development.

Influencing AI development to address areas of societal interest is a key challenge for the future. In some ways, it is possible to point to common agreement about what these interests are: international conventions on human rights or sustainable development point to areas of agreed need, for example. However, in other ways, the things that societies value seem to be highly contested: recent elections in the UK and US, for example, have revealed social or ideological divisions across society.

This workshop considered three areas where core societal values or interests are at stake as AI progresses, but where values or approaches might be contested: fairness and equality; transparency and interpretability; and democracy.

Fairness and equality

Bias comes from data and from people

A machine learning system combines data with a model – and objective function – in order to make a prediction. Each stage of this process can give rise to questions or concerns about fairness and equality.

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 for users from minority ethnic groups, for example. Without 'fair' data, these systems do not learn the statistical cues to recognise faces of people from a wide range of backgrounds.

The models created by a machine learning system can also generate issues of fairness or bias, even if trained on accurate data. 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.

When access to data is limited, for example in healthcare, baseline datasets may not be representative of the populations the AI systems hope to serve. Much of today's evidence based medical research data is limited to the average 50-year-old Caucasian male human or male rats, and the sex of stem cells is not always reported. AI-scaling in healthcare using systems that use only these limited datasets seems unlikely to equitably and inclusively serve the needs of a range of ages, races, genders and more.

To create systems that work well for a diverse range of users, researchers are grappling with questions such as: what populations are included in the training datasets and do they well-represent the populations that will interact with the system? Do the historical datasets represent the values we aspire to, vs. the historical discriminations of gender, race, age, and so on?

Graphic facilitation by Collective Next

Both technical and human interventions may be necessary

In seeking to resolve these issues, both technology-enabled and human-led solutions can play a role. Recent initiatives to address issues of bias in datasets, for example, include those such as *Datasheets for Datasets*, which seeks to provide details about standard operating characteristics and recommended usage for datasets that are made available for open use.

A challenge in designing effective interventions or approaches to manage fairness and bias issues in machine learning is in specifying desirable outcomes or objective functions. Terms including bias and discrimination can mean different things to different communities. Meanwhile there are instances where discrimination is acceptable, or desirable, in order to achieve equality of outcomes, and there are cases where there may be competing values at stake.

The bias-variance dilemma in machine learning provides a lens through which the trade-offs involved in addressing questions about bias can be considered. Machine learning can make different types of error, which need to be considered in specifying its objective functions. These can be simplified to:

- Overcomplicating the analysis, fitting models that are more complex than can be justified, given the data; or
- Producing a simple model, which over-simplifies the analysis.

Humans tend to make the first category of error, fitting models to the world that are more complex than can be justified, given the data. Human decision-making is inconsistent, as a result. Machine learning systems, meanwhile, err towards creating biased models, which give highly consistent results. While it can be tempting to view this consistency as desirable, the result can be consistent errors, with certain people consistently being treated incorrectly by the system. Such errors can be compounded in the creation of large-scale systems, with new categories of discrimination that are unobserved, and therefore not addressed.

In a dynamic system with much uncertainty – which might describe any area of human life – making consistent errors seems more likely to result in discrimination or unfair outcomes than allowing for inconsistency, or diversity of outcome. Variance in decision-making could help maintain fairness: different rules can be applied to different cases.

The need for diversity amongst developers and teams

A lack of diversity in the tech community can compound technical issues, as it can influence the extent to which developers are aware of potential biases as they create machine learning systems. A more diverse tech community would be better-placed to identify and tackle issues of fairness and equality at a human-level, rather than relying on post-hoc technical responses.

Diversity of expertise can also improve the effectiveness of AI systems: in many cases, knowledge of the environment into which a system will be deployed is important in ensuring that system works well. A combination of technical and domain insights is necessary to create systems that can be applied to 'real world' problems. This requires teams of people from a range of backgrounds and areas of expertise.

Access to AI

Further issues of equity arise in considering who is able to access different types of AI. For some, AI-enabled systems offer the hope of extending access to products or services to a wider range of users, who, for example, might not previously have been able to access legal advice. However, this potential comes with the risk of embedding current inequalities of access. In healthcare, for example, high-quality care combines clinical judgement with data analysis, bringing in second opinions and exploring multiple treatment options. While some might have access to such care, could AI-enabled systems that are optimised for efficiency or number of users ultimately reduce the care quality that is available to the wider population?

Key questions or issues

- Who defines when an AI system is unfair or unethical? How? What type of stakeholder engagement is necessary to ensure that AI systems are developed fairly and inclusively?
- If there is discrimination or unfairness, and if people are harmed, then what forms of redress are available?
- Which data should be available for use, and what do researchers need to understand in order to manage the risks in using biased data? How can this data be curated in a way that takes into account potential concerns about equity, bias or fairness?

Interpretability and transparency

The terms 'interpretability' or 'transparency' mean different things to different people. For some, they relate to how an algorithm works, or refer to data use, while for others they might relate to why an algorithm reaches an individual decision, or how that decision affects the system into which a machine learning algorithm is deployed. Words such as *interpretable, explainable, intelligible, transparent* or *understandable* are often used interchangeably, or inconsistently by those developing AI, and the expectations that come alongside these words might also vary by research discipline. In law, for example, explanations might focus on the intent of an actor, an idea that is not transferable to AI systems. In this context, interpretability risks becoming a 'catch-all' term, which ultimately loses its meaning.

A different approach to understanding what interpretability means is to approach it functionally. Interpretability can be desirable for a range of reasons:

- 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. Designing interpretable systems allows engineers to find the issue and develop a solution.
- Assessing risk: Understanding how a system works can be important in assessing risk, or creating confidence in the system. 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. In healthcare, this understanding might be necessary to inform decisions of underwriting of liability insurance for the use of AI. This can be particularly important if a system is deployed in a new environment, where the user cannot be sure of its effectiveness.
- Dealing with an adversarial environment: Interpretability can help navigate the challenges associated with adversarial examples or challenges, in which – for example – a small number of carefully-chosen pixel perturbations are used to influence the way in which an image recognition system works.
- Meeting regulatory or legal standards: Transparency or explainability can be important in enforcing legal rights surrounding a system, or in identifying cases of unfairness in how a system works. It can also be important in allowing those who are subject to the decisions made by a system to have a sense of agency in how they are treated. In this context, interpretability is not about placating users, but allowing those affected by decisions to feel a sense of dignity. Some form of interpretability may also be helpful in navigating questions about liability.

Graphic facilitation by Collective Next

- Enabling verification and validation: Interpretability can also be desirable in demonstrating the reproduceability of results, by tracing how data has been used by a model. There can be important links between ways of interrogating the micro-decisions of an AI system and software verification.

There are different approaches to creating interpretable systems. Some AI is interpretable by design; these tend to be shallow decision trees. An issue with these systems is that they do not get the leverage from vast amounts of data that techniques such as deep learning allow. This creates a performance-accuracy trade-off when using these systems, meaning they might not be desirable for those applications where high accuracy is prized over other factors.

Another approach is to create tools that can interrogate complex AI systems. These approaches take a deep neural network, for example, and apply ideas from neuroscience about how to analyse how the outputs from each 'neuron' vary with each data point used a stimulus. This allows the developer to understand, for example, which parts of a system are stimulated by different images.

In a similar vein, some researchers have created decomposable systems. For example, researchers at DeepMind created a system to diagnose eye diseases based on scans. This system split analysis into two parts: the first part segmented the image with meaningful labels. The second performed the diagnoses. This allowed doctors to look at the output of the system, then see what type of segment this output corresponded to. Similarly, many self-driving car systems split their world into segments, make predictions about each, then decide on actions as a result.

A current area of work for the *Automated Statistician* project is the creation of AI systems that produce interpretations for human users, building tools that approximate complex answers with more interpretable ones. Such as system might write a report to explain how data has been interpreted.

These different methods and approaches have benefits and limitations.

Risks and challenges associated with interpretability

Interpretability is not always desirable or positive, and can come with its own risks or challenges.

One such risk is deception. A system that creates plausible interpretations can secure false confidence from its users, which can then be used to fool or manipulate people. Such misplaced trust might also encourage users to invest too much confidence in the effectiveness or safety of systems.

Depending on the stage of analysis at which interpretability is required, there can be challenges associated with handling proprietary data or algorithms. In a healthcare setting, for example, treatments might be recommended on the basis of similarities between patients. However, it is unlikely that users would consider it acceptable to reveal personal medical details to explain why some data points are more or less similar to each other. Many algorithms, too, are proprietary technology, which creates challenges around accessing information.

The significance attached to interpretability – and the type of interpretability that is desirable – also depends on the context at hand. People's expectations will vary according to the type of application and the risks or benefits involved.

Human decision-making can often be opaque. It may be that algorithmic decisions are understood more clearly than human decisions, though it is unclear whether this will be sufficient to ensure confidence in their deployment.

In this context, the absence of a clear definition of interpretability might ultimately be helpful. It will be important as AI develops that researchers reach out to a range of communities to understand the nature of interpretability that is desirable to different groups, and the lessons that might come from a range of disciplines in understanding what interpretability might mean in different contexts.

Key questions or issues

- What role does interpretability play in a trustworthy system?
- In what contexts is interpretability important?
- How can researchers work across disciplinary boundaries and with user communities to develop useful understandings of what types of interpretability might be necessary in different contexts?
- What types of interaction between users and AI systems can support interpretability? How might trade-offs in creating interpretable systems be negotiated?

Democracy and civil society

Alone, together, at scale: AI and democratic debate

Early in the development of digital technologies, a great hope had been that they would enable people to connect and build communities in new ways, strengthening society and promoting new forms of citizen engagement. People would have access to more information from more diverse viewpoints, more opportunities for dialogue across social boundaries, and opportunities for common endeavour.

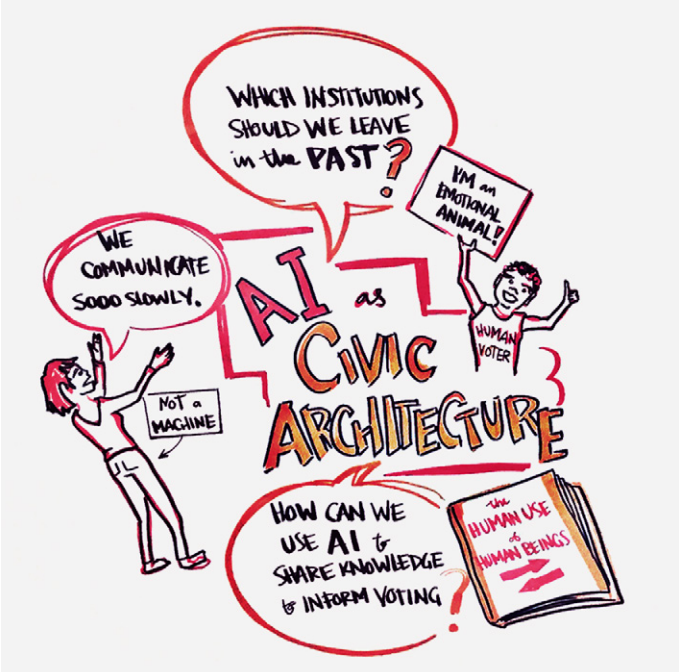
To some extent, this goal has been achieved: people have an opportunity to communicate to and with much broader – or much narrower – groups in ways that were not previously possible, exchanging views and gathering information from sources to which they would not previously have been exposed. However, these interactions play into a broader trend of the individuation of public life, a topic that has long been the concern of sociologists. AI potentially exacerbates this trend: people are simultaneously 'massified' and individuated – alone, together, at scale.

As new expressions of individualism and expression at scale emerge, there are new opportunities for collaboration, sometimes with unintended consequences. It is now possible for groups with extreme political views to connect and raise the profile of their cause in ways that previously would have required centralisation of resources. The information echo chambers that have existed in the physical world have found new manifestations in algorithmically-enabled filter bubbles, and the anonymity afforded by digital interactions has contributed to the coarsening of online political debate. New queries also arise about the trustworthiness of online information, in the absence of traditional gatekeepers.

These concerns intersect with debates about the role of private corporations in influencing the development of both AI and democracy. The dominance of a small number of technology companies in collecting data and public debate prompts concerns about power asymmetries between civic institutions and private corporations, and between consumers and providers of digital services.

The public and policy debates about AI and democracy that follow propose that these changes to the information environment might significantly influence the practice of democracy. Changing patterns in information exchange lead to changing political opinions, influencing voting patterns, electoral outcomes, and the institutions that sustain them. This pattern draws from enlightenment theories that suggest that people speak, listen, and act rationally, based on the information they receive. The actions suggested to respond to these challenges to democracy then concentrate on using AI to improve the circulation of information, providing people with the information necessary to make a rational decision.

However, insights from behavioural economics and associated disciplines show that people are not rational. Responses to information, evidence, and political debate are rooted in other behaviours: people find ways of rationalising their pre-existing beliefs, and the process of making political choices is influenced by emotions and social forces. AI technologies can exploit these base instincts, if designed to maximise user engagement, but through mechanisms that involve more than the exchange of information.



Graphic facilitation by Collective Next

Civic architecture: where does democracy live?

Democracy, too, is more than the exchange of information in campaigns and elections. It draws from a collection of institutions and civic interactions. In the context of shifting technologies, social norms, and behaviours, the rules and principles on which these institutions are founded can inject rationality into the democratic system. These institutions outlast any individual politician or leader. Democracy persists because institutions preserve it: in the press and the electoral process, but also in courts, in schools, in hospitals, in prisons, and more.

If democracy resides in institutions, then how can AI support them? What institutions serve the collective good? To make positive the impact of AI on democracy, it is necessary to identify those institutions that sustain democracy, and to create AI that can play a role in sustaining their missions. There is a need for spaces where people can develop civic networks or new civic institutions that allow people from different backgrounds to engage as citizens on common endeavours. Such spaces might enable community-development, but would be bolstered by an institutional element that provided rules and practices that outlast individuals or the changes to which communities are vulnerable.

In this context, questions about information exchange become less about the means of circulation, but instead about the ways in which institutions – including the press – can ensure that citizens have access to trustworthy sources of information. This requires architectures that sustain and validate information: libraries might be one example.

One response might be to develop datasets as a form of civic infrastructure. The open data movement in the UK has pursued this cause, demanding institutional data repositories with high standards for input and maintenance. An extension of this idea is the notion of a data trust, which holds data as a public good in an institution based on an ethical relationship of trust between trustees and the beneficiaries. Such trusts might provide a way of engaging citizens in collective action that can shape the development of AI technologies, by deciding with which trust they wish to engage, and with what data uses they are content.

Key questions or issues

- Which are the key institutions that will support democracy in an AI-enabled world?
- In what ways might AI support public service delivery?
- What new civil institutions or architectures might be required?
- How can AI facilitate collective action?

Participant list

(* = Academy Member)

Workshop Co-chairs

Peter Donnelly FRS, Professor of Statistical Science, University of Oxford; CEO, Genomics plc

*John Mitchell, Mary and Gordon Crary Family Professor, Stanford University

Participants

Michael Ananny, Associate Professor of Journalism and Communications, USC Annenberg School for Communication and Journalism

*John Seely Brown, Visiting Scholar and Advisor to the Provost, USC, Independent Co-Chairman, Deloitte Center for the Edge

Jenna Burrell, Associate Professor, School of Information, Co-director of the Algorithmic Fairness and Opacity Working Group (AFOG), University of California, Berkeley School of Information

Federica Carugati, Program Director, Center for Advanced Study in the Behavioral Sciences (CASBS), Stanford University

André Dua, Senior Partner, McKinsey & Company

John W. Etchemendy, Patrick Suppes Family Professor in the School of Humanities and Sciences, Stanford University

Zoubin Ghahramani FRS, Professor of Information Engineering, University of Cambridge; Chief Scientist, Uber

Yolanda Gil, Research Professor in Computer Science, Director of the Center for Knowledge-Powered Data Science, University of Southern California; President, Association for the Advancement of Artificial Intelligence (AAAI)Sharad Goel, Executive Director, Computational Policy Lab, Stanford University

Thore Graepel, Research Lead, DeepMind; Professor of Machine Learning, University College, London

Raia Hadsell, Senior Research Scientist, DeepMind

Sabine Hauert, Assistant Professor in Robotics, University of Bristol, President, Robohub.org

Jerry Jacobs, Professor of Sociology, University of Pennsylvania

Joseph Kahne, Ted and Jo Dutton Presidential Professor for Education Policy and Politics; Director of the Civic Engagement Research Group, University of California, Riverside

Neil Lawrence, Director, IPC Machine Learning, Amazon

Patrick Lin, Professor of Philosophy, Director of the Ethics + Emerging Sciences Group, California Polytechnic State University

Zachary Lipton, Assistant Professor, Carnegie Mellon University

Natasha McCarthy, Head of Policy – Data, The Royal Society

Angela McLean FRS, Professor of Mathematical Biology, University of Oxford

Vivienne Ming, Co-Founder, Socos Labs

Jessica Montgomery, Senior Policy Advisor, The Royal Society

*Peter Norvig, Director of Research, Google

Peter O'Hearn FRS, Professor of Computer Science, UCL; Research Scientist, Facebook

John Randell, John E. Bryson Director of Science, Engineering, and Technology Programs, American Academy of Arts and Sciences

Richard Re, Assistant Professor of Law, UCLA

Brendan P. Roach, Morton L. Mandel Presidential Fellow, American Academy of Arts and Sciences

*Londa Schiebinger, John L. Hinds Professor of History of Science, Stanford University

Nigel Shadbolt FRS, Principal, Jesus College, and Professorial Research Fellow, Department of Computer Science, University of Oxford; Chairman, Open Data Institute

Sonoo Thadaney Israni, Executive Director, Stanford Presence Center + Program in Bedside Medicine, Stanford University; co-chair National Academy of Medicine's Working Group on AI in Healthcare

Fred Turner, Harry and Norman Chandler Professor of Communication, Stanford University